Robust Bayesian fusion of continuous segmentation maps

Benoît Audelan, Dimitri Hamzaoui, Sarah Montagne, Raphaële Renard-Penna, Hervé Delingette

Robust Bayesian fusion of continuous segmentation maps

Benoît Audelan[a,*], Dimitri Hamzaoui[a], Sarah Montagne[b], Raphaële Renard-Penna[b], Hervé Delingette[a]

[a] *Université Côte d'Azur, Inria, Epione project-team, Sophia Antipolis, France*
[b] *Academic Department of Radiology, Hôpital Pitié-Salpétrière, Sorbonne Université, Assistance Publique des Hôpitaux de Paris, Paris, France*

ARTICLE INFO

Keywords: Image segmentation, data fusion, consensus, mixture

ABSTRACT

The fusion of probability maps is required when trying to analyse a collection of image labels or probability maps produced by several segmentation algorithms or human raters. The challenge is to weight the combination of maps correctly, in order to reflect the agreement among raters, the presence of outliers and the spatial uncertainty in the consensus. In this paper, we address several shortcomings of prior work in continuous label fusion. We introduce a novel approach to jointly estimate a reliable consensus map and to assess the presence of outliers and the confidence in each rater. Our robust approach is based on heavy-tailed distributions allowing local estimates of raters performances. In particular, we investigate the Laplace, the Student's t and the generalized double Pareto distributions, and compare them with respect to the classical Gaussian likelihood used in prior works. We unify these distributions into a common tractable inference scheme based on variational calculus and scale mixture representations. Moreover, the introduction of bias and spatial priors leads to proper rater bias estimates and control over the smoothness of the consensus map. Finally, we propose an approach that clusters raters based on variational boosting, and thus may produce several alternative consensus maps. Our approach was successfully tested on MR prostate delineations and on lung nodule segmentations from the LIDC-IDRI dataset.

1. Introduction

The fusion of probability maps is necessary to solve at least two important problems related to image segmentation. The first is to establish the underlying ground truth segmentation given several binary or multi-class segmentations provided by human raters or segmentation algorithms (e.g., in the framework of multi-atlas segmentation (Sabuncu et al., 2010)). This is especially important in the medical domain, where manual contour delineations are known to suffer from potentially large inter-observer variability, due to objective factors like the image quality, but also to more subjective ones, such as the observer level of expertise (Joskowicz et al., 2019). The generated segmentation masks might have a direct impact on clinical decisions, for example in cancer radiotherapy planning where delineation discrepancies could result in significant differences regarding the definition of the target region (Petersen et al., 2007). Moreover, in the computer vision domain, accurate consensus estimations are needed for the performance assessment of segmentation algorithms, as comparison with expert delineations is the gold standard in the absence of physical or virtual phantoms. Indeed, the data fusion method used to build the reference can significantly impact the ranking result when comparing several segmentation algorithms (Lampert et al., 2016). Another domain requiring robust segmentation estimation is radiomics analysis. For instance, radiomics models can be used to make predictions about a tumor. These predictions are based on features extracted from the tumor region in the image, which is typically defined by the segmentation. It has been shown that variations in the delineation of the tumor volume lead to a poor reproducibility of the radiomics results (Kocak et al., 2019), thus highlighting the importance of robust consensus estimation to mitigate the adverse effects of inter-rater variability.

The second related problem is the fusion of probability maps that are outputted by several segmentation algorithms such as neural networks. For instance, in (Wang et al., 2019), a 3D segmentation is obtained from several 2D maps generated by a neural network using a statistical fusion approach. Similarly, data fusion is needed in (Tang et al., 2021) to aggregate results obtained at a patch level into a final segmentation. Finally, it has also been shown experimentally that combining the outputs of several segmentation algorithms often leads to improved performances (Menze et al., 2015). One can note that this problem is related to Bayesian model averaging, which consists in making predictions according to a weighted combination of models instead of relying on a single one, thus reducing the risk of overconfidence (Hoeting et al., 1999).

*Corresponding author
Email address: benoit.audelan@inria.fr

Prior work has mainly focused on the fusion of binary masks. Majority voting is perhaps the most simple method and consists in choosing pixel-wise the most predominant label among raters. A major limitation of this approach is the equal contribution of all raters to the consensus thus neglecting their potentially varying levels of performance. One of the most well known method proposed to address this issue is the STAPLE algorithm (Warfield et al., 2004). It implements a weighting strategy based on the estimated level of performance of each expert. In this case, the raters' binary segmentations are described by Bernoulli distributions and an expectation-maximization (EM) scheme allows a consensus to be built and the raters' performances to be assessed at the same time. Spatial correlation between voxels is taken into account by the introduction of a Markov random field (MRF) prior over the consensus segmentation.

