Bayesian inference of structural brain networks with region-specific Dirichlet parametrisation

Louis Onrust

Abstract

In this paper we present an extension to a Bayesian framework for inference of structural brain networks. This framework provides a generative model that explicitly describes how structural brain networks lead to observed streamline distributions. Our extension consists of adding a hyperprior on the latent Dirichlet variables, such that we can capture global and region-specific behaviour within the streamline distributions. We apply these models on both simulated and empirical data. We show that the added flexibility of region-specific parametrisation is not needed for inference of the underlying structure of a brain network, and that the global model with less parameters is still sufficiently flexible to represent data even when generated from a model with region-specific Dirichlet parameters.

1 Introduction

Different brain areas are thought to be integrated into large-scale networks to support cognitive function. These networks can be categorised in terms of the characteristics that are modelled, such as functional, structural, and effective connectivity (Bullmore and Sporns, 2009).

Recent approaches for investigating structural organisation and functional coordination within these networks involve measures of connectivity among brain areas (see Rubinov and Sporns, 2010, for an overview). Structural brain networks presuppose white-matter tracts (edges) which are then used to model connectivity between spatially segregated brain regions (nodes). Functional coordination is concerned with understanding the relationship between activity in certain brain areas, constrained by structural connectivity. If human behaviour arises through interaction between multiple brain regions that together form networks, it is indispensable to model the brain network (e.g. Sporns et al., 2004; Stam and Reijneveld, 2007; Bassett and Bullmore, 2006). For both functional and structural connectivity, inferring the networks encompasses two steps. In the first step the white matter tracts are estimated. In the second step, the underlying network that captures which regions are connected is estimated, by using the tracts identified in the first step. In this paper we focus on structural connectivity, and particularly on the second step.

To infer networks from functional magnetic resonance imaging (fMRI) data, we first have to estimate white-matter tracts. With these tracts we can obtain the network that captures which regions are connected (Jones, 2010). The first step is diffusion-weighted imaging (DWI), which is a way to estimate structural connectivity of whole-brain networks in vivo. DWI is a magnetic resonance imaging (MRI) protocol, and it measures the restricted diffusion of water molecules, which is an indirect measure of the presence and orientation of white-matter tracts. Streamlines that represent the structure of fiber bundles can be determined by following the principal diffusion direction in individual voxels. This process is also known as deterministic tractography (Conturo et al., 1999). The data used in this paper is based on probabilistic tractographic methods, as it takes the uncertainty of the streamlining process in account. The main difference is that in the probabilistic method, the principal diffusion direction is used to sample distributions of streamlines. Traditionally, streamline distributions were obtained by iteratively running the tractography algorithm. These distributions are then thresholded to reflect the presence or absence of tracts, which populate the underlying brain network. Since probabilistic streamlining relies on the number of samples that are drawn per voxel, as more samples are drawn, more regions are likely to become connected. Also, these weights do not have a straightforward (probabilistic) interpretation. See Hinne et al. (2013) where this discussion is more
elaborate. The authors also provide a Bayesian framework for the inference of whole-brain networks from streamline distributions. They consider the distribution of networks that are supported by the data, rather than generating a single network based on an arbitrary threshold. The framework relies on a generative model which comprises two parts. The first part is a network prior, and second, they propose a forward model based on a Dirichlet compound multinomial distribution which views the streamline distributions produced by probabilistic tractography as noisy data.

In this work we build on this recent work on estimating structural connectivity from diffusion data. Hinne et al. (2013) investigated a model where the probability of streamlining is determined by two parameters: \( d_0 \) and \( d_1 \).\(^1\) However, further investigations showed that the model exhibits behaviour that cannot be modelled properly with two global parameters. For example, the brain is not fully connected. Some parts are highly connected, some parts are only indirectly connected. This difference in connectedness between regions, is also apparent in the streamline counts. Some regions show very sparse but intense streamlining to some regions, while there are no streamline counts found between other regions. On the other hand, there are also regions that have outgoing streamlines to many other regions, albeit with a lower intensity. At the expense of having a more complicated model with more parameters and less data to train the model on, one expects that the added richness of the regional model, where we parametrise each region, results in more sensitive estimations.

2 Method

In this paper we extend the current model to capture the difference in regions with different streamline distributions. To that end, we first extend the current model with a hyperprior on the parameters. Then, we specialise this global model to a regional model, where each region’s parameters are governed by the hyperprior.

2.1 Generative Model

We are interested in the posterior over structural networks represented by the adjacency matrix \( \mathbf{A} = (a_1, \ldots, a_k) \) given observed probabilistic streamlining data \( \mathbf{C} \) and hyperparameters \( \eta = (\mathbf{D}, \alpha, \beta) \), where \( k \) is the number of regions of interest. An element \( a_{ij} \) represents the absence or presence of an edge between region \( i \) and region \( j \). The posterior for \( \mathbf{A} \) is:

\[
P(\mathbf{A}|\mathbf{C}, \alpha, \beta, \mathbf{D}) \propto P(\mathbf{C}|\mathbf{A}, \mathbf{D})P(\mathbf{A}|\alpha, \beta)
\]

with Dirichlet parameters \( \mathbf{D} = (d_1; d_0) = (d_1^1, \ldots, d_1^k; d_0^1, \ldots, d_0^k) \). To infer the posterior distribution over the latent structural network \( \mathbf{A} \), we must specify a prior \( P(\mathbf{A}|\alpha, \beta) \) and a forward model \( P(\mathbf{C}|\mathbf{A}, \mathbf{D}) \). Together they define a generative model of probabilistic streamlining data.

2.1.1 Network Prior

The prior of the adjacency matrix \( P(\mathbf{A}|\alpha, \beta) \) is taken to be a Beta-Binomial distribution with hyperparameters \( \alpha \) and \( \beta \):

\[
P(\mathbf{A}|\alpha, \beta) = \int P(p|\alpha, \beta)P(\mathbf{A}|p)dp
\]

The adjacency matrix has a density \( p \) given by the mode of the Beta function with hyperparameters \( \alpha \) and \( \beta \), such that

\[
P(p|\alpha, \beta) = \text{Beta}(p|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} p^{\alpha-1}(1-p)^{\beta-1},
\]

\(^1\)In Hinne et al. (2013) the parameters \( d_0 \) and \( d_1 \) are called \( a^- \) and \( a^+ \) respectively.
for which the mode is defined as \( \frac{\alpha - 1}{\alpha + \beta - 2} \).

Since we do not model directed connectivity, but only whether there is a connection between two brain areas, the matrices are symmetrical. Furthermore, we assume that the edges are irreflexive, hence the diagonal of the matrix is populated with zeros. We model this with an indicator function \( 1_{\{A \in M\}} \) where \( M \) denotes the set of all symmetric matrices whose diagonals consists of all zeros, such that we can extend the original definition of the prior:

\[
P(A|\alpha, \beta) = 1_{\{A \in M\}} \int P(p|\alpha, \beta)P(A|p)dp.
\]  

(4)

2.1.2 Forward Model

The forward model describes how observed streamline counts \( C \) depend on the underlying network \( A \) through latent streamline probabilities \( X \). With probabilistic tractography, streamlines are drawn over a distribution of target nodes resulting in streamline counts \( C = (c_1, \ldots, c_k) \). \( S \) is the upperbound of drawn streamlines from one node, with \( b_i = \sum_j c_{ij} \leq S \).

The distribution \( c_i \) depends on the underlying latent streamline probabilities \( x_i \), which is a probability vector with \( \sum_j x_{ij} = 1 \). Each \( x_{ij} \) represents the probability of drawing a streamline from region \( i \) to region \( j \) and depends on whether a physical tract exists between region \( i \) and region \( j \). This idea can be captured with a multinomial distribution:

\[
P(c_i|x_i) \propto \prod_{j=1}^{K} x_{ij}^{c_{ij}}
\]  

(5)

The physical tracts are modelled by \( A \) as actual white-matter connectivity, where \( a_{ij} \) indicates the connectivity between region \( i \) and all other regions. The intuition is that we expect a high streamline probability if there is an edge between two regions. The other way around, we also expect a low probability if there is no connection between the two regions. The streamline probabilities \( X \) are modelled from an underlying adjacency matrix with a Dirichlet distribution:

\[
P(x_i|a_i, \eta) \propto \prod_{j=1}^{K} x_{ij}^{\delta_{ij} - 1}
\]  

(6)

with \( \delta'_{ij} = a_{ij}d_i^1 + (1 - a_{ij})d_0^1 \). An extension of the original model to a global model is to place hyperpriors on \( d_0 \) and \( d_1 \). In this work we compare such an extension with a regional approach, where each region \( r \) has its own \( d_r^0 \) and \( d_r^1 \).

