Dirichlet Process Mixture Models: Application to Brain Image Segmentation

Daniel C. Castro
Department of Computing
Imperial College London
London SW7 2AZ, UK
dc315@imperial.ac.uk

Ben Glocker
Department of Computing
Imperial College London
London SW7 2AZ, UK
b.glocker@imperial.ac.uk

Abstract

The ability of nonparametric models to automatically adapt to the complexity of data makes them particularly suitable for neuroimaging applications, where it is often preferable to avoid assumptions on the correct model structure. We have applied a multivariate Dirichlet process Gaussian mixture model (DPGMM) for segmenting main cerebral tissues (grey matter, white matter and cerebrospinal fluid) by learning from multiple MRI modalities (T1, T2 and PD). We experimentally show that a multivariate DPGMM produced significantly more consistent and accurate segmentations than an equivalent univariate DPGMM trained on a single modality (T1). This is also the first known attempt at performing lesion segmentation with DP mixture models. Our preliminary results show great promise, as the DPGMMs were able to correctly identify most traumatic brain injury lesions in a multimodal MRI dataset (MPRAGE, FLAIR, GE, T2 and PD) and have demonstrated their capability to learn intricate multidimensional probability distributions.

1 Introduction

In medical image segmentation, common approaches are based on supervised learning (e.g. Kamnitsas et al., 2015), atlas-based methods (Cabezas et al., 2011), unsupervised techniques, such as mixture models (e.g. Ashburner and Friston, 2005), or combinations thereof (e.g. Ledig et al., 2012, 2015).

Unsupervised methods are attractive because labelled medical imaging data is relatively scarce, and they essentially attempt to discover structure in the data without any explicit constraints on the output mapping. Furthermore, nonparametric methods make no hard assumptions about this structure, dynamically adapting their complexity as more data is observed (Ghahramani, 2013). In particular, Dirichlet process mixture models (DPMMs) (Antoniak, 1974; Ferguson, 1983) can be seen as an infinite extension of finite mixture models. They place no bound on the number of mixture components, and can, in principle, fit arbitrarily complex distributions. This property makes DPMMs very attractive for modelling elaborate data such as multimodal MRI.

DPMMs have recently been used in a variety of applications in medical imaging. Thirion et al. (2007) and Kim and Smyth (2006) have used DPMMs for modelling spatial activation patterns in functional imaging. Jbabdi et al. (2009) proposed using hierarchical Dirichlet processes for modelling intra- and inter-subject variability in connectivity-based parcellation and Wang et al. (2011) used DPMMs for tractography segmentation. In the context of MRI segmentation, Wachinger and Golland (2014) presented a formulation in which DPMMs were used to model background variation when segmenting small structures to reduce bias. Ferreira da Silva (2007, 2009) performed whole-brain MRI segmentation with a univariate DPMM. More generally, DP-based models have also been applied in computer vision for natural image segmentation (e.g. Torralba et al., 2006; Sudderth and Jordan, 2009).

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

Ferreira da Silva's work was restricted to a single MRI modality (T1) and all the univariate DPGMM components were constrained to share the same variance. In this work, we have addressed *multimodal* brain image segmentation with a multivariate DPGMM, allowing full and untied covariances for all mixture components. Because each imaging modality carries complementary information about the underlying object (in this case, a human brain), we expect that a model capable of harnessing this wealth of information could generate more accurate segmentations.

2 Methods

In this work, we have studied a simple voxel-wise intensity model. Let $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ represent the N voxels of an image, where $\mathbf{x}_i \in \mathbb{R}^D$ are the intensities of voxel i across the D available modalities (assumed perfectly registered). The generative process for the Dirichlet process Gaussian mixture model (DPGMM) – otherwise known as the infinite Gaussian mixture model (Rasmussen, 2000) – used in this work is formulated as follows:[1]

$$\begin{aligned}
\boldsymbol{\pi} &\sim \mathrm{GEM}(\alpha) \\
\boldsymbol{\Lambda}_k &\sim \mathcal{W}(a, \mathbf{B}), & k &= 1, 2, \ldots \\
\boldsymbol{\mu}_k \mid \boldsymbol{\Lambda}_k &\sim \mathcal{N}(\boldsymbol{\mu}_0, (\nu \boldsymbol{\Lambda}_k)^{-1}), & k &= 1, 2, \ldots \\
z_i \mid \boldsymbol{\pi} &\sim \mathrm{Cat}(\boldsymbol{\pi}), & i &= 1, \ldots, N \\
\mathbf{x}_i \mid z_i, \{\boldsymbol{\mu}_k\}_k, \{\boldsymbol{\Lambda}_k\}_k &\sim \mathcal{N}(\boldsymbol{\mu}_{z_i}, \boldsymbol{\Lambda}_{z_i}^{-1}), & i &= 1, \ldots, N,
\end{aligned} \tag{1}$$