Among the known shortcomings of STAPLE, there is the constraint of having only global performance estimations of raters, and thus ignoring local variations (Commowick et al., 2012; Asman and Landman, 2012, 2011). One proposed solution (Commowick et al., 2012) is to apply STAPLE in a sliding window fashion or to extend the performance parameters to the pixel level (Asman and Landman, 2012). Another limitation is that STAPLE only considers binary masks as input and is thus agnostic to the image content and especially to the presence of large image gradients (Asman and Landman, 2013; Liu et al., 2013; Akhondi-Asl et al., 2014). In (Liu et al., 2013), the authors proposed to include in the STAPLE approach simple appearance models, such as Gaussian distributions for the background and foreground, but this approach is only applicable to simple salient structures. Other extensions of STAPLE consider the case of missing data or repeated labels (Landman et al., 2012; Commowick and Warfield, 2010).

A first extension of the STAPLE algorithm for continuous inputs, which is the focus of this paper, was proposed in (Warfield et al., 2008). Raters' performances were captured by a set of biases and variances while assuming a Gaussian distribution for their continuous labels. This model was further studied in (Xing et al., 2016) and the authors demonstrated that to properly estimate rater bias, the introduction of a bias prior was required. An additional limitation of this model is the absence of a spatial prior for regularizing the consensus estimate. Furthermore, rater performances are not estimated locally but assumed to be global for the whole image, which was a limitation also shared by its binary counterpart, as noted above. Another model developed for probabilistic maps is PSTAPLE proposed by (Akhondi-Asl and Warfield, 2013). This approach is closer to the binary STAPLE formalism than (Warfield et al., 2008) and also uses an MRF prior to regularize the consensus. However, raters performances are again estimated globally for the whole image.

In this paper, we introduce a comprehensive probabilistic framework that addresses many shortcomings of approaches proposed in the literature for the fusion of continuous or categorical labels. Our baseline is the continuous STAPLE model introduced in (Warfield et al., 2008). First, we propose replacing the Gaussian likelihood with heavy-tailed distributions to model the rater input maps. In this paper, heavy-tailed distributions are broadly defined as distributions whose tails decline more slowly than the Gaussian distribution. Heavy-tailed distributions, unlike the Gaussian, are not very sensitive to outliers and, importantly, allow a spatial assessment of rater performances. Thus, image regions that differ greatly from the consensus segmentation will be considered as outliers and the contribution of that rater to the consensus will be reduced in the problematic area. In particular, the Laplace, Student's t and generalized double Pareto distributions are investigated. These distributions were used in prior works for their attractive robustness and sparsity-inducing properties. For instance, the Bayesian lasso that enables variable selection is based on the Laplace distribution (Park and Casella, 2008). A robust Bayesian clustering approach was proposed in (Archambeau and Verleysen, 2007) using Student's t distributions and a framework based on the generalized double Pareto distribution was developed for compressive sensing in (Sadeghigol et al., 2016). In this paper, we employ these distributions in a multivariate setting, which has never been done before for the generalized double Pareto distribution, to the best of our knowledge. In addition, we introduce a bias prior and take into account spatial correlation between voxels with a label smoothness prior, defined as a generalized linear model of spatially smooth kernels. We propose a common inference scheme based on variational calculus that allows the latent posterior distributions and the model parameters to be estimated in a data-driven fashion. Tractability is ensured for all heavy-tailed distributions by the use of scale mixture representations.

Last but not least, we address the unexplored issue of dissensus rather than consensus among raters. Indeed, fusing several probability maps into a single consensus map may not be meaningful when consistent patterns appear among raters. In (Langerak et al., 2010), the worse performing raters' masks were removed from the consensus estimation process at each iteration. In (Commowick and Warfield, 2009), a comparison framework for the raters' maps based on the continuous STAPLE parameters was developed. In the approach presented in this paper, several consensuses are iteratively estimated through a technique similar to variational boosting (Miller et al., 2017) and clusters of raters are identified.

Finally, although our framework is particularly suitable for the fusion of continuous probability maps generated as is by segmentation algorithms, it can also be used for merging binary masks once they are transformed to the continuous domain using, for instance, signed distance maps (Pohl et al., 2007).

We summarize the main contributions of our work below:

- The classical Gaussian likelihood used in prior work is replaced by heavy-tailed distributions to model the input rater maps. This allows raters' performances to vary locally and their contributions to the consensus to be weighted differently depending on the region in the image.
- Heavy-tailed distributions are employed in a multivariate setting, which is novel for the generalized double Pareto distribution, to the best of our knowledge.
- Bias and spatial priors are introduced, allowing a proper

bias estimation and a control over the smoothness of the consensus map.

- Tractability is ensured with a common variational inference scheme and scale mixture representations.
- The concept of a mixture of consensuses is introduced with a proper model and inference framework.

This paper is built upon an earlier work of the authors (Audelan et al., 2020). The initially proposed framework relying on a Student't distribution is expanded with the introduction of two other heavy-tailed distributions, namely the Laplace and generalized double Pareto distributions. The relationships between these distributions is discussed and a common inference framework is proposed. Moreover, we also provide more extensive experiments and further analysis. In particular, we investigate for the mixture of consensuses a new application to raters clustering. The code used to perform the experiments reported in this paper is available in this repository: `https://gitlab.inria.fr/epione/promfusion`.