In the global model, each Dirichlet parameter \( d \) comes from a Gamma distribution with hyperparameters \( k \) and \( \theta \). We assume that the parameters that indicate the presence of an edge are from the same distribution given by \( \text{Gamma}(d_1|k_1, \theta_1) \), where

\[
\text{Gamma}(d|k, \theta) = \frac{1}{\theta^k \Gamma(k)} d^{k-1} \exp \left( -\frac{d}{\theta} \right).
\]  

(7)

Similarly, we assume that all parameters that indicate the absence of an edge are from \( \text{Gamma}(d_0|k_0, \theta_0) \).

The likelihood of the network \( A \) is thus expressed as

\[
P(C|A, D) = \int P(C|X)P(X|A, D)dX
\]  

(8)

In Equation (8) we can recognise a product of Dirichlet compound multinomial (DCM) distributions, since the Dirichlet distribution is the conjugate prior for the multinomial distribution:

\[
P(c_i|x_i, \eta) \propto \int \left( \prod_{j=1}^{K} x_{ij}^{c_{ij}} \right) \left( \prod_{j=1}^{K} x_{ij}^{\delta'_{ij} - 1} \right) dX = \int \prod_{j=1}^{K} x_{ij}^{c_{ij} + \delta'_{ij} - 1} dX
\]  

(9)
A DCM distribution assumes that, given a network, a probability vector can be drawn such that large values occur where the network has edges, and small values where the network has no edges. With this probability vector we can sample from a multinomial distribution, such that the outcome reflects that of probabilistic tractography. With this sampling method we can mimic the burstiness behaviour we observe in streamline distributions, where some pairs of regions are connected by many streamlines, whereas most pairs have few or even zero streamlines.

The complete generative model for the global model is given by

\[ a_{ii} \leftarrow 0 \quad \text{for } i = 1, \ldots, K \]
\[ a_{ij} \leftarrow a_{ji} \quad \text{for } i = 1, \ldots, K, j = 1, \ldots, i - 1 \]
\[ p|\alpha, \beta \sim \text{Beta}(\alpha, \beta) \]
\[ a_{ij}|p \sim \text{Bernoulli}(p) \quad \text{for } i = 1, \ldots, K, j = i + 1, \ldots, K \]
\[ d_0|k_0, \theta_0 \sim \text{Gamma}(k_0, \theta_0) \]
\[ d_1|k_1, \theta_1 \sim \text{Gamma}(k_1, \theta_1) \]
\[ \delta_i \leftarrow a_i d_1 + (1 - a_i) d_0 \quad \text{for } i = 1, \ldots, K \]
\[ y_i \sim \text{Dirichlet}(\delta_i) \quad \text{for } i = 1, \ldots, K \]
\[ c_i \sim \text{Multinomial}(b_i, y_i) \quad \text{for } i = 1, \ldots, K \]

Extending the global model to a regional model is trivial, with small differences for the variables that are directly involved with the Dirichlet parameters:

\[ d_0^i|k_0, \theta_0 \sim \text{Gamma}(k_0, \theta_0) \quad \text{for } i = 1, \ldots, K \]
\[ d_1^i|k_1, \theta_1 \sim \text{Gamma}(k_1, \theta_1) \quad \text{for } i = 1, \ldots, K \]
\[ \delta_i \leftarrow a_i d_1^i + (1 - a_i) d_0^i \quad \text{for } i = 1, \ldots, K \]

Notice that although each region \( i \) has its own set of Dirichlet parameters \( d_0^i \) and \( d_1^i \), the Gamma distribution from which these parameters are sampled, is the same for all regions.

### 2.2 Approximate inference

Since the posterior (Equation (1)) cannot be calculated analytically, we use a single-component sampler, a Markov Chain Monte Carlo (MCMC) scheme, to sample from this distribution. Iteratively each of the parameters is updated in three situations. First, sampling proceeds by flipping each of the pairs \((a_{pq}, a_{qp})\) for \( p \neq q \) while keeping \( d_0 \) and \( d_1 \) fixed. Second, \( d_0 \) is updated while keeping \( A \) and \( d_1 \) fixed. And last, \( d_1 \) is updated and \( A \) and \( d_0 \) are kept fixed. The acceptance of the proposed samples is determined by the ratio

\[
\gamma = \frac{P(C, A', d_0^i, d_1^i, p|\alpha, \beta, k_0, k_1, \theta_0, \theta_1)}{P(C, A, d_0, d_1, p|\alpha, \beta, k_0, k_1, \theta_0, \theta_1)},
\]

where \((A', d_0^i, d_1^i)\) has one of the parameters flipped with respect to \((A, d_0, d_1)\). A proposed sample is accepted with probability \( \min(1, \gamma) \).

In general, \( P(C, A, d_0, d_1, p|\alpha, \beta, k_0, k_1, \theta_0, \theta_1) \) can be decomposed as follows:

\[
P(C, A, d_0, d_1, p|\alpha, \beta, k_0, k_1, \theta_0, \theta_1) = P(p|\alpha, \beta) \prod_j P(d_0^j|k_0, \theta_0) \prod_j P(d_1^j|k_1, \theta_1) \times \\
\prod_j P(c_j|b_i, a_j, d_0^j, d_1^j) \prod_{i<j} P(a_{ij}|p) 1_{\{A \in M\}}
\]

where \(1_{\{A \in M\}} \) is the indicator function, and \( M \) denotes the set of all symmetric matrices whose diagonal consists of all zeros.
Note that we can integrate out the dependence on $p$ since

$$P(A|\alpha, \beta) = \int P(p|\alpha, \beta) \left( \prod_{i<j} P(a_{ij}|p) \right) dp$$

$$= \int \frac{1}{B(\alpha, \beta)} p^{\alpha-1}(1-p)^{\beta-1} \prod_{i<j} p^{a_{ij}}(1-p)^{1-a_{ij}} dp$$

$$= \frac{1}{B(\alpha, \beta)} \int p^{\alpha-1+\sum_{i<j} a_{ij}}(1-p)^{\beta-1+\sum_{i<j} (1-a_{ij})} dp$$

$$= \frac{B(\alpha + \sum_{i<j} a_{ij}, \beta + \sum_{i<j} (1-a_{ij}))}{B(\alpha, \beta)} .$$

The new distribution is thus

$$P(d_0, d_1, A, C|\alpha, \beta, k_0, \theta_0, k_1, \theta_1) = \prod_j P(d'_0|k_0, \theta_0) \prod_j P(d'_1|k_1, \theta_1) \times$$

$$\prod_j P(c_{ij}|b_i, a_i, d'_0, d'_1) P(A|\alpha, \beta) 1_{\{A \in M\}} .$$

It proves to be more convenient to compute the log acceptance ratio $\gamma$:

$$\log \gamma = \log P(d'_0, d'_1, A', C|\alpha, \beta, k_0, \theta_0, k_1, \theta_1) - \log P(d_0, d_1, A, C|\alpha, \beta, k_0, \theta_0, k_1, \theta_1)$$

$$= \log \left( \sum_i \left( \log P(c_{ij}|b_i, a'_i, d'_0, d'_1) - \log P(c_{ij}|b_i, a_i, d'_0, d'_1) \right) \right)$$

$$= \sum_i T'_i$$

where

$$T'_i = \log P(c_{ij}|b_i, a'_i, d'_0, d'_1) - \log P(c_{ij}|b_i, a_i, d'_0, d'_1)$$

$$= - \sum_j \log \frac{\Gamma(\delta_{ij})}{\Gamma(\delta'_{ij})} + \log \frac{\Gamma(\sum_j \delta_{ij})}{\Gamma(\sum_j \delta'_{ij})} + \sum_j \log \frac{\Gamma(c_{ij} + \delta_{ij})}{\Gamma(c_{ij} + \delta'_{ij})} - \log \frac{\Gamma(\sum_j c_{ij} + \sum_j \delta_{ij})}{\Gamma(\sum_j c_{ij} + \sum_j \delta'_{ij})} .$$