where $\mathrm{GEM}(\alpha)$ is the distribution of weights from the one-parameter stick-breaking process, defined as $\pi_k = \beta_k \prod_{l<k}(1 - \beta_l)$, where $\beta_k \sim \mathrm{Beta}(1, \alpha)$ for k from 1 to infinity (Sethuraman, 1994). We consider the parametrisation of the Wishart distribution $\mathcal{W}$ as given in Bernardo and Smith (2000).

In both our experiments, we have set hyperparameters $\nu = 1$ and $a = D + 1$, while $\boldsymbol{\mu}_0$ and $\mathbf{B}$ were set to match the first and second moments of the prior predictive distribution, $p(\mathbf{x})$, to the data. Inference was done in the truncated DPMM variational framework presented in Blei and Jordan (2006), based on which we have derived the parameter updates for the DPGMM (Appendix A).

2.1 Evaluation

To assess the compatibility of the cluster partitioning with the ground-truth labels, we have computed a number of clustering consistency metrics: adjusted Rand index (ARI) (Rand, 1971; Hubert and Arabie, 1985), normalised mutual information (NMI), v-measure (homogeneity, completeness and validity) (Rosenberg and Hirschberg, 2007) and purity.

We are mostly concerned with correctly delineating the structures of interest, regardless of inner subdivisions. To compute spatial accuracy metrics, which require matching labels, we have merged clusters by reassigning their labels to the majority ground-truth label in each cluster. The chosen spatial metrics were the Dice similarity coefficient (DSC) (Dice, 1945), the average symmetric surface distance (ASSD), the Hausdorff distance (HD) and the relative absolute volume difference (RAVD).

3 Results

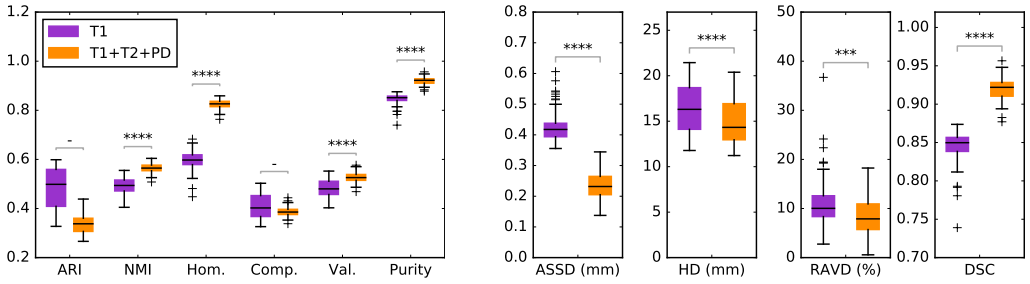
3.1 Normal Tissue Segmentation

In our first experiment, we aimed to assess whether the Dirichlet process Gaussian mixture model could successfully segment grey matter (GM), white matter (WM) and cerebrospinal fluid (CSF) in multimodal images of healthy brains, using a subset of the IXI database.[2] It comprises registered T1- and T2-weighted and proton density (PD) MRI scans of 86 healthy subjects aged 35.8 ± 11.9, with dimensions $256 \times 256 \times 150$ and $.9375 \times .9375 \times 1.2\,\mathrm{mm}^3$ resolution. Since there are no manual segmentations available, the GM, WM and CSF probability maps were computed with SPM12.[3]

[1] We refer readers unfamiliar with Dirichlet processes to the comprehensive overview given in Teh (2010).