The rest of the paper is organized as follows. Section 2 begins with the introduction of the robust probabilistic framework and the presentation of the heavy-tailed distributions investigated. Then, the common inference scheme based on variational calculus is developed, with details specific to each distribution. Section 3 explores the concept of a mixture of consensuses with a novel fusion algorithm similar to variational boosting. Finally, the last section gives qualitative and quantitative results on two datasets of prostate and lung nodule segmentations. We show that local variations in rater performance were successfully identified and that improved segmentation performances were obtained after fusing probability maps.

2. Robust estimate of consensus probability maps

2.1. Baseline probabilistic framework

The starting point of our work is the probabilistic framework proposed in (Warfield et al., 2008). We are given as input a set of P probability maps $\mathbf{D}_n^p$, each map consisting of N categorical probability values in K classes, i.e. $\mathbf{D}_n^p \in S^{K-1} \in \mathbb{R}^K$ where S^{K-1} is the K unit simplex space such that $\sum_{k=1}^K \mathbf{D}_{nk}^p = 1$. P is the number of raters. In our paper, a rater denotes either a human expert or a segmentation algorithm. Our objective is to estimate a consensus probability map $\mathbf{T}_n \in S^{K-1}$ over the input maps.

Each probability map is supposed to be derived from a consensus map through a random process. Let F be a link function $F(\mathbf{p}) \in \mathbb{R}^K$, where $\mathbf{p} \in S^{K-1}$, which maps probability S^{K-1} space into Euclidean space, and its inverse $F^{-1}(\mathbf{r})$ such that $F^{-1}(F(\mathbf{p})) = \mathbf{p}$. We write $\tilde{\mathbf{D}}_n^p = F(\mathbf{D}_n^p)$ and $\tilde{\mathbf{T}}_n = F(\mathbf{T}_n)$. In this paper, we follow (Pohl et al., 2007) and consider the logit function and its inverse as link functions. For instance, we have for $K = 2$:

$$F((\mathbf{p}_1, \mathbf{p}_2)^T) = \left(\log \frac{\mathbf{p}_1}{1-\mathbf{p}_1}, \log \frac{\mathbf{p}_2}{1-\mathbf{p}_2}\right)^T , \quad (1)$$

$$F^{-1}((\mathbf{r}_1, \mathbf{r}_2)^T) = \left(\frac{\sigma(\mathbf{r}_1)}{\sigma(\mathbf{r}_1)+\sigma(\mathbf{r}_2)}, \frac{\sigma(\mathbf{r}_2)}{\sigma(\mathbf{r}_1)+\sigma(\mathbf{r}_2)}\right)^T , \quad (2)$$

where $\sigma(x) = 1/(1+\exp(-x))$ is the sigmoid function.

Our baseline model follows prior works (Warfield et al., 2008; Xing et al., 2016) and assumes that the observed rater probability maps $\tilde{\mathbf{D}}^p$ are Gaussian distributed with a mean given by the consensus plus a rater bias:

$$p(\tilde{\mathbf{D}}_n^p | \tilde{\mathbf{T}}_n^p, \mathbf{b}_p, \mathbf{\Sigma}_p) = \mathcal{N}\left(\tilde{\mathbf{D}}_n^p; \tilde{\mathbf{T}}_n + \mathbf{b}_p, \mathbf{\Sigma}_p\right) . \quad (3)$$

The rater bias $\mathbf{b}_p$ and variance $\mathbf{\Sigma}_p$ do not depend on the location in the image. Together, they characterize the rater performance at the whole image level, large biases and variances being associated with poor performances.

In (Xing et al., 2016), the authors demonstrated that the absence of a bias prior leads to an indeterminate estimation. Therefore, we define a zero mean Gaussian prior over the bias, with precision β:

$$p(\mathbf{b}_p|\beta) = \mathcal{N}(\mathbf{b}; 0, \beta^{-1}\mathbf{I}_K) . \quad (4)$$

Moreover, spatial smoothness is generally considered as a desirable characteristic of segmentation maps. In the binary case (Warfield et al., 2004), a Markov random field (MRF) prior was introduced to allow a connectivity-based regularization of the discrete consensus map. Spatial consistency was also enforced through an MRF prior in PSTAPLE (Akhondi-Asl and Warfield, 2013), which is another approach extending STAPLE to continuous inputs. The main limitation of MRF priors is the impossibility of a data-driven estimation of the MRF hyperparameter β controlling the level of regularization. Because inference cannot be done in closed-form, it has to be set manually. In the context of our probabilistic framework, prior works (Warfield et al., 2008; Xing et al., 2016) did not include any smoothness prior.