If we update region $i$, then for all regions $j \neq i$, $T'_i = 0$, and therefore $T_1 = T'_i$.  

$$T_2 = \log P(A'|\alpha, \beta) - \log P(A|\alpha, \beta)$$

$$= \log \frac{B(\alpha + \sum_{i<j} a'_{ij}, \beta + \sum_{i<j} (1-a'_{ij}))}{B(\alpha + \sum_{i<j} a_{ij}, \beta + \sum_{i<j} (1-a_{ij}))} .$$
Similar to the simplification of $T_1 = T_1'$, $T_3$ can be reduced to $T_3'$ for a sample update in region $i$.

$$T_3' = \log P(d_0' | k_0, \theta_0) - \log P(d_0' | k_0, \theta_0)$$

$$= (k_0 - 1) \log \frac{d_0'}{d_0} + \frac{d_0' - d_0}{\theta_0}$$

(26)

(27)

The derivation of $T_4$ is analogous to $T_3$, such that

$$T_4' = (k_1 - 1) \log \frac{d_1'}{d_1} + \frac{d_1' - d_1}{\theta_1}$$

(28)

Now, if we want to compute whether we should accept an update $A'$ for network $A$, we should compute $T_1$ through $T_4$. However, since $d_0$ and $d_1$ are fixed, we do not have to compute the acceptance ratios $T_3$ and $T_4$. Therefore, for a flip from $a_{pq}$ to $a_{eq}$, log $\gamma = T_{1,A'} + T_2$, where $T_{1,A'}$ is

$$T_{1,A'} = \log \frac{P(c_i | b_p, a'_p, d_0, d_1)}{P(c_i | b_q, a'_q, d_0, d_1)} + \log \frac{P(c_i | b_q, a'_q, d_0, d_1)}{P(c_i | b_q, a'_q, d_0, d_1)}$$

$$= \log \frac{\Gamma(\delta_{pq} + c_{pq})}{\Gamma(\delta_{pq} + c_{pq})} + \log \frac{\Gamma(\delta_{pq} + c_{pq})}{\Gamma(\delta_{pq} + c_{pq})} + \log \frac{\Gamma(\sum_j c_{pj})}{\Gamma(\sum_j c_{pj})} + \log \frac{\Gamma(\sum_j c_{pq})}{\Gamma(\sum_j c_{pq})} + \log \frac{\Gamma(\delta_{pq} + c_{pq})}{\Gamma(\delta_{pq} + c_{pq})} - \log \frac{\Gamma(\delta_{pq} + c_{pq})}{\Gamma(\delta_{pq} + c_{pq})}$$

(29)

(30)

To compute the acceptance rate for $d_0$ samples in region $i$, while keeping $A$ and $d_1$ fixed, we have to compute $\gamma_{d_0} = T_{1,d_0} + T_3$, where

$$T_{1,d_0} = \log \frac{P(c_i | b_i, a, d_0', d_1)}{P(c_i | b_i, a, d_0, d_1)}$$

$$= -\sum_j \log \frac{\Gamma(\delta_{ij})}{\Gamma(\delta_{ij})} + \log \frac{\Gamma(\sum_j c_{ij} + \delta_{ij})}{\Gamma(\sum_j c_{ij} + \delta_{ij})} + \sum_j \log \frac{\Gamma(c_{ij} + \delta_{ij})}{\Gamma(c_{ij} + \delta_{ij})} - \log \frac{\Gamma(\sum_j c_{ij} + \delta_{ij})}{\Gamma(\sum_j c_{ij} + \delta_{ij})}$$

(31)

(32)

The procedure for $T_{1,d_1}$ is similar ($\gamma_{d_1} = T_{1,d_1} + T_4$), except for an obvious difference in the update rule:

$$T_{1,d_1} = \log \frac{P(c_i | b_i, a, d_0, d_1)}{P(c_i | b_i, a, d_0, d_1)}$$

$$= -\sum_j \log \frac{\Gamma(\delta_{ij})}{\Gamma(\delta_{ij})} + \log \frac{\Gamma(\sum_j c_{ij} + \delta_{ij})}{\Gamma(\sum_j c_{ij} + \delta_{ij})} + \sum_j \log \frac{\Gamma(c_{ij} + \delta_{ij})}{\Gamma(c_{ij} + \delta_{ij})} - \log \frac{\Gamma(\sum_j c_{ij} + \delta_{ij})}{\Gamma(\sum_j c_{ij} + \delta_{ij})}$$

(33)

(34)

### 3 Evaluation and Results

In the previous section we introduced the global and regional model. In this section we present experimental setup, to see if the regional model can outperform the global model. Furthermore, we present the results, and evaluate the results. The discussion of these results is the subject of the last chapter (Section 4.1).

#### 3.1 Experimental setup

To show whether the region-based model improves over the global model, we run four experiments on simulated data:

- Global Dirichlet parameters with inference based on the regional model ($G_{simRinf}$);
- Global Dirichlet parameters with inference based on the global model ($G_{simGinf}$);
- Regional Dirichlet parameters with global inference ($R_{simGinf}$);
• Regional Dirichlet parameters with regional inference (RsimRinf).

The goal of these experiments is first, can we show that if the data is generated with global Dirichlet parameters, we can also use a regional model to estimate the underlying network? Second, can we show that if one extends the approach to a regional model, global inference can be improved on?

Since the objective is to estimate the underlying adjacency matrix, we focus the evaluation on metrics of similarity on the estimated and ground truth adjacency matrix. We use metrics known from Receiver Operating Characteristic (ROC) curves, where each estimated edge that also occurs in the ground truth, is a true positive. Similarly, every edge that is correctly estimated to be absent, is a true negative. The four metrics that we use are:

- **density** the number of edges estimated to be present, divided by the total number of edges;
- **sensitivity** the number of true positives divided by the total number of edges estimated to be present;
- **specificity** the number of true negatives divided by the total number of edges estimated to be absent;
- **accuracy** the sum of true results (true positives and true negatives), divided by the sum of absent and present estimated edges.

Because each step in MCMC is only a sample, without context about previous samples, we collect micro statistics. Each step, we compute the four metrics, and we average over those values in the end. For the posterior distributions over densities and accuracies, we also provide the 95% credible interval. For all experiments we use 20% of the samples as burnin period. These samples are excluded from the statistics.

### 3.2 Generating Simulated Streamline data

We start with a uniformly distributed random matrix $R$ of size $k \times k$. The matrix is transformed into a symmetric matrix with zeros on the diagonal, to respectively express the bidirectionality and irreflexivity of real data. $R$ is then thresholded into a (binary) adjacency matrix $T$, such that if $r_{ij} > \tau$, $t_{ij}$ is marked as having an edge. $\tau$ is chosen such that $\sum_{m<n} t_{mn} / \sum_{m,n} t_{mn} \approx \hat{p}$ where $\hat{p}$ is the mode of the prior Beta distribution for $p$. In the remainder of this paper, we use the same matrix $T$ in the experiments, unless noted otherwise.

The parameters for the simulation of streamline data are equivalent to the parameters for the generative model: $\alpha, \beta$ for the density of the adjacency matrix; $k_0, \theta_0, k_1, \theta_1$ for the Gamma distribution to sample $d_0$ and $d_1$; $k$ for the size of the matrix, and $S$ for the maximum number of streamlines to draw from each node.

Based on the adjacency matrix $T$ we can now construct the Dirichlet parameters with $v_i = d_i^0 t_i + d_i^0 (1 - t_i)$ which can then be used in a multinomial sample $s_i$. First we draw a sample $y_i$ from the Dirichlet distributions with our parameters $v_i$: $y_i = \text{Dirichlet}(v_i)$ which are in turn used as parameter for a Multinomial distribution: $s_i = \text{Multinomial}(s_i; S, y_i)$.