[2] `http://brain-development.org/ixi-dataset/`

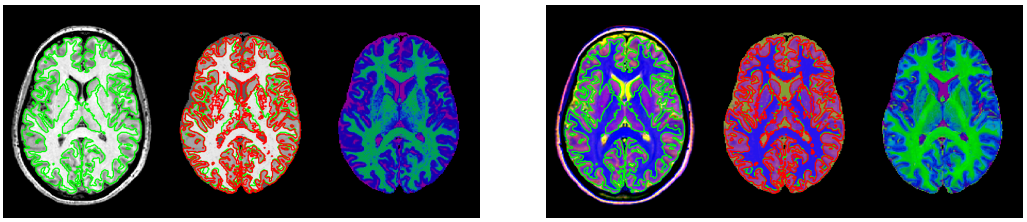
[3] `http://www.fil.ion.ucl.ac.uk/spm/software/spm12/`



(a) Clustering consistency metrics

(b) Spatial accuracy metrics

Figure 1: Results for the normal tissue segmentation experiment. The boxplots indicate quartiles, median, range and outliers. One to four stars indicate $p < 0.05$, 0.01, 10^{-3} and 10^{-4}, respectively, and '–' is insignificant (one-sided Wilcoxon signed-rank test).



(a) T1 slice. $K = 7$ components, $ASSD = 0.49\,\text{mm}$, $HD = 16.0\,\text{mm}$, $RAVD = 5.47\,\%$, $DSC = 0.822$.

(b) PD, T2 and T1 slice (RGB). $K = 12$ components, $ASSD = 0.30\,\text{mm}$, $HD = 16.0\,\text{mm}$, $RAVD = 3.99\,\%$, $DSC = 0.906$.

Figure 2: Uni- and multimodal grey matter (GM) segmentations of IXI subject 52. *Left:* Original image and overlaid reference segmentation (green). *Middle:* Fuzzy image reconstruction from cluster means, overlaid with reference (green) and obtained merged segmentation (red). *Right:* Fuzzy cluster label map. ("Fuzzy" here refers to being weighted by the mixing proportions.)

As a baseline, we fitted a univariate DPGMM to the single T1 image volumes, as was done in the work of Ferreira da Silva (2007, 2009). We then trained multivariate DPGMMs on the combined T1, T2 and PD volumes and compared the results.

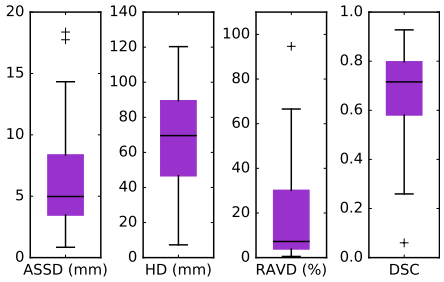
The consistency results are plotted in Fig. 1a. The model found 11.9 ± 0.3 components when using multimodal data and 8.1 ± 1.1 components with a single modality – even though all runs had the same fixed concentration, $\alpha = 0.5$ and truncation level 20. This more fragmented segmentation would explain the observed drop in completeness and in Rand index, because pairs of voxels belonging to the same tissue are more likely to be split across cluster borders.

On the other hand, the multimodal clusters are substantially purer and more homogeneous – which pulls up the validity, counteracting the slight decrease in completeness –, and convey more information about the reference tissue labels. These effects indicate that the availability of the additional modalities allowed the model to infer more specialised clusters.

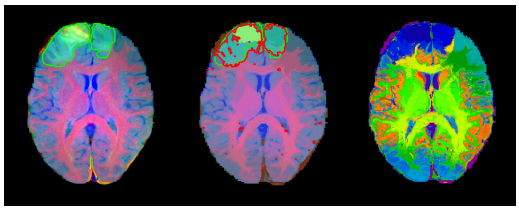
Regarding spatial accuracy (Fig. 1b), the multimodal model provided dramatically better matches to the reference segmentation, as indicated by the average surface distance and Dice coefficient. For each subject, the boundaries of the multimodal clusters follow the tissue boundaries $44\,\%$ more closely and present $8.9\,\%$ higher overlap than in the single modality case, on average. The subtle decrease in Hausdorff distance and volume difference suggests slightly fewer gross mistakes.

3.2 Lesion Discovery

This experiment was intended as a proof of concept – rather than a thorough assessment – of the applicability of DPGMMs to the task of lesion segmentation. As such, it was run on a single axial slice for each subject instead of using the full volumes, which would have taken considerably longer to compute due to the higher dimensionality (five input modalities) and much greater number of components to estimate ($\alpha = 5$ and truncation level 40). This relatively high concentration was



(a) Spatial accuracy metrics



(b) TBI subject 17744. MPRAGE, FLAIR and GE overlaid as RGB. $K = 31$ components, $ASSD = 4.9\,\mathrm{mm}$, $HD = 118\,\mathrm{mm}$, $RAVD = 4.7\,\%$, $DSC = 0.854$.

