Published in final edited form as:
Med Image Comput Comput Assist Interv. 2012 ; 15(0 3): 50–57.

Incorporating Parameter Uncertainty in Bayesian Segmentation Models: Application to Hippocampal Subfield Volumetry

Juan Eugenio Iglesias[1], **Mert Rory Sabuncu**[1], **Koen Van Leemput**[1,2,3], and **the Alzheimer's Disease Neuroimaging Initiative**

[1]Martinos Center for Biomedical Imaging, MGH, Harvard Medical School, USA [2]Department of Informatics and Mathematical Modeling, DTU, Denmark [3]Departments of Information and Computer Science and of Biomedical Engineering and Computational Science, Aalto University, Finland

Abstract

Many successful segmentation algorithms are based on Bayesian models in which prior anatomical knowledge is combined with the available image information. However, these methods typically have many free parameters that are estimated to obtain point estimates only, whereas a faithful Bayesian analysis would also consider all possible alternate values these parameters may take. In this paper, we propose to incorporate the uncertainty of the free parameters in Bayesian segmentation models more accurately by using Monte Carlo sampling. We demonstrate our technique by sampling atlas warps in a recent method for hippocampal subfield segmentation, and show a significant improvement in an Alzheimer's disease classification task. As an additional benefit, the method also yields informative "error bars" on the segmentation results for each of the individual sub-structures.

1 Introduction

Many segmentation algorithms in medical image analysis are based on Bayesian modeling, in which generative image models are constructed and subsequently "inverted" to obtain automated segmentations. Such methods have a *prior* that makes predictions about where anatomical structures typically occur throughout the image, such as Markov random field models or probabilistic atlases [1, 2]. They also include a *likelihood* term that models the relationship between segmentation labels and image intensities, often incorporating explicit models of imaging artifacts [3]. Once the prior and likelihood have been specified, segmentation of a particular image proceeds by inferring the posterior distribution over all possible segmentations using Bayes' rule, and searching for the segmentation that maximizes this posterior, or estimating the volumes of specific structures.

Although these methods are clearly "Bayesian", an issue that is usually overlooked is that they only apply Bayesian analysis in an *approximate* sense. In particular, these models typically have many free parameters for which suitable values are unknown a priori. In a true Bayesian approach, such parameters need to be integrated over when inferring the segmentation posterior. But, in practice, their *optimal* values are first estimated and only the resulting point estimates are used to compute the segmentation posterior instead. In recent years generative models have started to include deformable registration methods that warp probabilistic atlases into the domain of the image being analyzed, often adding thousands of free parameters to the model [4–7]. Since many plausible atlas warps beside the truly

optimal one may exist, computing segmentations based on a single warp may lead to biased results. Furthermore, the numerical optimizers computing such high-dimensional atlas warps may not necessarily find the global optimum, further contributing to segmentation errors.

In this paper, we investigate the effect of using a more accurate approximation of the segmentation posterior in Bayesian segmentation models than the point estimates of the free model parameters. In particular, we will approximate the integral over atlas deformations in a recently proposed method for hippocampal subfield segmentation [7] using Markov chain Monte Carlo (MCMC) sampling, and compare the results to those obtained using the most probable warp only. We show that MCMC sampling yields hippocampal subfield volume estimates that better discriminate controls from subjects with Alzheimer's disease, while providing informative "error bars" on those estimates as well.

To the best of our knowledge, the issue of integrating over free parameters in Bayesian segmentation models has not been addressed before in the literature. The closest work related to the techniques used in this paper infers the posterior distribution of deformation fields in the context of computing location-specific smoothing kernels [8], quantifying registration uncertainties [9], or constructing Bayesian deformable models [10].