### 3.3 Streamline quality

The quality of the generated streamlines is determined by the values of the Dirichlet parameters. Recall that $d_1$ and $d_0$ can be interpreted as the probability of streamlining between region $i$ and $j$ if an edge $a_{ij}$ is present, respectively the probability of streamlining when such an edge is absent.

If we then assume that $d_1 = d_0$, there is no distinction anymore, and it models a uniform streamlining probability. This means that the quality of streamlines is very poor, since the influence of the adjacency matrix is rendered useless, hence, there is too much noise in the streamlining data. However, if we take $d_1 \gg d_0$, the quality of the streamlines is too good due to the absence of noise. Streamlines only occur when there is an edge, and $d_0$ has too little influence to give streamlines to regions where an edge was thought to be absent.
We can illustrate this problem with an experiment where we vary $k$ and $\theta$ from $\hat{P}(d_0|k_0, \theta_0) > \hat{P}(d_1|k_1, \theta_1)$ to $\hat{P}(d_0|k_0, \theta_0) \ll \hat{P}(d_1|k_1, \theta_1)$, where $\hat{P}$ is the mode (argmax) of the probability distribution. In the experiment we sample from a 25×25 matrix, in 25000 steps (of which 5000 are used as burnin), and $d_0$ is drawn from Gamma(5, 0.12) (mode is 0.48) and we vary the Gamma distribution over $d_1$.

Table 1: In the first experiment, with an accuracy of 56.1%, it only performs marginally better than random assignment of edges. As $d_1$ moves from $d_0$, the quality of the streamlines improves, to a point where a given streamline count matrix is not improved upon, and it is implicitly converted to an adjacency matrix. For example for the last experiment, only three network proposals were accepted out of 7.5M proposals (matrix with size 25, and 25000 steps with Metropolis). The experiment was done with region-based simulating and global inference (RsimGinf). The numbers are macro-averages taken from the micro-averages of 3 runs. $k_1$ and $\theta_1$ are the parameters of the Gamma distribution, with $\hat{d}_1$ being the mode of that distribution. The density is the mean density of the estimated adjacency matrices. Sensitivity, specificity and accuracy denote respectively the true positive rate, true negative rate, and the proportion of true results.

<table>
<thead>
<tr>
<th>$k_1$</th>
<th>$\theta_1$</th>
<th>$\hat{d}_1$</th>
<th>density (95% CI)</th>
<th>sensitivity</th>
<th>specificity</th>
<th>mean accuracy (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>0.03</td>
<td>0.45</td>
<td>37.3 [10.56,58.56]</td>
<td>38.5</td>
<td>52.8</td>
<td>56.1 [45.60,67.68]</td>
</tr>
<tr>
<td>20</td>
<td>0.05</td>
<td>0.95</td>
<td>36.5 [16.00,58.88]</td>
<td>49.7</td>
<td>72.6</td>
<td>63.3 [54.56,71.84]</td>
</tr>
<tr>
<td>25</td>
<td>0.06</td>
<td>1.44</td>
<td>36.7 [18.24,30.08]</td>
<td>68.5</td>
<td>70.5</td>
<td>74.0 [82.08,88.48]</td>
</tr>
<tr>
<td>39</td>
<td>0.05</td>
<td>1.90</td>
<td>35.8 [18.88,34.56]</td>
<td>76.9</td>
<td>80.8</td>
<td>79.7 [79.84,87.20]</td>
</tr>
<tr>
<td>50</td>
<td>0.05</td>
<td>2.45</td>
<td>32.8 [25.28,35.84]</td>
<td>86.0</td>
<td>88.8</td>
<td>88.0 [85.60,92.00]</td>
</tr>
<tr>
<td>200</td>
<td>0.025</td>
<td>4.98</td>
<td>28.5 [26.88,28.80]</td>
<td>96.0</td>
<td>98.8</td>
<td>98.0 [97.76,99.68]</td>
</tr>
<tr>
<td>999</td>
<td>0.01</td>
<td>9.98</td>
<td>28.8 [28.80,29.12]</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0 [99.68,100.00]</td>
</tr>
</tbody>
</table>

In Table 1 we list the results of experiments with simulated data with region-specific Dirichlet parameters, and global inference. It is clear that an increasing mode, also yields an increase in accuracy. As the credible intervals show, for $k = 999$ and $\theta = 0.01$, the posterior densities and accuracies are very concentrated around their respective modes. At the same time, a higher mode of the prior also reduces the number of samples that is accepted for the adjacency matrix. In the last case, only 3 out of 7.5M proposals were accepted.

As said earlier, the streamlines are too “good”, and in this experiment even to such an extent that the model does not have any doubts that the incoming count matrix is correct. Each edge in the count matrix is then represented by a streamline count. The prior on $d_1$ is so strong, that if there is a streamline count $c_{ij} > 0$, then there must also be an underlying edge $a_{ij}$.

Not only are the densities and accuracies converging, the credible intervals also decrease with a stronger prior on the Dirichlet parameters (see Figure 1a). This also indicates that with a stronger prior there is less uncertainty. We will discuss the results further in the next section.

However, it is not only the mode that influences the posterior performance such as density and accuracy, also the width of the distribution, as we show with the following example in Table 2. We use the same experimental setup as the ceiling experiment described earlier. Now, we fix the mode, and vary the prior on the Dirichlet parameters, such that the width changes. The width of the interval which contains 95% of the sampled parameters is denoted by column $w$ in Table 2. As the width of the distribution increases, the distribution flattens (this is illustrated in Figure 1b).

The reported values seem to suggest that with an increasing width, the accuracy seems to rise as well. The sensitivity is comparable, but the specificity is higher. This in turn means that the inferred matrix is sparser, or put stronger, populated with more absent edges (which seems to hold according to the reported densities). Of these absent edges most seem to be correct as well (specificity). With a width greater than 4, the same ceiling effect seems to appear as in the previous experiments.
(a) Graphical representation of the first experiment. In the topmost image the prior densities over the different $d$ are plotted. The left-most black distribution is the prior on $d_0$. The other distributions are from left to right the distribution from top to bottom as in Table 1. In the middle image, the posterior distribution of $d_1$ parameters is plotted. The colours of the lines are the same as for the distributions in the upper image. In the bottom image the posterior values of the accuracy and density are plotted with error bars denoting the range of the 95% credible interval.

(b) Graphical representation of the different Gamma distributions denoted in Table 2. Each Gamma distribution’s mode is at 1.44. The 95% credible interval of the respective Gamma distributions are denoted with the horizontal bars.

Figure 1: Figures for the streamline quality experiments.
Table 2: In this experiment we gradually lowered the peak of the distribution, or rather, we choose a flatter distribution. The goal is to see whether this leads to the same unwanted behaviour as in the mode experiment. The experiment was done with data generated with region-specific Dirichlet parameters and global inference (RsimGinf). The width of the distribution is the width of the interval between 2.5% and 97.5% based on 100000 random samples. The reported values are macro-averages over the micro-statistics from 3 runs.

<table>
<thead>
<tr>
<th>k</th>
<th>θ</th>
<th>d₁</th>
<th>w</th>
<th>density (95% CI)</th>
<th>sensitivity</th>
<th>specificity</th>
<th>mean accuracy (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>0.06</td>
<td>1.44</td>
<td>1.16</td>
<td>37.8 [17.92, 64.00]</td>
<td>65.2</td>
<td>73.3</td>
<td>70.9 [58.72, 79.52]</td>
</tr>
<tr>
<td>13</td>
<td>0.12</td>
<td>1.44</td>
<td>1.70</td>
<td>33.2 [11.52, 52.48]</td>
<td>63.0</td>
<td>78.8</td>
<td>73.3 [63.52, 81.12]</td>
</tr>
<tr>
<td>7</td>
<td>0.24</td>
<td>1.44</td>
<td>2.46</td>
<td>40.6 [10.88, 56.32]</td>
<td>71.9</td>
<td>72.1</td>
<td>72.1 [60.96, 79.84]</td>
</tr>
<tr>
<td>4</td>
<td>0.48</td>
<td>1.44</td>
<td>3.64</td>
<td>34.9 [20.48, 48.96]</td>
<td>65.9</td>
<td>77.7</td>
<td>74.3 [64.48, 82.08]</td>
</tr>
<tr>
<td>2.5</td>
<td>0.96</td>
<td>1.44</td>
<td>5.67</td>
<td>24.1 [14.40, 36.48]</td>
<td>56.9</td>
<td>89.1</td>
<td>79.9 [73.12, 84.64]</td>
</tr>
<tr>
<td>1.75</td>
<td>1.92</td>
<td>1.44</td>
<td>9.68</td>
<td>24.5 [17.60, 30.08]</td>
<td>70.5</td>
<td>94.1</td>
<td>87.3 [79.20, 93.28]</td>
</tr>
<tr>
<td>1.375</td>
<td>3.84</td>
<td>1.44</td>
<td>16.60</td>
<td>22.6 [12.48, 31.04]</td>
<td>65.1</td>
<td>94.6</td>
<td>86.1 [82.08, 89.44]</td>
</tr>
</tbody>
</table>

3.4 Comparing global and regional inference

In the previous experiments, we used a streamline count matrix generated with region-specific Dirichlet parameters, and did inference according to the global model, in order to set a baseline. If we repeat the first experiment, but now with only 5000 steps (of which 1000 are used as burnin), we get the results reported in Table 3.