Figure 3: Lesion segmentation results and one example

chosen for increased sensitivity to subtle intensity variations and potentially very small clusters. The selected slices were the ones containing the most lesion voxels.

We have analysed images of a cohort of 63 moderate-to-severe traumatic brain injury (TBI) patients. The dataset consists of MPRAGE, FLAIR, GE, T2 and PD scans, plus manual lesion segmentations. Each modality has been skull-stripped, affinely registered to MNI space and resampled to $193 \times 229 \times 193$ isotropic $1\,\mathrm{mm}^3$ resolution. This data was originally collected and prepared for Kamnitsas et al. (2016), where further details can be found.

For this task, because there is only one label, we computed only spatial accuracy metrics, plotted in Fig. 3a. Note that the simple majority merging procedure described in Section 2.1 failed to match any lesion for 17 of the 63 subjects, which have been excluded from the results. Interestingly, these subjects were among those with the smallest lesions.

For most cases, the annotated lesions were correctly localised, and the surface distances and volume differences are remarkably low, considering the difficulty of the task. On Fig. 3b we see that the model is able to correctly capture some of the finer inner structure of lesions. Another particularly appealing feature which was observed is that the model has allocated additional clusters to accommodate bias field inhomogeneities, as is clear again in Fig. 3b. Manual annotations involve substantial prior knowledge from the experts. Therefore, in some occasions our intensity-based model clustered together regions which had similar intensity profiles but were not both annotated as lesions, resulting in excessive false positives or false negatives.

4 Conclusion

To the best of the authors' knowledge, this is the first application of Dirichlet process mixture models to multimodal medical image segmentation. Under our experimental conditions, we have observed that leveraging multiple imaging modalities drives the DPMM towards better segmentations than those obtained with a single modality, which was confirmed in terms of both labelling consistency and spatial accuracy.

A further contribution of this work is that this is the first known attempt at applying Dirichlet process mixture models to lesion segmentation in medical images. We have shown that, at least in the 2D case, this nonparametric approach finds a good correspondence to the manual segmentations for large enough lesions. Additionally, we have verified that DPGMMs are able to provide convincing representations of multidimensional probability distributions (see example in Appendix B).

In future work, we will extend the experimental evaluation to a 3D analysis also on other brain pathologies such as brain tumours and stroke lesions. Additionally, as with any purely intensity-based model, we expect that the incorporation of spatial context – such as patch features, atlases or MRF-like constraints – could lead to smoother, more accurate and more robust segmentations. Finally, we believe that DPGMMs will prove to be a useful approach in many other medical imaging applications.

Acknowledgments

This work was supported by the CAPES Foundation, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brazil.

References

Antoniak, C. E. (1974). Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems. *The Annals of Statistics*, 2(6):1152–1174.

Ashburner, J. and Friston, K. J. (2005). Unified Segmentation. *NeuroImage*, 26(3):839–851.

Bernardo, J. M. and Smith, A. F. M. (2000). *Bayesian Theory*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Chichester, US.

Blei, D. M. and Jordan, M. I. (2006). Variational Inference for Dirichlet Process Mixtures. *Bayesian Analysis*, 1(1):121–144.

Cabezas, M., Oliver, A., Lladó, X., Freixenet, J., and Bach Cuadra, M. (2011). A Review of Atlas-Based Segmentation for Magnetic Resonance Brain Images. *Computer Methods and Programs in Biomedicine*, 104(3):e158–e177.

Dice, L. R. (1945). Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26(3):297–302.

Ferguson, T. S. (1983). Bayesian Density Estimation by Mixtures of Normal Distributions. *Recent Advances in Statistics*, 24(1983):287–302.

Ferreira da Silva, A. R. (2007). A Dirichlet Process Mixture Model for Brain MRI Tissue Classification. *Medical Image Analysis*, 11(2):169–182.

Ferreira da Silva, A. R. (2009). Bayesian Mixture Models of Variable Dimension for Image Segmentation. *Computer Methods and Programs in Biomedicine*, 94(1):1–14.

Ghahramani, Z. (2013). Bayesian non-parametrics and the probabilistic approach to modelling. *Philosophical Transactions of the Royal Society A*, 371(20110553).

Hubert, L. and Arabie, P. (1985). Comparing Partitions. *Journal of Classification*, 2(1):193–218.