In our model, spatial regularity of the consensus map is enforced by a smoothness prior defined as a generalized linear model of a set of L spatially smooth functions $\{\Phi_l(\mathbf{x})\}$, whose hyperparameters can be estimated. Let $\mathbf{x}_n \in \mathbb{R}^D$ be the position of voxel n, where D is the image dimension. Then the prior on the variables $\tilde{\mathbf{T}}_n$ is defined as:

$$p(\tilde{\mathbf{T}}_n|\mathbf{W}_l) = \mathcal{N}\left(\tilde{\mathbf{T}}_n; \sum_{l=1}^{L} \Phi_l(\mathbf{x}_n)\mathbf{W}_l; \mathbf{\Sigma}_T\mathbf{I}_K\right) , \quad (5)$$

where $\mathbf{W}_l$ are vectors of size K and where $\mathbf{\Sigma}_T \in \mathbb{R}^+$ is the prior variance. For computational convenience, we write the prior using $\mathbf{W}_k \in \mathbb{R}^L$, such that $p(\tilde{\mathbf{T}}_{nk}|\mathbf{W}_k) = \mathcal{N}(\tilde{\mathbf{T}}_{nk}; \mathbf{W}_k^T\mathbf{\Phi}_n, \mathbf{\Sigma}_T)$ where $\mathbf{\Phi}_n^T = [\Phi_1(\mathbf{x}_n), \cdots, \Phi_L(\mathbf{x}_n)]$. The weights $\mathbf{W}_k$ are placed in a weight matrix $\mathbf{W} \in \mathbb{R}^{K\times L}$ such that we can write more compactly:

$$p(\tilde{\mathbf{T}}_n|\mathbf{W}) = \mathcal{N}(\tilde{\mathbf{T}}_n; \mathbf{W}\mathbf{\Phi}_n; \mathbf{\Sigma}_T\mathbf{I}_K) \quad (6)$$

To obtain a robust description, the weights $\mathbf{W}_k$ are equipped with a zero mean Gaussian prior and precision α:

$$p(\mathbf{W}_k|\alpha) = \mathcal{N}(\mathbf{W}_k; 0, \alpha^{-1}\mathbf{I}_L) . \quad (7)$$

The spatial prior will be denoted by GLSP (Generalized Linear Spatial Prior) in the remainder of the paper. The graphical model of this baseline framework is shown in Fig. 1.

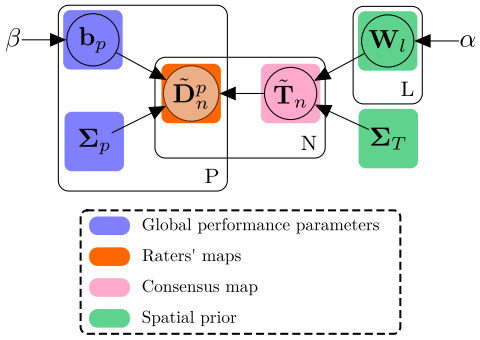


Fig. 1. Graphical representation of the baseline model with a Gaussian likelihood.

2.2. *Heavy-tailed distributions and scale mixture representation*

The main limitation of the baseline model presented in the last section is the global estimation of rater performances, thus neglecting local variations. In this paper, we address this issue by replacing the Gaussian with heavy-tailed distributions that can be written as Gaussian scale mixtures. Thus, compared with the baseline Gaussian model, we introduce an additional multiplicative scale variable τ, which varies spatially and introduces a way to weight the rater variance differently according to the location in the image.

More precisely, we introduce the Laplace, Student's t and generalized double Pareto (GDP) distributions as heavy-tailed substitutes to the Gaussian distribution. A relationship between these three distributions can be established by first introducing the power exponential distribution, also known as generalized Gaussian distribution (Pascal et al., 2013; Gómez et al., 1998). The density function of a multivariate power exponential distribution is written:

$$\mathrm{PE}(x;\tau,M,\theta) = |M|^{-\frac{1}{2}} h_{\theta,\tau}((x-\mu)^T M^{-1}(x-\mu)) \;, \tag{8}$$

for $x \in \mathbb{R}^K$ where M is a $K \times K$ covariance matrix, $\theta > 0$, $\tau > 0$, and

$$h_{\theta,\tau}(y) = \frac{\theta\Gamma\left(\frac{K}{2}\right)}{\pi^{\frac{K}{2}}\Gamma\left(\frac{K}{2\theta}\right)} \tau^{\frac{K}{2\theta}} \exp\left(-\tau y^{\theta}\right) \;. \tag{9}$$

Power exponential scale mixtures are distributions that can be represented in a hierarchical fashion using a scale mixture as follows:

$$p_X(x) = \int_\tau p_{X|\tau}(x) p_\tau(\tau) d\tau = \int_\tau \mathrm{PE}(x;\tau,M,\theta) p_\tau(\tau) d\tau \;. \tag{10}$$

Depending on the choice of parameter θ and mixing density $p_\tau(\tau)$, various distributions can be obtained. In this paper, we consider the case where the mixing density is a Gamma distribution $p_\tau(\tau) = \mathrm{Ga}(\tau;\nu,\nu)$ with shape and scale parameter $\nu > 0$:

$$\mathrm{Ga}(x;\nu,\nu) = \frac{\nu^\nu}{\Gamma(\nu)} x^{\nu-1} \exp\left(-\nu x\right) \;. \tag{11}$$