Table 3: Similar experiment as described in Table 2, but now with regional-based inference (RsimRinf). The reported values are macro-averages over the micro-statistics from 3 runs, with each run being 5000 steps, of which 1000 burnin. The underlined values denote improvements over the global model (not statistically significant).

<table>
<thead>
<tr>
<th>k</th>
<th>θ</th>
<th>d₁</th>
<th>w</th>
<th>density (95% CI)</th>
<th>sensitivity</th>
<th>specificity</th>
<th>mean accuracy (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>0.06</td>
<td>1.44</td>
<td>1.16</td>
<td>27.0 [12.48, 38.08]</td>
<td>54.9</td>
<td>84.3</td>
<td>75.8 [70.56, 80.80]</td>
</tr>
<tr>
<td>13</td>
<td>0.12</td>
<td>1.44</td>
<td>1.70</td>
<td>18.6 [7.04, 30.08]</td>
<td>40.0</td>
<td>90.0</td>
<td>75.6 [71.84, 79.52]</td>
</tr>
<tr>
<td>7</td>
<td>0.24</td>
<td>1.44</td>
<td>2.46</td>
<td>35.3 [3.84, 56.32]</td>
<td>58.5</td>
<td>74.0</td>
<td>69.6 [60.64, 78.24]</td>
</tr>
<tr>
<td>4</td>
<td>0.48</td>
<td>1.44</td>
<td>3.64</td>
<td>25.1 [8.32, 41.92]</td>
<td>54.0</td>
<td>86.6</td>
<td>77.2 [71.52, 82.40]</td>
</tr>
<tr>
<td>2.5</td>
<td>0.96</td>
<td>1.44</td>
<td>5.67</td>
<td>14.3 [1.92, 26.24]</td>
<td>37.9</td>
<td>95.3</td>
<td>78.8 [71.52, 86.24]</td>
</tr>
<tr>
<td>1.75</td>
<td>1.92</td>
<td>1.44</td>
<td>9.68</td>
<td>14.4 [2.88, 25.28]</td>
<td>40.2</td>
<td>96.1</td>
<td>80.0 [73.12, 86.24]</td>
</tr>
<tr>
<td>1.375</td>
<td>3.84</td>
<td>1.44</td>
<td>16.60</td>
<td>17.3 [3.20, 25.28]</td>
<td>49.5</td>
<td>95.8</td>
<td>82.5 [73.12, 88.16]</td>
</tr>
</tbody>
</table>

If we consider the density, regional inference structurally estimates a sparser network. This is indirectly reflected in the specificity scores, which are better than for global inference. However, with sensitivity scores hovering around or even below random guessing, it also confirms that the edges that were predicted are not of good quality. Therefore, although the reported accuracy is slightly better than the global model in two cases (but not statistically significant), regional inference generally performs worse than global inference. In the remainder of this section we will do more experiments to compare between the two models, and try to find a regime where regional inference qualitatively outperforms global inference, which is our baseline.

3.5 Model comparison

For the main experiments (as introduced in Section 3.1) we used a small (25 nodes) adjacency matrix. With 25 nodes, sparseness of edges is not a problem (whereas this is for example for 10 nodes\(^2\)), but we gain the

\(^2\)A 10 \times 10 matrix has 45 unique nodes (due to symmetry), and with a density of 0.3 this gives only around 14 edges.
advantage of running the experiments in reasonable time. We also experimented with different matrix sizes \((k = 50\) and \(k = 75\)), yielding comparable results. Later, we will report experiments on real life data with over 100 nodes (Section 3.7). All experiments are run on the same ground truth adjacency matrix, with a desired density of 30\% \((\alpha = 4, \beta = 8, 28.80\% \text{ in practice})\). The gamma distributions for the Dirichlet parameters are described by two gamma distributions: \(\text{Gamma}(k_1, \theta_1)\) with \(k_1 = 2.5\) and \(\theta_1 = 0.96\) and \(\text{Gamma}(k_0, \theta_0)\) with \(k_0 = 5\) and \(\theta_0 = 0.12\). The choice of these parameters is based on earlier experiments in this section, such that the quality of streamlines is not too good. Furthermore, we draw a maximum of 50,000 outgoing streamlines per region. The results are reported in Table 4.

Table 4: The four main experiments as described in Section 3.1. The reported values are macro-averages over the micro-statistics from 3 runs, with each run being 5000 steps, of which 1000 burnin. The density of \(\text{RsimRinf}\) is consistently low, this is not coincidental, but it is not clear why.

<table>
<thead>
<tr>
<th>model</th>
<th>density (95% CI)</th>
<th>sensitivity</th>
<th>specificity</th>
<th>mean accuracy (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GsimGinf</td>
<td>28.5 [26.24,30.40]</td>
<td>94.7</td>
<td>98.3</td>
<td>97.3 [95.20,99.04]</td>
</tr>
<tr>
<td>GsimRinf</td>
<td>30.9 [19.52,44.48]</td>
<td>76.3</td>
<td>87.4</td>
<td>84.2 [78.24,90.40]</td>
</tr>
<tr>
<td>RsimGinf</td>
<td>29.4 [14.72,43.52]</td>
<td>57.2</td>
<td>82.0</td>
<td>74.9 [66.72,81.12]</td>
</tr>
<tr>
<td>RsimRinf</td>
<td>14.5 [1.92,26.88]</td>
<td>34.3</td>
<td>93.5</td>
<td>76.4 [71.20,81.44]</td>
</tr>
</tbody>
</table>

From these results we can conclude that global inference outperforms regional inference irrespective of whether the data was generated with global- or region-specific Dirichlet parameters. The small improvement on the latter with regional inference, is overshadowed by its low density, as we know many connections between brain areas are missing.

We conclude this section with two experiments to see if we can force the regional model to outperform, in specific situations, where we expect the global model to fail.

### 3.6 Out-of-prior experiment

In the previous section we saw that global inference could still outperform region-based inference, even when the streamline counts were generated with the regional-based approach where each region has its own set of Dirichlet parameters. In this section we perform some additional experiments to see whether this also holds in situations where we assume that we do not know from which Gamma distributions the Dirichlet parameters are sampled.

To this end, we distinguish four scenarios. For scenarios 1, 2, and 4, we use a two-valued approach where \(d \in \{ \delta_1, \delta_2 \}\). In scenario 1, \(\delta_1\) is the value on 2.5\% interval boundary of \(d\)’s Gamma distribution, and \(\delta_2\) is on the 97.5\% interval boundary Gamma distribution. In scenario 2, \(\delta_1\) is the same, and \(\delta_2\) is the mode of the Gamma distribution. Scenario 4 is similar with \(\delta_1\) being the mode, and \(\delta_2\) the 97.5\% interval boundary. The idea behind these scenarios is that global inference does not have as many degrees of freedom as region-based inference. We named these experiments out-of-prior experiments because although the values fall within the rangle of possible values from the prior distribution, they do not follow the distributional properties of the prior distribution. Now that there are only 2 different values, region-based inference should be advantageous, as it can learn a Dirichlet parameter per region. The global model can only learn 1 set of global hyperparameter, and is expected to perform worse. Scenario 3 uses different Gamma distributions (with the same mode) for generating data and sampling: a peaked distribution with \(k_p = 25\) and \(\theta_p = 0.06\), and a flatter distribution with \(k_f = 1.75\) and \(\theta_f = 1.92\). With this experiment we test the performance of a situation where there is a (slightly) wrong prior on the Dirichlet parameters, and the real values are outside the estimated prior. In all four experiments, and invariant of the model (global and regional) used, the values of \(d_0\) are the same; the values of \(d_1\) are the same within a model.