Jbabdi, S., Woolrich, M. W., and Behrens, T. E. J. (2009). Multiple-Subjects Connectivity-Based Parcellation Using Hierarchical Dirichlet Process Mixture Models. *NeuroImage*, 44(2):373–384.

Kamnitsas, K., Chen, L., Ledig, C., Rueckert, D., and Glocker, B. (2015). Multi-scale 3D convolutional neural networks for lesion segmentation in brain MRI. In *MICCAI Brain Lesion Workshop 2015*, pages 13–16, Munich, Germany.

Kamnitsas, K., Ledig, C., Newcombe, V. F. J., Simpson, J. P., Kane, A. D., Menon, D. K., Rueckert, D., and Glocker, B. (2016). Efficient Multi-Scale 3D CNN with Fully Connected CRF for Accurate Brain Lesion Segmentation.

Kim, S. and Smyth, P. (2006). Hierarchical Dirichlet Processes with Random Effects. *Advances in Neural Information Processing Systems 19 (NIPS 2006)*, pages 697–704.

Ledig, C., Heckemann, R. A., Hammers, A., Lopez, J. C., Newcombe, V. F., Makropoulos, A., Lötjönen, J., Menon, D. K., and Rueckert, D. (2015). Robust whole-brain segmentation: Application to traumatic brain injury. *Medical Image Analysis*, 21(1):40–58.

Ledig, C., Wolz, R., Aljabar, P., Lotjonen, J., Heckemann, R. A., Hammers, A., and Rueckert, D. (2012). Multi-class brain segmentation using atlas propagation and EM-based refinement. In *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 896–899. IEEE.

Rand, W. M. (1971). Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, 66(336):846–850.

Rasmussen, C. E. (2000). The Infinite Gaussian Mixture Model. In *Advances in Neural Information Processing Systems 12 (NIPS 2000)*, pages 554–560. MIT Press.

Rosenberg, A. and Hirschberg, J. (2007). V-Measure: A Conditional Entropy-Based External Cluster Evaluation Measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 410–420, Prague, Czech Republic.

Sethuraman, J. (1994). A Constructive Definition of Dirichlet Priors. *Statistica Sinica*, 4(2):639–650.

Sudderth, E. B. and Jordan, M. I. (2009). Shared Segmentation of Natural Scenes Using Dependent Pitman-Yor Processes. In *Advances in Neural Information Processing Systems 21 (NIPS 2008)*, pages 1585–1592.

Teh, Y. W. (2010). Dirichlet Process. In Sammut, C. and Webb, G. I., editors, *Encyclopedia of Machine Learning*, pages 280–287. Springer US, Boston, MA.

Thirion, B., Tucholka, A., Keller, M., Pinel, P., Roche, A., Mangin, J.-F., and Poline, J.-B. (2007). High Level Group Analysis of FMRI Data Based on Dirichlet Process Mixture Models. In *Information Processing in Medical Imaging - IPMI 2007*, pages 482–494, Berlin, Heidelberg. Springer Berlin Heidelberg.

Torralba, A., Willsky, A. S., Sudderth, E. B., and Freeman, W. T. (2006). Describing Visual Scenes using Transformed Dirichlet Processes. In *Advances in Neural Information Processing Systems 18 (NIPS 2005)*, pages 1297–1304.

Wachinger, C. and Golland, P. (2014). Atlas-Based Under-Segmentation. In *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2014*, pages 315–322. Springer International Publishing.

Wang, X., Grimson, W. E. L., and Westin, C.-F. (2011). Tractography Segmentation Using a Hierarchical Dirichlet Processes Mixture Model. *NeuroImage*, 54(1):290–302.