Then, the resulting distribution $p_X(x)$ obtained after marginalization of τ is a generalized t distribution (Giri, 2016) whose density function is given by:

$$p_X(x) = \frac{\theta\Gamma\left(\frac{K}{2}\right)\nu^\nu}{\pi^{\frac{K}{2}} B\left(\nu, \frac{K}{2\theta}\right)} |M|^{-\frac{1}{2}} \times \frac{1}{\left(\nu + \left((x-\mu)^T M^{-1}(x-\mu)\right)^\theta\right)^{\nu+\frac{K}{2\theta}}} \;, \tag{12}$$

where $\Gamma(x)$ and $B(a,b)$ are the Gamma and Beta functions, respectively (Arslan, 2004). Depending on the values of θ and ν, different situations can arise (Giri, 2016):

- If $\theta = 1$, we get a multivariate Student's t-distribution. Moreover, if $\nu \longrightarrow \infty$ then we recover the multivariate Gaussian.
- If $\theta = \frac{1}{2}$, we obtain a multivariate generalized double Pareto distribution. Moreover, if $\nu \longrightarrow \infty$ then we recover the multivariate Laplace distribution.

Together, the θ and ν parameters control the shape of the distribution tails. Large parameters values lead to thinner tails while smaller values lead to heavier tails (McDonald and Newey, 1988). Fig. 2 shows the four distributions and compares the tails for different parameter values. The Laplace distribution spikes at zero and has fatter tails than the baseline Gaussian. The Student's t and GDP distributions have with the parameter ν a supplementary degree of freedom in comparison with the Gaussian and Laplace distributions, allowing the level of robustness to be adapted.

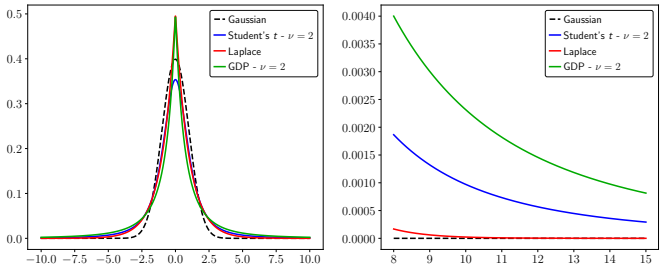


Fig. 2. Density plot of the zero-mean heavy-tailed distributions (2a), with a focus on the tail behaviors (2b).

Importantly, the three heavy-tailed distributions can all be written as Gaussian scale mixtures, namely $p_X(x) = \int_\tau \mathcal{N}\left(x;\mu,\frac{M}{\tau}\right) p_\tau(\tau) d\tau$. This re-writing is attractive for 2 reasons. First, it introduces a new variable, the dimensionless scale factor τ, that can be used to weight the rater performance depending on the image location. Indeed, when $p_\tau(\tau)$ is set to a degenerate constant distribution, i.e. $p_\tau(\tau) \propto 1$, the Gaussian scale mixture is equivalent to the Gaussian distribution. Thus by forcing the scale variable to belong to specific parametric laws (whose parameters can be estimated), we allow local variations in the raters' variance compared with the baseline Gaussian model for which it is constant. Second, the specific choice of heavy-tailed Gaussian scale mixtures enables the tractability of the inference since it leads to closed-form analytical solutions within a variational Bayes framework.

The derivation of the scale mixture for the Student's t is straightforward as the power exponential distribution of Eq. 10 amounts to a Gaussian for $\theta = 1$. The same equation for $\theta = \frac{1}{2}$ corresponds to a Laplace scale mixture. Yet, (Gómez-Sánchez-Manzano et al., 2008) showed that for any $\theta \in]0, 1]$, the power exponential can be written as a Gaussian scale mixture. However, as the mixing densities involve stable distributions, they cannot generally be written analytically, except for a few cases and in particular for $\theta = \frac{1}{2}$. The Laplace and generalized double Pareto distributions can thus be written as Gaussian scale mixtures, with an additional level of hierarchy for the latter.

Tab. 1 summarizes how the rater input map distributions, $p(\tilde{\mathbf{D}}_n^p)$, are written as scale mixtures after replacement of the Gaussian with the heavy-tailed distributions. The corresponding graphical models are presented in Fig. 3. The scale factors $\{\tau_n^p\} \in \mathbb{R}^{+NP}$ are additional latent variables not present in the Gaussian model, that separately weight each data point $\tilde{\mathbf{D}}_n^p$, allowing local variations in the performance of rater p to be taken into account. The degree of freedom ν_p^{-1} characterizes the number of data outliers that it is necessary to discard in the estimation of the consensus, i.e., a small degree of freedom ν_p indicates that rater p contributes a lot of outliers.

One can note that prior knowledge could be incorporated over the model parameters α and β by introducing, for example, Gamma hyperpriors. However, this is not the choice made in this paper, where we consider a simpler situation with uniform priors.

2.3. Model inference

To estimate the consensus, previous works used an EM algorithm. However, this approach does not lead to closed-form solutions after replacing the Gaussian with heavy-tailed distributions. Instead, we propose a common inference framework based on variational calculus (a.k.a. variational Bayes) allowing the true posterior distribution $p(U|\tilde{\mathbf{D}})$ of the model variables $U = \{\tilde{\mathbf{T}}, \mathbf{b}, S, \mathbf{W}\}$ to be approximated by a chosen family of distributions $q(U)$. We recall that $S = \{\tau\}$ for the Student's t and Laplace distributions, and $S = \{\tau, z\}$ for the generalized double Pareto distribution.