In Figure 2 we can see the differences between the inferred and the expected Dirichlet parameters for \(d_0\) and \(d_1\) per node for each of the four scenarios. Generally, for \(d_0\), the sampled values are quite accurate. The only exception is for node 17 and 18, because the inferred (mean) value is 0.3876, whereas the \(d_0^{17} = 1.0859\)
Figure 2: Differences between sampled and true Dirichlet parameters. The heatmaps depict the differences between the expected and the sampled values for \( d_0 \) and \( d_1 \) per node. G stands for global sampling, R stands for regional sampling. The last digit in the model name represents the strategy. Strategy 1 picks either of two values for \( d_1 \), namely the values that are on the boundary of the 95% interval of the Gamma distribution. We used \( k_0 = 5 \) and \( \theta_0 = 0.12 \) as hyperparameters for the prior on \( d_0 \). For \( d_1 \) we used \( k_1 = 1.75 \) and \( \theta_1 = 1.92 \), resulting in a mode of 1.44, and the 95% intervals are on \([0.3295, 9.8951]\) (see the black line in Figure 1b). Strategy 2 uses two values as well: the 2.5% interval and the mode. Strategy 4 is similar, but uses the mode and the 97.5% interval. Strategy 3 uses a more peaked Gamma distribution \((k_1 = 25, \theta_1 = 0.06)\) to simulate the data, and the flatter distribution \((k_1 = 1.75, \theta_1 = 1.92)\) for sampling.

and \( d_0^{18} = 0.9664 \). This is inherent to global inference, since it lacks the flexibility of fitting to the nodes. However, we see the same behaviour for the regional models. It overestimates less, but it underestimates a bit more. Overall, the \( d_0 \) parameters are estimated quite well, there are not many differences between the methods, and the region of variation is reasonable given the expected mode of 0.48, and absolute differences being mainly around 0.2.

Table 5: The results for the out-of-prior experiments. See Figure 2 and the text for descriptions of the settings.

<table>
<thead>
<tr>
<th>model</th>
<th>setting</th>
<th>density (95% CI)</th>
<th>sensitivity</th>
<th>specificity</th>
<th>mean accuracy (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Global</td>
<td>1</td>
<td>25.8 [23.68,28.16]</td>
<td>83.9</td>
<td>97.7</td>
<td>93.7 [92.00,85.52]</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>29.0 [25.92,32.00]</td>
<td>82.2</td>
<td>92.5</td>
<td>89.6 [87.20,91.68]</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>16.2 [13.12,19.52]</td>
<td>48.2</td>
<td>96.8</td>
<td>82.8 [80.48,84.96]</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>15.5 [12.36,18.48]</td>
<td>21.9</td>
<td>87.1</td>
<td>68.3 [56.16,73.44]</td>
</tr>
<tr>
<td>Regional</td>
<td>1</td>
<td>26.7 [22.72,30.72]</td>
<td>77.2</td>
<td>93.7</td>
<td>88.9 [84.96,92.32]</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>27.6 [23.04,32.32]</td>
<td>72.8</td>
<td>90.7</td>
<td>85.6 [82.40,88.80]</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>16.3 [11.84,19.84]</td>
<td>44.3</td>
<td>95.0</td>
<td>80.4 [76.64,83.36]</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>12.2 [7.36,17.92]</td>
<td>21.2</td>
<td>91.4</td>
<td>71.2 [68.32,73.76]</td>
</tr>
</tbody>
</table>

In the same light as the previous experiments, we also look at the posterior metrics for the inferred adjacency matrices. For \( d_1 \), the results are different. Model 3 and 4 exhibit similar behaviour for estimating the Dirichlet parameters, as are the results for models 1 and 2. However, the similarity is only partly reflected in the matrix metric density. For sensitivity, specificity and accuracy, this similarity does not hold. Besides the outlier in \( d_1^{0} \) and \( d_1^{1} \), the global and regional are comparable. Both in distances to the expected Dirichlet parameter, as well as in the matrix metrics.

Although model 3 and 4 appear to be completely different models, they likeness in behaviour can be explained. Both models do not use low-sampled Dirichlet parameters. Combined with a decent mode (see Table 3), the model can estimate the Dirichlet parameters reasonably well. However, in terms of posterior density and accuracy, the models cannot compete with models 1 and 2.
Figure 3: Applying the models on empirical streamline counts. The streamlining data (Hinne et al., 2013) comprises 116 parcelated regions (AAL, Tzourio-Mazoyer et al., 2002), and has been acquired with probabilistic tractography. We used the same parameters as for other runs in this paper: a prior density of 0.3, $k_1 = 2.5$, $\theta_1 = 0.96$, $k_0 = 5$, and $\theta_0 = 0.12$. On the left we plotted the matrix with edge probabilities for the global model for 4000 sweeps (after 1000 burnin steps). The centre figure shows the edge probabilities for the regional model. On the right, we show the differences with respect to the regional model. A blue-marked edge indicates that the edge was not found with the regional model; a red-marked edge indicates that the edge was not found by the global model.

Table 6: The (bidirectional) connections between brain regions that only have been found by the global model, and vice versa. The table lists the top 10 of the connections ordered by their edge probabilities (in %). Whether these connections are biologically plausible, has yet to be determined in future research.

<table>
<thead>
<tr>
<th>prob. connections found only with regional</th>
<th>prob. connections found only with global</th>
</tr>
</thead>
<tbody>
<tr>
<td>85.5 Hippocampus_L - Lingual_R</td>
<td>90.4 Frontal_Sup_L - Precentral_L</td>
</tr>
<tr>
<td>76.1 Cingulum_Ant_R - Olfactory_R</td>
<td>67.3 Cerebelum_6_L - Lingual_L</td>
</tr>
<tr>
<td>63.2 Precuneus_R - ParaHippocampal_R</td>
<td>59.0 Paracentral_Lobule_L - Precentral_L</td>
</tr>
<tr>
<td>52.0 Temporal_Sup_L - Rolandic_Oper_L</td>
<td>56.2 Cerebelum_10_R - Cerebelum_6_R</td>
</tr>
<tr>
<td>51.0 Temporal_Mid_R - Temporal_Pole_Sup_R</td>
<td>56.0 Cerebelum_4_5_L - Lingual_L</td>
</tr>
<tr>
<td>49.5 Temporal_Mid_L - SupraMarginal_L</td>
<td>48.7 Occipital_Mid_L - Calcarine_L</td>
</tr>
<tr>
<td>48.8 Frontal_Med_Orb_L - Frontal_Sup_Orb_L</td>
<td>47.4 Temporal_Pole_Mid_L - Temporal_Mid_L</td>
</tr>
<tr>
<td>48.7 Cerebelum_6_R - Cerebelum_3_R</td>
<td>46.9 Temporal_Mid_L - Occipital_Mid_L</td>
</tr>
<tr>
<td>48.1 Vermis_7 - Cerebelum_Crus1_R</td>
<td>46.0 Cerebelum_Crus1_L - Lingual_L</td>
</tr>
<tr>
<td>45.0 Putamen_L - Caudate_L</td>
<td>37.1 Caudate_R - Frontal_Mid_R</td>
</tr>
</tbody>
</table>

3.7 Estimating empirical data

We used simulated data throughout the paper for two reasons. First, to reduce the time-complexity, as we were free to use smaller matrices. Second and most important is that for empirical data, we do not have a true underlying structure. Otherwise we would not have to estimate this structure.

To show that the regional model can also be applied in a real situation, we run both the global and regional model on parcellated (AAL, Tzourio-Mazoyer et al., 2002) empirical probabilistic tractography data. In Figure 3 we show the results and the differences between the two models.