2 Methods

2.1 Baseline segmentation method

We start from the Bayesian method for hippocampal subfield segmentation [7] that is publicly available as part of the FreeSurfer software package[4]. In this method, a segmentation prior is defined in the form of a tetrahedral mesh-based probabilistic atlas in which each mesh vertex has an associated vector of probabilities for the different hippocampal subfields and surrounding tissues (fimbria, presubiculum, subiculum, CA1, CA2/3, CA4/DG, hippocampal fissure, white matter, gray matter, and CSF). The resolution and topology of the mesh are locally adaptive to the level of shape complexity of each anatomical region, e.g., it is coarse in uniform regions and fine around convoluted boundaries. The mesh can be deformed according to a probabilistic model on the location of the mesh nodes $p(\mathbf{x}) \propto \exp(-\varphi(\mathbf{x}))$, where $\mathbf{x}$ is a vector containing the coordinates of the mesh nodes, and $\varphi(\mathbf{x})$ is an energy function that penalizes mesh positions in which the tetrahedra are deformed [11]. This function goes to infinity if the Jacobian determinant of any tetrahedron's deformation approaches zero, and therefore ensures that the mesh topology is preserved. For a given $\mathbf{x}$, the prior probability $p_i(k|\mathbf{x})$ of tissue k occurring in voxel i is obtained by interpolating the probability vectors in the vertices of the deformed mesh. Assuming conditional independence of the labels between voxels given $\mathbf{x}$, the prior probability of a segmentation is then given by $p(\mathbf{l}|\mathbf{x}) = \Pi_i p_i(l_i|\mathbf{x})$, where $\mathbf{l} = (l_1, \ldots, l_I)^T$, $l_i \in \{1, \ldots, K\}$ is a segmentation of an image with I voxels into K tissue types.

For the likelihood, we model the intensity of voxels in tissue k as a Gaussian distribution with parameters μ_k, σ_k^2: $p(\mathbf{y}|\mathbf{l}, \theta) = \prod_i \mathcal{N}(y_i; \mu_{l_i}, \sigma_{l_i}^2)$, where the vector $\mathbf{y} = (y_1, \ldots, y_I)^T$ contains the image intensities, and $\theta = (\mu_1, \sigma_1^2, \ldots, \mu_K, \sigma_K^2)^T$ represents the Gaussian distribution parameters. A non-informative prior for $\boldsymbol{\theta}$ (i.e., $p(\boldsymbol{\theta}) \propto 1$) completes the model.

Given an image to segment, the posterior over possible segmentations is given by $p(\mathbf{l}|\mathbf{y}) = \int_{\boldsymbol{\theta}} \int_{\mathbf{x}} p(\mathbf{l}|\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}) p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y}) d\mathbf{x} d\boldsymbol{\theta}$, which takes into account the contribution of all possible values for the model parameters $\{\mathbf{x}, \boldsymbol{\theta}\}$, each weighted by their posterior probability $p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})$. In [7], this integral is approximated by estimating the parameters with maximal

[4] http://surfer.nmr.mgh.harvard.edu/

weight$\{\hat{\mathbf{x}}, \hat{\boldsymbol{\theta}}\} = \arg\max_{\{\mathbf{x},\boldsymbol{\theta}\}} p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})$, and using the contribution of those parameters only, yielding

$$p(\mathbf{l}|\mathbf{y}) \simeq p(\mathbf{l}|\mathbf{y}, \widehat{\mathbf{x}}, \widehat{\theta}) = \prod_i p_i(l_i|y_i, \widehat{\mathbf{x}}, \widehat{\theta}) \quad (1)$$

$$\text{with} \quad p_i(k|y_i, \widehat{\mathbf{x}}, \widehat{\theta}) \propto \mathcal{N}(y_i; \widehat{\mu}_k, \widehat{\sigma}_k^2) p_i(k|\widehat{\mathbf{x}}). \quad (2)$$

The segmentation maximizing this approximate posterior is obtained by simply assigning each voxel to the tissue class that maximizes Eq. (2). Furthermore, the volume of class k also has an (approximate) posterior distribution, with mean

$$v_k = \sum_i p_i(k|y_i, \widehat{\mathbf{x}}, \widehat{\theta}) \quad (3)$$

and variance

$$\gamma_k^2 = \sum_i p_i(k|y_i, \widehat{\mathbf{x}}, \widehat{\theta})[1 - p_i(k|y_i, \widehat{\mathbf{x}}, \widehat{\theta})]. \quad (4)$$

2.2 Incorporating parameter uncertainty

The approximation of Eq. (1) will be a good one if the posterior of the model parameters, $p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})$, is very peaked around $\{\hat{\mathbf{x}}, \hat{\boldsymbol{\theta}}\}$. Although this is a reasonable assumption for the Gaussian distribution parameters $\boldsymbol{\theta}$ – one cannot alter them much without considerably decreasing the likelihood of the model – assuming a sharp peak for the mesh position $\mathbf{x}$ is not necessarily accurate, since moving vertices in areas with low image contrast does not drastically change $p(\mathbf{x}, \boldsymbol{\theta}|\mathbf{y})$.