The objective is to maximize the marginal log likelihood of the data by minimizing the Kullback-Leibler divergence between the true posterior $p(U|\tilde{\mathbf{D}})$ and the approximation $q(U)$, or equivalently by maximizing the lower bound $\mathcal{L}(q)$:

$$\log p(\tilde{\mathbf{D}}) = \underbrace{\int_U q(U) \log \frac{p(\tilde{\mathbf{D}}, U)}{q(U)}}_{\mathcal{L}(q)} + \underbrace{\mathrm{KL}[q(U)\|p(U|\tilde{\mathbf{D}})]}_{\geq 0} . \quad (13)$$

Furthermore, we assume a mean field approximation leading to a factorization of the posterior approximation as follows:

$$q(U) = q(\tilde{\mathbf{T}})q(\mathbf{b})q(S)q(\mathbf{W}) . \quad (14)$$

The lower bound can be re-written as:

$$\log p(\tilde{\mathbf{D}}) \geq \mathcal{L}(q) = \int_{\tilde{\mathbf{T}}} \int_{\mathbf{b}} \int_S \int_{\mathbf{W}} q(\tilde{\mathbf{T}})q(\mathbf{b})q(S)q(\mathbf{W}) \log \frac{p(\tilde{\mathbf{D}}, \tilde{\mathbf{T}}, \mathbf{b}, S, \mathbf{W})}{q(\tilde{\mathbf{T}})q(\mathbf{b})q(S)q(\mathbf{W})} d\tilde{\mathbf{T}}\, d\mathbf{b}\, dS\, d\mathbf{W} . \quad (15)$$

If q_i denotes any of the factors in Eq. 14 and q_{-i} the product of the remaining factors, we know by variational calculus that the distribution q_i^* maximizing Eq. 15 has the form:

$$\log q_i^* = \mathbb{E}_{q_{-i}}[\log p(\tilde{\mathbf{D}}, U)] + cst , \quad (16)$$

when fixing the other distributions q_{-i}. This results leads to an iterative algorithm where the lower bound is optimized with respect to each approximate distribution q_i in turn. We present in the following sections the main results for each posterior distribution approximation. Details about the derivations can be found in Appendix A and the values of some expectations are compiled in Appendix C.

2.3.1. Consensus posterior approximation

Using Eq. 16, the consensus posterior approximation is found to be Gaussian distributed $\mathcal{N}(\tilde{\mathbf{T}}_n; \mu_{\tilde{\mathbf{T}}_n}, \mathbf{\Sigma}_{\tilde{\mathbf{T}}_n})$, with parameters given by:

$$\mathbf{\Sigma}_{\tilde{\mathbf{T}}_n} = \left[\sum_{p=1}^{P} \mathbb{E}[\tau_{np}] \mathbf{\Sigma}_p^{-1} + \Sigma_T^{-1} \mathbf{I}_K \right]^{-1} , \quad (17)$$

$$\mu_{\tilde{\mathbf{T}}_n} = \mathbf{\Sigma}_{\tilde{\mathbf{T}}_n} \left[\sum_{p=1}^{P} \mathbb{E}[\tau_{np}] \mathbf{\Sigma}_p^{-1} \left(\tilde{\mathbf{D}}_n^p - \mathbb{E}[\mathbf{b}_p] \right) + \Sigma_T^{-1} \mathbb{E}[\mathbf{W}] \mathbf{\Phi}_n \right] . \quad (18)$$

With a Gaussian likelihood, the consensus mean vector at voxel n was given by $\mu_{\tilde{\mathbf{T}}_n} = \mathbf{\Sigma}_{\tilde{\mathbf{T}}_n} \left[\sum_{p=1}^{P} \mathbf{\Sigma}_p^{-1} \left(\tilde{\mathbf{D}}_n^p - \mathbb{E}[\mathbf{b}_p] \right) + \Sigma_T^{-1} \mathbb{E}[\mathbf{W}] \mathbf{\Phi}_n \right]$. Thus, the consensus is now computed as a weighted mean of the raters' values corrected with the bias, where the weights vary spatially through the variable τ according to the raters' local performances.

2.3.2. Rater bias posterior approximation

The posterior approximation of the rater bias is also a Gaussian distribution $\mathcal{N}(\mathbf{b}_p; \mu_{\mathbf{b}_p}, \mathbf{\Sigma}_{\mathbf{b}_p})$, whose parameters are given below:

$$\mathbf{\Sigma}_{\mathbf{b}_p} = \left[\beta \mathbf{I}_K + \sum_{n=1}^{N} \mathbb{E}[\tau_{np}] \mathbf{\Sigma}_p^{-1} \right]^{-1} , \quad (19)$$

$$\mu_{\mathbf{b}_p} = \mathbf{\Sigma}_{\mathbf{b}_p} \sum_{n=1}^{N} \mathbb{E}[\tau_{np}] \mathbf{\Sigma}_p^{-1} \left(\tilde{\mathbf{D}}_n^p - \mathbb{E}[\tilde{\mathbf{T}}_n] \right) . \quad (20)$$

2.3.3. Posterior approximations of the scale variables

The scale mixture representation introduced supplementary latent variables, the scale factor τ common to the three distributions and, in addition, z for the generalized double Pareto distribution.