A blue-marked edge indicates that the edge was only found in the global model, likewise, a red-marked edge indicates that the edges was only found in the regional model. In Table 6 we list for both models the connections found uniquely by either of the models, together with their respective edge probability. However, further research is needed to verify whether these connections are biologically plausible.

We also looked at the symmetrical properties of the data, following from the fact that there are two hemispheres in the brain, left and right. The average edge probabilities are similar: for the global model this is 3.6% and for the regional model 3.7%. Within the hemispheres (based on the first 45 left/right labelled
Figure 4: The distribution of edge probabilities for the different models on empirical data. The y-axis is
cutoff at 400 the keep the other bars visible. The original values are indicated in the plots. The distributions
are taken from the $116 \times 116$ matrix.

Figure 5: Side by side comparison of edges within either left and right hemisphere for global and regional
sampling. We only used the first 90 nodes of AAL (45 for left, and 45 for right), and only looked at edges
within one hemisphere.

nodes in AAL) the averages are also comparable: 9.6%, 9.6% for the global model and 9.7% and 9.5%
regional model, respectively left and right. If we take into account the distribution of edge probabilities as
depicted in Figure 4, then the global model shows more confident behaviour: the centre of the distribution
shows mainly low probabilities. Although the global model seem to prefer sparser networks than the regional
model, the number of edges with a edge probability around zero is almost the same for both models.

We used the brain regions that clearly belong to one of those hemispheres to generate the figures in
Figure 5, and looked at the connections within either hemisphere. One would expect that if two regions are
connected in the left hemisphere, their counterparts in the right hemisphere are also connected. We used a
weighted Jaccard similarity

$$J(X, Y) = \begin{cases} \frac{\sum_{i=1}^{n} \min(X_i, Y_i)}{\sum_{i=1}^{n} \max(X_i, Y_i)} & \text{if } \sum_{i=1}^{n} \max(X_i, Y_i) > 0 \\ 1 & \text{if } \sum_{i=1}^{n} \max(X_i, Y_i) = 0 \end{cases}$$

(35)

where $X$ and $Y$ are non-negative $n$-dimensional real vectors (Chierichetti et al., 2010). We are mainly
interested in how many edges from the left hemisphere are mirrored in the right hemisphere, and vice versa.
Since generally the brain structures are taken to be symmetrical, we expect to see the same pattern in the
results. There are 3 scenarios that we consider. First, how many connections from region $i$ in the left
hemisphere go to region $i$ in the right hemisphere? Second, how many connections from region $i$ in the left
hemisphere go to region $j$ in the right hemisphere, while region $i$ in the right hemisphere has a connection
with $j$ in the left hemisphere (crosslinked)? And last, how symmetric are the connections from $i$ to $j$ in the
Figure 6: We divided the full edge probability matrix of 116 parcellated regions into two parts, where each part is about either the left or the right hemisphere. We used the labels from AAL, and limited to the brain regions for which it makes sense to distinguish between left and right (the first 90 nodes from AAL). In the first scenario we looked at direct connections between homologous brain regions. In the second scenario we looked at symmetric crosslinks between homologous brain regions. In the third scenario we looked at the parallel connections between brain regions, which indicates how symmetric the brain structure is. In scenario 3 we plotted the weighted Jaccard similarity (Equation (35)) against the cutoff (marked with an ‘x’).

left hemisphere in the right hemisphere from $i$ to $j$, and how many of these connections are present? We plotted the results in Figure 6.

From the left figure we see that there are few connections between the left and the right hemispheres. The global model finds 2 connections (Supp_Motor_Area and Frontal_Sup_Medial) with edge probability 1.0 to connect two homologous regions. Regional finds more symmetric connections, but with less confidence, as with an increasing cutoff, the number of connections decreases. We observe the same behaviour in scenario 2, which is a superset of scenario 1. Again the regional model finds more symmetric crosslinks, but generally with less confidence. Scenario 3 is about the symmetry not ‘between’, but within, the brain hemispheres. The global model finds 52 connections with probability 1.0 in the left hemisphere which can also be found in the right hemisphere, whereas the regional model finds none at that certainty level. If we consider the similarity of the edge probability matrix between the left and right hemisphere, we can see that the models perform comparable until a probability around 0.75. In all three scenarios we already saw that cutoff at a higher probability, the global model is better than regional model. This is also reflected in the similarity score. Although these results might indicate that the global model is only slightly better, the figures show that beyond chance level, the global model structurally finds more symmetric edges and therefore reaches better results than the regional model.

4 Discussion and Conclusion

In this section we reflect on the work, discuss the results, and conclude the paper with a conclusion.

4.1 Discussion

The transition from a model with fixed parameters to a global and regional model seems a logical step. The global model allowed for such an extension, and intuitively we expected that the streamlining properties are not necessarily global, as regions may have different roles and functions. The global model fails to address these issues, and therefore, we expected that overcoming this globalness by introducing region-based interactions, these issues could be overcome. However, the current results do not show the expected
behaviour. In this section we will discuss the results.

Estimating the Dirichlet parameters is only one aspect of the model. In the previous section we also saw that even in situations that were expected to be difficult for the global model, it still outperforms the regional model. However, in the end, we are not as much interested in estimating the Dirichlet parameters, but more in the inferred structural brain network.

Initially, we started with priors on the Dirichlet parameters that were clearly separated and peaked. As shown in Section 3.3, this led to a ceiling effect, in which it was difficult to interpret the results. With an average accuracy over 99.99% with the global model, it would be hard to get closer to 100%. However, we also knew from previous experiments, and from the original paper by Hinne et al., 2013, that the problem of inferring structure in brain networks was not this easy. Therefore we performed the experiments to assess the influence of the mode of the prior for \(d_1\) with respect to a ceiling. We found (see Table 1) that depending on the mode of the prior, the accuracy would range from almost random guessing, to almost 100% correct. In the first case, the problem was too difficult. Because the values of \(d_1\) were generally smaller than \(d_0\), there was no discriminative element in the values. If any, the roles would be reversed, where \(d_1\) would take the role of \(d_0\), and vice versa. For the inference this does not matter, however from the perspective of the original model where we want to analyse structural brain networks, it does not make sense. It is not likely that regions which are not connected with each other, have a higher chance of streamlining than two regions which are connected.

A similar phenomenon is reported for the wide priors in Table 2 where we changed the width of the distribution, but kept the mode of the prior fixed. Where the width of the distributions in the previous experiment was around 2, we now increased up to above 16. Apparently the influence was stronger for \(d_1\), as we see the same behaviour for specificity as in the varying mode experiment. We do not see the same rise for sensitivity, which means that the lower values in \(d_1\) still have some influence, but not enough to stop the specificity from converging towards 100%. The same behaviour for the accuracy is also visible here, but it is slowed down by the slower growth of the sensitivity.

These two experiments look at the quality of streamlines, and they show that although it is not critical which priors are chosen, it is important to stay away from the cases in which there is no room for the model to learn the underlying structures. In Table 3 we showed the same width experiment, but now with regional inference. The estimated networks are sparser. The reported specificity values are a bit better due to the lower density, the sensitivity values are more alarming, as they are close to random guessing. We are not only interested in the accuracy, but more in an accurate estimation of the original network. With the reported densities, only about half the number of edges are estimated. Hence, the regional model does not perform better than the global model, even though the data used in the experiments is generated with region-specific Dirichlet parameters.

We picked a prior for \(d_1\) that was neither causing the inferring of the underlying structure to be a too easy problem nor a difficult problem, and a reasonable width: Gamma\((k = 2.5, \theta = 0.96)\). With this prior we look at the four possible combinations of data generating and inference, with both regional and global. It turns out (see Table 4) that the global model models a strong method, and also performs well on data generated with region-specific parameters. The posterior accuracy of regional inference is slightly better, but with the penalty that the density is only half of the expected density. One of the reasons that global inference performs better is because it has more data to train on. The regional model only uses data regarding a specific region, whereas the global model can use all data about all regions. Although the density of the networks used in this paper is around 0.3, this does not mean that this holds for each region as well. In regions with sparsely outgoing edges, the density of a graph is easily underestimated. However, this does not explain the difference between the performance on globally generated data and regionally generated data, as the first is a subset of the latter.