We therefore propose to use a computationally more demanding but more accurate way of approximating $p(\mathbf{l}|\mathbf{y})$. Specifically, we propose to draw a number of samples $\mathbf{x}(n)$, $n = 1, \ldots, N$ from the posterior distribution $p(\mathbf{x}|\mathbf{y}, \hat{\boldsymbol{\theta}})$ using Monte Carlo sampling, and approximate the segmentation posterior by

$$p(\mathbf{l}|\mathbf{y}) \simeq \int_{\mathbf{x}} p(\mathbf{l}|\mathbf{y}, \mathbf{x}, \widehat{\theta}) p(\mathbf{x}|\mathbf{y}, \widehat{\theta}) \mathrm{d}\mathbf{x} \simeq \frac{1}{N} \sum_{n=1}^{N} p(\mathbf{l}|\mathbf{y}, \mathbf{x}(n), \widehat{\theta}), \quad (5)$$

where in the first step we have used the mode approximation in the direction of $\boldsymbol{\theta}$, as before, but in the second step the remaining integral is approximated by summing the contributions of many possible atlas warps (with more probable warps occurring more frequently), rather than by the contribution of a single point estimate $\hat{\mathbf{x}}$ only. Given enough samples, this approximation can be made arbitrarily close to the true integral.

Once N samples $\mathbf{x}(n)$ are available, it follows from Eqs. (3–5) that the approximate posterior for the volume of tissue class k has mean and variance

$$v_k = \frac{1}{N}\sum_{n=1}^{N} v_k(n) \quad (6)$$

$$\gamma_k^2 = \frac{1}{N}\left[\sum_{n=1}^{N}[v_k(n)-v_k]^2+\gamma_k^2(n)\right], \quad (7)$$

respectively, where $v_k(n) = \Sigma_i\, p_i(k|y_i, \mathbf{x}(n), \hat{\boldsymbol{\theta}})$ and $\gamma_k^2(n)=\sum_i p_i(k|y_i, \mathbf{x}(n), \widehat{\theta})[1-p_i(k|y_i, \mathbf{x}(n), \widehat{\theta})]$.

2.3 MCMC sampling

In order to obtain the required samples $\mathbf{x}(n)$, we use a MCMC sampling technique known as the Hamiltonian Monte Carlo (HMC) method [12], which is more efficient than traditional Metropolis schemes because it uses gradient information to reduce random walk behavior. Specifically, it facilitates large steps in $\mathbf{x}$ with relatively few evaluations of the target distribution $p(\mathbf{x}|\mathbf{y}, \hat{\boldsymbol{\theta}})$ and its gradient, by iteratively assigning a random momentum to each component of $\mathbf{x}$, and then simulating the Hamiltonian dynamics of a system in which $-\log p(\mathbf{x}|\mathbf{y}, \hat{\boldsymbol{\theta}})$ acts as an internal "force". In our implementation, we discretize the Hamiltonian trajectories using the so-called leapfrog method [12], and simulate the Hamiltonian dynamics for a number of time steps sampled uniformly from [1, 50] to obtain a proposal for the Metropolis algorithm. Discretization step sizes that are adequate for some tetrahedra might be too large or small for others, leading to either slow convergence or too many rejected moves. We therefore use the following heuristic stepsize for each vertex: $\eta/\max[\partial^2(-\log p(\mathbf{x}))/\partial \mathbf{x}_j^2|_{\bar{\mathbf{x}}}]$, where η is a global adjustment factor and $\partial^2/\mathbf{x}_\mathbf{j}^2$ denotes the second derivatives with respect to the three spatial coordinates of vertex j. Two samples of $p(\mathbf{x}|\mathbf{y}, \hat{\boldsymbol{\theta}})$ obtained using the proposed scheme are displayed in Fig. 1.

3 Experiments and Results

To investigate the effect of approximating the true posterior over the segmentations using parameter sampling instead of point estimates, we compared the performance of the estimated subfield volumes for both methods (Eq. (3) vs. Eq. (6)) in an Alzheimer's disease classification task[5]. In particular, we collected the volume estimates for all seven subfields (averaged over the left and right hemispheres) into a feature vector $\mathbf{v}$ for each subject, and trained and tested a simple multivariate classifier to discern between elderly controls (EC) and Alzheimer's disease patients (AD) in the corresponding feature space. We also compared the variance ("error bars") on the subfield volume estimates for both methods (Eq. (4) vs. Eq. (7)), and investigated the effect of incorporating this information in the training of the classifier as well.