Applying Eq. 16 for the first scale factor leads to a Gamma distribution for the Student's t framework and to an inverse Gaussian distribution for the other two. Formula are given in Tab 2.

The GDP model has a supplementary level of hierarchy with the other scale variable z. Eq. 16 leads to the following equation

Table 1. Heavy-tailed distributions and scale mixture representations. $S_n^p = \{\tau_n^p\}$ for the Student's t and Laplace distributions, and $S_n^p = \{\tau_n^p, z_n^p\}$ for the generalized double Pareto distribution.

Likelihood	Parameters	$p(\tilde{\mathbf{D}}_n^p \vert \tilde{\mathbf{T}}_n, \mathbf{b}_p, \mathbf{\Sigma}_p, S_n^p)$
Student's t	$\theta = 1$ $\nu > 0$	$\int_{\tau_n^p} \mathcal{N}\left(\tilde{\mathbf{D}}_n^p; \tilde{\mathbf{T}}_n + \mathbf{b}_p, \frac{\mathbf{\Sigma}_p}{\tau_n^p}\right) \mathrm{Ga}\left(\tau_n^p; \frac{\nu_p}{2}, \frac{\nu_p}{2}\right) d\tau_n^p$
Laplace	$\theta = \frac{1}{2}$ $\nu \longrightarrow \infty$	$\int_{\tau_n^p} \mathcal{N}\left(\tilde{\mathbf{D}}_n^p; \tilde{\mathbf{T}}_n + \mathbf{b}_p, \frac{\mathbf{\Sigma}_p}{\tau_n^p}\right) \mathrm{InvGa}\left(\tau_n^p; \frac{K+1}{2}, \frac{1}{8}\right) d\tau_n^p$
GPD	$\theta = \frac{1}{2}$ $\nu > 0$	$\int_{\tau_n^p} \int_{z_n^p} \mathcal{N}\left(\tilde{\mathbf{D}}_n^p; \tilde{\mathbf{T}}_n + \mathbf{b}_p, \frac{\mathbf{\Sigma}_p}{\tau_n^p}\right) \mathrm{InvGa}\left(\tau_n^p; \frac{K+1}{2}, \frac{(z_n^p)^2}{2}\right) \mathrm{Ga}\left(z_n^p; \nu_p, \nu_p\right) dz_n^p d\tau_n^p$

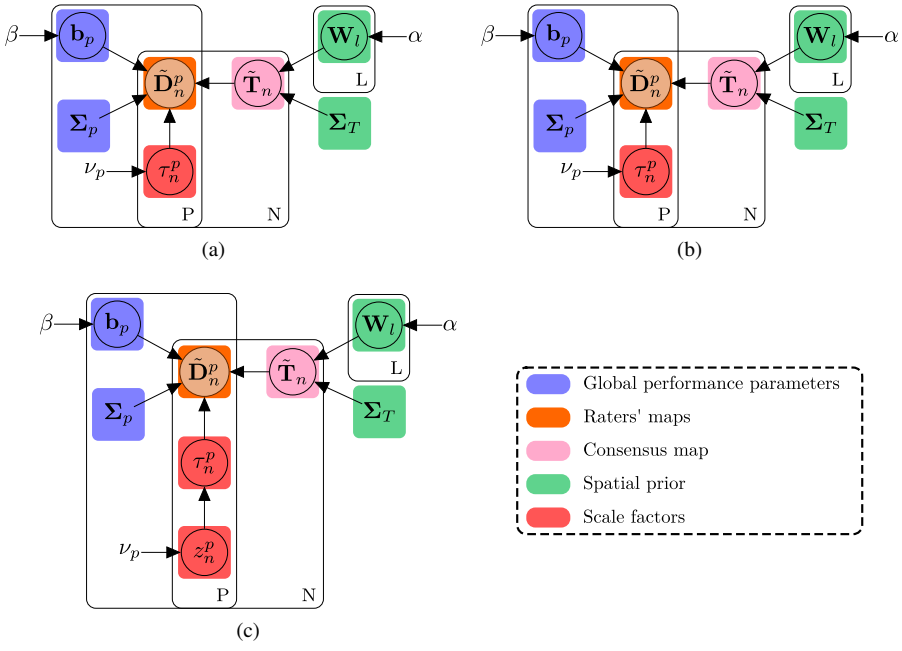


Fig. 3. Graphical models of the probabilistic framework with a Student's t-distribution (3a), a Laplace distribution (3b) and a generalized double Pareto distribution (3c).

Table 2. Posterior approximation of the scale factor τ depending on the chosen likelihood. E is given by $E = \mathbb{E}[(\tilde{\mathbf{D}}_n^p - \tilde{\mathbf{T}}_n - \mathbf{b}_p)^T \mathbf{\Sigma}_p^{-1} (\tilde{\mathbf{D}}_n^p - \tilde{\mathbf{T}}_n - \mathbf{b}_p)]$.