In the last experiments, the out-of-prior experiments, we try to force certain regimes in which we would expect regional inference to outperform global inference. The main goal of scenarios 1, 2, and 4, was to see what would happen if the global model was forced to choose between either values \(\delta_1\) and \(\delta_2\). As we can see in the right figure of Figure 2, the global model does not fit to either of the two values, but fits to the some value in the middle. The same holds for the regional model, which is peculiar, since the regional model has
the flexibility to adapt per region to the given \(d_1\) value. In general, the regional model performs comparable to the global model, albeit a bit worse. Why the first region is structurally underestimated is not known, since the sampling process is randomised, hence the first region is not necessarily the first region for which a sample is proposed.

Based on these experiments, we were not able to find a regime in which the regional model structurally outperforms the global model. Although it may be the case that the added flexibility, by adding region-specific parameters, does not help inferring the underlying structure. In most cases it does not perform unusably poor, however, with the disadvantage of added time complexity, it does not seem to be an improvement over the global model. On the other hand, most experiments performed in this paper are with a fixed set of parameters. Based on the ceiling and width experiment (Section 3.3) we chose a set of parameters that seemed reasonable. They provided room to either improve or make mistakes (by avoiding the ceiling effect), but also seemed to have a plausible width of the prior distribution. But by picking a fixed set of parameters, we also greatly reduce the area in which the presented results are valid. To this end, and this was implicit throughout the paper, the results are only valid under the conditions given by the chosen parameters. We did some small experiments and saw that the claims would also hold under different circumstances. However, we did not and cannot exhaustively test with every set of parameters.

From these experiments with simulated data we saw that the global model, with less parameters, is still flexible enough to capture the underlying graph structure better than the regional model, which has more parameters and thus more flexibility. In the last section we presented the results of both models run on empirical data. There we see some interesting things. First, both models have similar average edge probabilities, although the global model distributes these probabilities more towards the boundaries 0 and 1, whereas the regional model has a flatter probability distribution. Furthermore, if we divide the parcellated data into regions in the left and right brain hemisphere we find that the regional model finds more symmetric patterns in the brain. However, the regional model is generally less confident of its estimations, to a degree where beyond chance level on the edge probability, the global model is better.

Unfortunately, we were not able to show that global inference can be improved upon by regional inference, despite the intuition and expectations. This probably means as well that even another extension of node-based inference, is not likely to work as well. For future research we recommend to explore other sets of parameters for the hyperpriors or to find a way to use more data than only region-specific data, for example by grouping regions.

4.2 Conclusion
In this paper we introduced an extension to a Bayesian framework for the inference of brain structures. We first added a global hyperprior on the Dirichlet parameters which are used as latent variables in the forward model, which we then enriched with region-specific parameters. However, the added flexibility of region-specific Dirichlet parameters is not needed for inference of the underlying graph structure. Apparently the probability distribution obtained when integrating out the global Dirichlet parameters is still sufficiently flexible to represent data even when generated from a model with region-specific Dirichlet parameters.

References


Stam, Cornelis J and Jaap C Reijneveld (2007). “Graph theoretical analysis of complex networks in the brain”. In: *Nonlinear biomedical physics* 1.1, p. 3.
A Worked-out derivations

\[ T_1' = \log P(c_i | b_i, a_i', d_0', d_1') - \log P(c_i | b_i, a_i, d_0, d_1) \]  
\[ = \log \text{DCM}(c_i | b_i, d_0' (1 - a_i') + d_1' a_i') - \log \text{DCM}(c_i | b_i, d_0' (1 - a_i') + d_1' a_i) \]  
\[ = \log \text{DCM}(c_i | b_i, \delta_i') - \log \text{DCM}(c_i | b_i, \delta_i) \]  
\[ = \log \left( \frac{b_i! \Gamma(c_i') \prod_j \Gamma(e_{ij} + \delta_{ij}')}{\prod_j c_{ij}! \Gamma(b_i + c_i') \prod_j \Gamma(e_{ij} + \delta_{ij}')} \right) - \log \left( \frac{b_i! \Gamma(c_i) \prod_j \Gamma(e_{ij} + \delta_{ij})}{\prod_j c_{ij}! \Gamma(b_i + c_i) \prod_j \Gamma(e_{ij} + \delta_{ij})} \right) \]  
\[ = \log \frac{b_i!}{\prod_j c_{ij}!} + \log \frac{\Gamma(c_i')}{\Gamma(b_i + c_i')} + \log \prod_j \frac{\Gamma(e_{ij} + \delta_{ij}')}{\Gamma(e_{ij} + \delta_{ij})} - \log \frac{b_i!}{\prod_j c_{ij}!} + \log \frac{\Gamma(c_i)}{\Gamma(b_i + c_i)} + \log \prod_j \frac{\Gamma(e_{ij} + \delta_{ij})}{\Gamma(e_{ij} + \delta_{ij})} \]  
\[ = \log \Gamma(\sum_j c_{ij}!') - \log \Gamma(\sum_j c_{ij}! + \sum_j \delta_{ij}') - \sum_j \log \Gamma(\sum_j c_{ij} + \sum_j \delta_{ij}) + \sum_j \log \Gamma(\sum_j c_{ij} + \sum_j \delta_{ij}) - \sum_j \log \Gamma(\sum_j c_{ij}! + \sum_j \delta_{ij}') \]  
\[ = - \sum_j \log \Gamma(\delta_{ij}') + \log \frac{\Gamma(\sum_j c_{ij}!')}{\Gamma(\sum_j \delta_{ij}')} + \sum_j \log \frac{\Gamma(c_{ij} + \delta_{ij}')}{\Gamma(c_{ij} + \delta_{ij})} - \log \frac{\Gamma(\sum_j c_{ij}! + \sum_j \delta_{ij})}{\Gamma(\sum_j c_{ij} + \sum_j \delta_{ij})} \]  
\[ T_3' = \log P(d_0' | k_0, \theta_0) - \log P(d_0' | k_0, \theta_0) \]  
\[ = \log \text{Gamma}(d_0' | k_0, \theta_0) - \log \text{Gamma}(d_0' | k_0, \theta_0) \]  
\[ = \log \left( \frac{1}{\theta_0^{k_0} \Gamma(k_0)} d_0^{k_0 - 1} \exp \left( -\frac{d_0'}{\theta_0} \right) \right) - \log \left( \frac{1}{\theta_0^{k_0} \Gamma(k_0)} d_0^{k_0 - 1} \exp \left( -\frac{d_0'}{\theta_0} \right) \right) \]  
\[ = \log \frac{1}{\theta_0^{k_0} \Gamma(k_0)} + \log d_0^{k_0 - 1} + \log \exp \left( -\frac{d_0'}{\theta_0} \right) - \log \frac{1}{\theta_0^{k_0} \Gamma(k_0)} - \log d_0^{k_0 - 1} - \log \exp \left( \frac{d_0'}{\theta_0} \right) \]  
\[ = \log d_0^{k_0 - 1} - \log d_0^{k_0 - 1} - \frac{d_0'}{\theta_0} + \frac{d_0'}{\theta_0} \]  
\[ = (k_0 - 1) \log d_0' - (k_0 - 1) \log d_0' - \frac{d_0'}{\theta_0} + \frac{d_0'}{\theta_0} \]  
\[ = (k_0 - 1) \log \frac{d_0'}{d_0'} + \frac{d_0'}{\theta_0} \]
\[ T_2 = \log P(A'|\alpha, \beta) - \log P(A|\alpha, \beta) \]  
\[ = \log 1_{\{A'|\in M\}} + \log \frac{B(\alpha + \sum_{i<j} a'_{ij}, \beta + \sum_{i<j} (1-a'_{ij}))}{B(\alpha, \beta)} \]  
\[ - \log 1_{\{A|\in M\}} - \log \frac{B(\alpha + \sum_{i<j} a_{ij}, \beta + \sum_{i<j} (1-a_{ij}))}{B(\alpha, \beta)} \]  
\[ = \log \frac{B(\alpha + \sum_{i<j} a'_{ij}, \beta + \sum_{i<j} (1-a'_{ij}))}{B(\alpha + \sum_{i<j} a_{ij}, \beta + \sum_{i<j} (1-a_{ij}))} \]