3.1 Data and experimental set-up

The 400 baseline T_1 scans from controls and AD subjects available in ADNI[6] where used in this study. The MRI pulse sequence is described elsewhere[6]. The volumes were preprocessed and parsed into 36 brain structures using FreeSurfer. We discarded 17 subjects for which FreeSurfer crashed. The demographics for the remaining 383 were: 56.2%

[5]Although this specific classification task is best performed using information from the whole brain [13], the goal of this paper is to show the effect of MCMC sampling.
[6]Online at http://www.adni-info.org/.

controls (age 76.1 ± 5.6), 43.8% Alzheimer's (age 75.5 ± 7.6); 53.6% males (age 76.1 ± 5.6), 46.4% females (age 75.9 ± 6.8).

After the segmentation of subcortical structures, the FreeSurfer hippocampal subfield segmentation routine (Section 2.1) was executed. The output $\{\hat{\mathbf{x}}, \hat{\boldsymbol{\theta}}\}$ was used to initialize the HMC sampler, which was then used to generate $N = 50$ samples per subject. The parameter η was tuned so that the average Metropolis rejection rate was approximately 25%. To decrease the correlation between successive samples, we recorded $\mathbf{x}$ at the end of every $200th$ Hamiltonian trajectory (chosen by visual inspection of the autocorrelation of subsequent runs). We allowed 300 initial "burn-in" runs before collecting samples. The running time of the sampling was roughly three hours.

3.2 Classification and ROC analysis

We used a Quadratic Discriminant Analysis (QDA) classifier, which assumes that the feature vectors $\mathbf{v}$ in each group are normally distributed according to $\mathcal{N}(\mathbf{v}|\boldsymbol{\mu}_{EC}, \boldsymbol{\Sigma}_{EC})$ and $\mathcal{N}(\mathbf{v}|\boldsymbol{\mu}_{AD}, \boldsymbol{\Sigma}_{AD})$, respectively. The means and covariances were estimated from the available training samples. In testing, a subject was classified as EC or AD by thresholding the likelihood ratio $\mathcal{N}(\mathbf{v}|\boldsymbol{\mu}_{EC}, \boldsymbol{\Sigma}_{EC})/\mathcal{N}(\mathbf{v}|\boldsymbol{\mu}_{AD}, \boldsymbol{\Sigma}_{AD}) \lessgtr \lambda$. The corresponding ROC curve (i.e., true positive rate vs. false positive rate) was obtained by sweeping the threshold λ, and the area under the curve (A_z) was then used as a measure of performance. The ROCs were computed using cross-validation with two randomly selected folds.

We also analyzed the accuracy when the volume of the whole hippocampus is thresholded to separate EC from AD. We compared two estimates of the volume: (1) the sum of the volumes of the subfields; and (2) the estimate from the FreeSurfer pipeline. Finally, to assess the effect of sampling on training and testing separately, we conducted an experiment in which the classifier was trained on point estimate volumes and evaluated on MCMC volumes, and vice versa.

3.3 Results

Fig. 2 shows the ROC curves and the areas under them (A_z) for the different methods. Also shown are the p-values of paired DeLong statistical tests [14] that evaluate if the differences in A_z are significant. At $p = 0.05$, sampling significantly outperformed point estimates in all cases (subfields and whole hippocampus). At the operating point closest to (0, 1), sampling provides a ~ 2% increase in classification accuracy. Using all the subfields performed significantly better than the whole hippocampal volume alone. All methods based on the subfield analysis outperformed the standard FreeSurfer hippocampal segmentation.

When the QDA was trained on the point estimate subfield volumes and tested on those obtained with sampling, we obtained $A_z = 0.875$, and when the roles were switched, $A_z = 0.876$. These values are better than when point estimate volumes were used for both training and testing, but worse than when sampling was used throughout, indicating that MCMC sampling is beneficial for both obtaining better discriminative directions and classifying individual subjects.

We also compared the variances of the hippocampal subfield volume posteriors (Table 1). The point estimates (Eq. (4)) clearly underestimate them, especially for the larger subfields; e.g., the standard error for CA2-3 is 0.4% of its volume, unrealistic given the poor image contrast (Fig. 1). In contrast, sampling (Eq. (7)) produces values between 5% and 10%, better reflecting the uncertainty in the estimated volumes.

In an attempt to take the MCMC volumetry uncertainty estimates into account in the classifier, we also trained a QDA by simply using all contributing volumes $v_k(n)$, $n = 1, \ldots,$

$N = 50$ in Eq. (6) for each subject – effectively using 50 times more training samples than there are training subjects. The ROC and the corresponding A_z are displayed in Fig. 2 (labeled as "error bars"), showing a modest further improvement compared to when the classifier is trained using the mean values only. Although the improvement was not statistically significant ($p \approx 0.1$), the ROC seems to be consistently better in the region that is closest to (0,1), where the operating point of the classifier would be typically defined.