A Variational Inference

Table 1: True posterior and variational distributions for the multivariate DPGMM

Posterior conditionals	Variational
$p(\mathbf{\Lambda}_k \mid \mathbf{X}, \mathbf{Z}) = \mathcal{W}(\mathbf{\Lambda}_k \mid a'_k, \mathbf{B}'_k)$	$q(\mathbf{\Lambda}_k) = \mathcal{W}(\mathbf{\Lambda}_k \mid c_k, \mathbf{D}_k)$
$p(\boldsymbol{\mu}_k \mid \mathbf{X}, \mathbf{Z}, \mathbf{\Lambda}_k) = \mathcal{N}\big(\boldsymbol{\mu}_k \,\big|\, \boldsymbol{\mu}_{0k}', (\nu'_k \mathbf{\Lambda}_k)^{-1}\big)$	$q(\boldsymbol{\mu}_k) = \mathcal{N}(\boldsymbol{\mu}_k \mid \mathbf{m}_k, \mathbf{\Psi}_k)$
$p(\beta_k \mid \mathbf{Z}) = \mathcal{B}(\beta_k \mid 1 + N_k, \alpha + N_{>k})$	$q(\beta_k) = \mathcal{B}(\beta_k \mid \gamma_{k,1}, \gamma_{k,2})$
$p(z_n = k \mid \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\theta}) \propto \pi_k \mathcal{N}_D\big(\mathbf{x}_n \,\big|\, \boldsymbol{\mu}_k, \mathbf{\Lambda}_k^{-1}\big)$	$q(z_n = k) = \phi_{nk}$

The primed variables are the parameters of the corresponding posterior distributions given the observations assigned to cluster k. In the definitions below, let $\widehat{N}_k = \sum_n \phi_{nk}$ and $\widehat{N}_{>k} = \sum_{l>k} \widehat{N}_l$.

Mean parameters: $q(\boldsymbol{\mu}_k) = \mathcal{N}(\boldsymbol{\mu}_k \mid \mathbf{m}_k, \mathbf{\Psi}_k)$

$$\mathbf{m}_k = \frac{\nu \boldsymbol{\mu}_0 + \sum_n \phi_{nk} \mathbf{x}_n}{\nu + \widehat{N}_k} \tag{2}$$

$$\mathbf{\Psi}_k = \frac{1}{(\nu + \widehat{N}_k) c_k} \mathbf{D}_k \tag{3}$$

Precision parameters: $q(\mathbf{\Lambda}_k) = \mathcal{W}(\mathbf{\Lambda}_k \mid c_k, \mathbf{D}_k)$

$$c_k = a + \frac{\widehat{N}_k + 1}{2} \tag{4}$$

$$\mathbf{D}_k = \mathbf{B} + \frac{1}{2}\left[\sum_n \phi_{nk} (\mathbf{x}_n - \mathbf{m}_k)(\mathbf{x}_n - \mathbf{m}_k)^\top + \nu (\mathbf{m}_k - \boldsymbol{\mu}_0)(\mathbf{m}_k - \boldsymbol{\mu}_0)^\top + (\nu + \widehat{N}_k) \mathbf{\Psi}_k\right] \tag{5}$$

Stick-breaking parameters: $q(\beta_k) = \mathcal{B}(\beta_k \mid \gamma_{k,1}, \gamma_{k,2})$

$$\gamma_{k,1} = 1 + \widehat{N}_k \tag{6}$$

$$\gamma_{k,2} = \alpha + \widehat{N}_{>k} \tag{7}$$

Assignment parameters: $q(z_n = k) = \phi_{nk}$

$$\begin{aligned} \phi_{nk} \propto \exp\Bigg\{ & \frac{1}{2}\Big(\psi_D(c_k) - \log|\mathbf{D}_k| - c_k \operatorname{tr}\Big[\mathbf{D}_k^{-1}\Big((\mathbf{x}_n - \mathbf{m}_k)(\mathbf{x}_n - \mathbf{m}_k)^\top + \mathbf{\Psi}_k\Big)\Big]\Big) \\ & + [\psi(\gamma_{k,1}) - \psi(\gamma_{k,1} + \gamma_{k,2})] + \sum_{l<k} [\psi(\gamma_{l,2}) - \psi(\gamma_{l,1} + \gamma_{l,2})] \Bigg\}, \end{aligned} \tag{8}$$

where $\psi(x) = \frac{d}{dx} \log \Gamma(x)$ and $\psi_D(x) = \sum_{d=1}^D \psi\big(x + \frac{1-d}{2}\big)$ are the digamma and multivariate digamma functions, respectively.