4 Discussion

In this paper we proposed to approximate the segmentation posterior in probabilistic segmentation models more faithfully by using Monte Carlo samples of their free parameters. We demonstrated our technique by sampling atlas warps in a Bayesian method for hippocampal subfield segmentation, and showed a significant improvement in an Alzheimer's disease classification task. The method is general and can also be applied to other Bayesian segmentation models. It yields realistic confidence intervals on the segmentation results of individual structures, which we believe will convey important information when these techniques are ultimately applied in clinical settings. Furthermore, such confidence information may also help select the most suitable scanning protocol for imaging studies investigating the morphometry of specific anatomical structures.

Acknowledgments

This research was supported by NIH NCRR (P41-RR14075), NIBIB (R01EB006758, R01EB013565, 1K25EB013649-01), NINDS (R01NS052585), NIH 1KL2RR025757-01, Academy of Finland (133611), TEKES (ComBrain), Harvard Catalyst, and financial contributions from Harvard and affiliations.

References

1. Zhang Y, Brady M, Smith S. Segmentation of brain MR images through a hidden Markov random field model and the Expectation-Maximization algorithm. IEEE Transactions on Medical Imaging. 2001; 20(1):45–57. [PubMed: 11293691]
2. Fischl B, Salat D, Busa E, Albert M, Dieterich M, Haselgrove C, van der Kouwe A, Killiany R, Kennedy D, Klaveness S, Montillo A, Makris N, Rosen B, Dale A. Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. Neuron. 2002; 33:341–355. [PubMed: 11832223]
3. Wells W, Grimson W, Kikinis R, Jolesz F. Adaptive segmentation of MRI data. IEEE Transactions on Medical Imaging. 1996; 15(4):429–442. [PubMed: 18215925]
4. Fischl B, Salat D, van der Kouwe A, Makris N, Segonne F, Quinn B, Dale A. Sequence-independent segmentation of magnetic resonance images. NeuroIm-age. 2004; 23:S69–S84.
5. Ashburner J, Friston K. Unified segmentation. NeuroImage. 2005; 26:839–851. [PubMed: 15955494]
6. Pohl K, Fisher J, Grimson W, Kikinis R, Wells W. A Bayesian model for joint segmentation and registration. NeuroImage. 2006; 31(1):228–239. [PubMed: 16466677]
7. Van Leemput K, Bakkour A, Benner T, Wiggins G, Wald L, Augustinack J, Dickerson B, Golland P, Fischl B. Automated segmentation of hippocampal subfields from ultra-high resolution in vivo MRI. Hippocampus. 2009; 19:549–557. [PubMed: 19405131]
8. Simpson I, Woolrich M, Groves A, Schnabel J. Longitudinal brain MRI analysis with uncertain registration. Proceedings of MICCAI. 2011:647–654.
9. Risholm P, Pieper S, Samset E, Wells W. Summarizing and visualizing uncertainty in non-rigid registration. Proceedings of MICCAI. 2010:554–561.
10. Allassonniére S, Amit Y, Trouvé A. Toward a coherent statistical framework for dense deformable template estimation. Journal of the Royal Statistical Society, Series B. 2007; 69:3–29.
11. Ashburner J, Andersson J, Friston K. Image registration using a symmetric prior – in three dimensions. Human Brain Mapping. 2000; 9(4):212–225. [PubMed: 10770230]

12. Duane S, Kennedy A, Pendleton B, Roweth D. Hybrid Monte Carlo. Physics letters B. 1987; 195(2):216–222.

13. Cuingnet R, Gerardin E, Tessieras J, Auzias G, Lehéricy S, Habert M, Chupin M, Benali H, Colliot O. Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. NeuroImage. 2011; 56(2):766–781. [PubMed: 20542124]

14. DeLong E, DeLong D, Clarke-Pearson D. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988:837–845. [PubMed: 3203132]

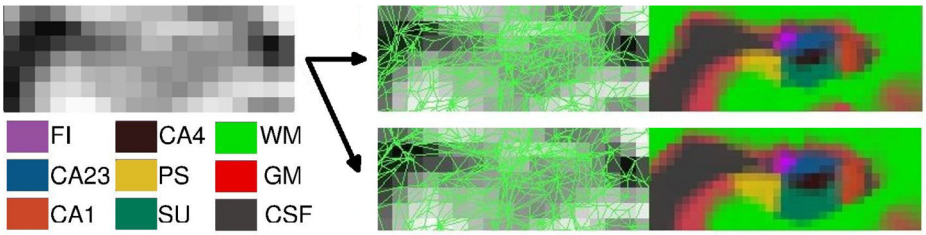
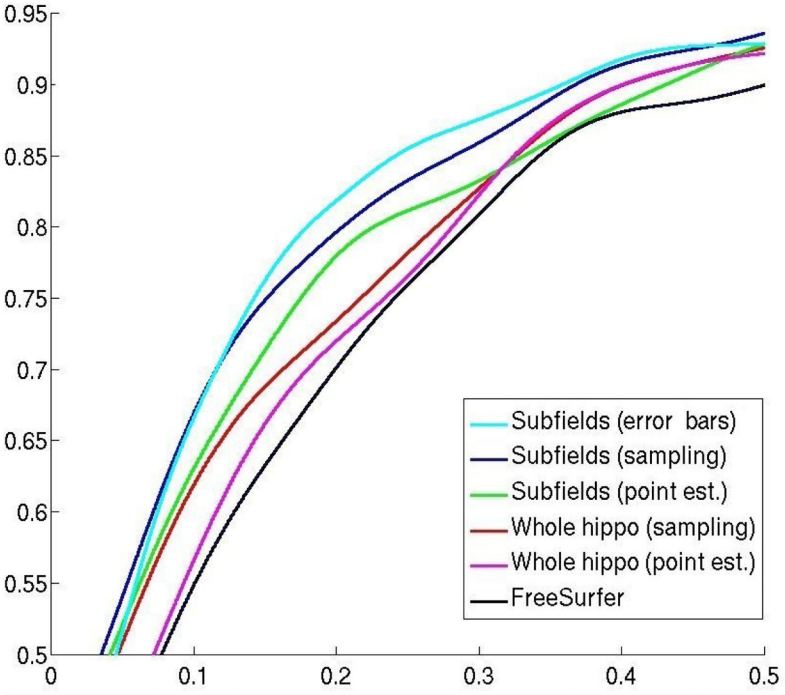


Fig. 1.

A coronal slice of an MR scan, zoomed in around the right hippocampus, and two different samples from $p(\mathbf{x}|\mathbf{y}, \hat{\boldsymbol{\theta}})$. Left: deformed mesh; right: corresponding priors $p(\mathbf{l}|\mathbf{x})$ (at the locations in which more than one class prior is greater than zero, the color is a linear combination of the class colors, weighted by their corresponding probabilities). The abbreviations in the color code are: FI: fimbria, PS: presubiculum, SU: subiculum, WM: white matter, GM: gray matter.



Method	A_z	$p_{SF_{sp}}$	$p_{SF_{pe}}$	$p_{WH_{sp}}$	$p_{WH_{pe}}$	p_{FS}
SF_{eb}	.884	.101	.022	.011	.005	.002
SF_{sp}	.882	-	.031	.019	.010	.006
SF_{pe}	.869	-	-	.046	.019	.009
WH_{sp}	.857	-	-	-	.032	.012
WH_{pe}	.844	-	-	-	-	.025
FS	.834	-	-	-	-	-

Fig. 2.

Top: ROC curves for the different methods. “FreeSurfer” refers to the whole hippocampus segmentation produced using the standard FreeSurfer pipeline. Note that only the region [0, 0.5] × [0.5, 0.95] is shown. Bottom: Area under the curve (A_z) for each method as well as p-values corresponding to DeLong tests comparing A_z for different methods. “SF” stands for subfields, “WH” for whole hippocampus, “pe” for point estimate, “sp” for sampling, “eb” for sampling with error bars (i.e. using all volumes $v_k(n)$ in Eq. (6)), and “FS” for FreeSurfer.

Table 1

Mean volumes and relative standard deviations (γ_k/v_k) for the different sub-fields, estimated using point estimates and MCMC samples of atlas deformations. HF stands for "hippocampal fissure"; the other abbreviations are as in Fig. 2.

Subfield	**HF**	**FI**	**CA4**	**CA1**	**PS**	**SU**	**CA23**
Volume (mm^3)	38	56	248	265	324	326	517
γ_k/v_k, point est. (%)	5.5	1.0	1.2	0.5	0.3	0.7	0.4
γ_k/v_k, sampling (%)	9.9	4.8	7.3	7.4	6.3	8.0	5.8