B Model Visualisation

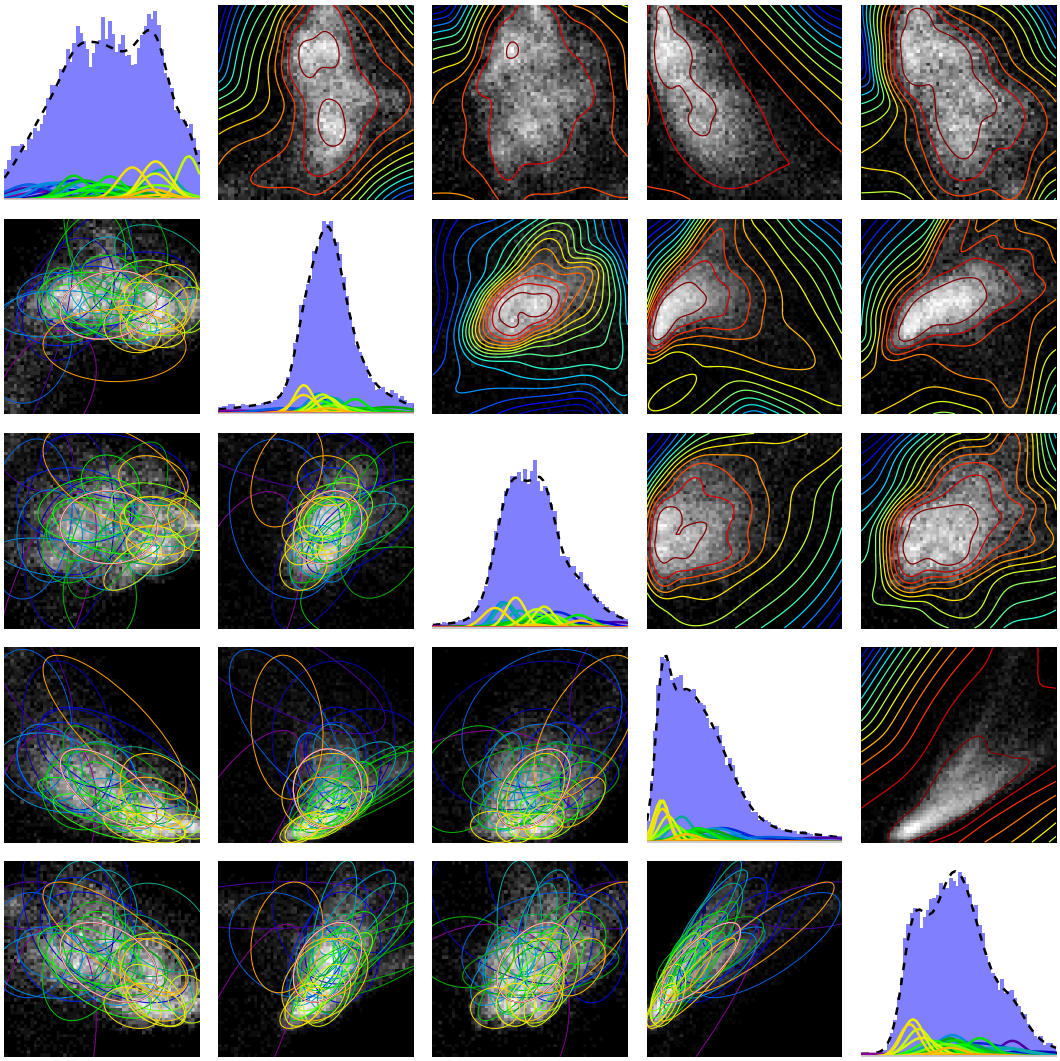


Figure 4: Empirical and fitted five-dimensional distributions for TBI subject 17770. The rows (top to bottom) and columns (left to right) correspond to MPRAGE, FLAIR, GE, T2 and PD modalities. Plots on the same column share their x-axis, and off-diagonal plots on the same row share the y-axis. Best viewed in colour.

Histograms: On the diagonal are the intensity histograms of each individual modality (in light blue). Off-diagonal, we have the two-dimensional log-histograms for each pair of modalities (column modality on x-axis vs. row modality on y-axis, lighter is higher).

Mixture components: The 1D and 2D marginals of the mixture components ($p(x^d \mid z = k)$ and $p(x^{d_1}, x^{d_2} \mid z = k)$) are overlaid on the histograms, rendered as bell curves on the diagonal and as ellipses on the lower left half, with matching colours. The mixture proportions (π_k) are represented by the scaling factor of the bell curves and thickness of the ellipse contours.

Likelihoods: The dashed line on the diagonal shows the 1D marginal likelihood ($p(x^d)$). On the upper right half, we see level curves of the 2D marginal log-likelihoods ($\log p(x^{d_1}, x^{d_2})$). Note the remarkably tight fit to the histograms in all 1D and 2D projections.