Likelihood	$q(\tau_n^p)$	Density	Parameters
Student's t	$\mathrm{Ga}(\tau_n^p; a_{np}, b_{np})$	$\frac{x^{a_{np}-1} b_{np}^{a_{np}}}{\Gamma(a_{np})} \exp(-b_{np} x)$	$a_{np} = \frac{\nu_p + K}{2}$, $b_{np} = \frac{\nu_p}{2} + \frac{E}{2}$
Laplace	$\mathcal{IG}(\tau_n^p; \mu_{np}, \lambda_{np})$	$\sqrt{\frac{\lambda_{np}}{2\pi x^3}} \exp\left(-\frac{\lambda_{np}(x-\mu_{np})^2}{2\mu_{np}^2 x}\right)$	$\mu_{np} = \frac{1}{2\sqrt{E}}$, $\lambda_{np} = \frac{1}{4}$
GDP	$\mathcal{IG}(\tau_n^p; \mu_{np}, \lambda_{np})$	$\sqrt{\frac{\lambda_{np}}{2\pi x^3}} \exp\left(-\frac{\lambda_{np}(x-\mu_{np})^2}{2\mu_{np}^2 x}\right)$	$\mu_{np} = \sqrt{\frac{\mathbb{E}[(z_n^p)^2]}{E}}$, $\lambda_{np} = \mathbb{E}[(z_n^p)^2]$

for $q^*(z_n^p)$:

$$q^*(z_n^p) = \frac{(\mathcal{T}_n^p)^{\frac{K+\nu_p+1}{2}} (z_n^p)^{K+\nu_p} \exp\left(-\nu_p z_n^p - \frac{(z_n^p)^2}{2}\mathcal{T}_n^p\right)}{\Gamma(K+\nu_p+1) \exp(\frac{\nu_p^2}{4\mathcal{T}_n^p}) D_{-K-\nu_p-1}\left(\frac{\nu_p}{\sqrt{\mathcal{T}_n^p}}\right)}, \quad (21)$$

where $\mathcal{T}_n^p$ stands for $\mathbb{E}\left[\frac{1}{\tau_n^p}\right] = \frac{1}{\mu_{np}} + \frac{1}{\lambda_{np}}$ and D_ν is the parabolic cylinder function of order $\nu \in \mathbb{R}$. The expectations $\mathbb{E}[z_n^p]$ and $\mathbb{E}[(z_n^p)^2]$ can be computed and are given in Appendix A.

2.3.4. *Spatial regularization variable*

We now present the posterior approximation for the variable $\mathbf{W}_k$ which controls the smoothness of the k-th consensus map. $q^*(\mathbf{W}_k)$ is a Gaussian distribution $\mathcal{N}(\mathbf{W}_k, \mu_{\mathbf{W}_k}, \mathbf{\Sigma}_{\mathbf{W}_k})$ whose pa-

rameters are:

$$\mathbf{\Sigma}_{\mathbf{W}_k} = \left[\mathbf{\Sigma}_T^{-1}\left(\sum_{n=1}^{N}\mathbf{\Phi}_n\mathbf{\Phi}_n^T\right) + \alpha\mathbf{I}_L\right]^{-1}, \tag{22}$$

$$\mu_{\mathbf{W}_k} = \mathbf{\Sigma}_{\mathbf{W}_k}\left[\sum_{n=1}^{N}\mathbf{\Phi}_n\Sigma_T^{-1}\mathbb{E}[\tilde{\mathbf{T}}_{nk}]\right]. \tag{23}$$

2.3.5. Update of the model parameters

Finally, a data-driven estimation of the model parameters can be performed. The parameters in question are α, which controls the extent of the spatial regularization, $\mathbf{\Sigma}_T$, the covariance of the consensus prior, β, the precision of the prior defined over the rater bias, $\mathbf{\Sigma}_p$, the rater variance and lastly ν_p, the degree of freedom of the Student's t and GDP distributions.

We assume that the posterior approximation of these parameters is a Dirac distribution. Applying Eq. 16 and taking the derivatives, we obtain the following update formula:

$$\alpha = \frac{LK}{\sum_{k=1}^{K}\mu_{\mathbf{W}_k}^T\mu_{\mathbf{W}_k} + \mathrm{Tr}(\mathbf{\Sigma}_{\mathbf{W}_k})}, \tag{24}$$

$$\mathbf{\Sigma}_T = \frac{\sum_{n=1}^{N}\sum_{k=1}^{K}(\mu_{\tilde{\mathbf{T}}_{nk}} - \mu_{\mathbf{W}_k}^T\mathbf{\Phi}_n)^2 + \mathbf{\Sigma}_{\tilde{\mathbf{T}}_{nk}} + \mathrm{Tr}(\mathbf{\Phi}_n\mathbf{\Phi}_n^T\mathbf{\Sigma}_{\mathbf{W}_k})}{NK}, \tag{25}$$

$$\beta = \frac{KP}{\sum_{p=1}^{P}\mu_{\mathbf{b}_p}^T\mu_{\mathbf{b}_p} + \mathrm{Tr}(\mathbf{\Sigma}_{\mathbf{b}_p})}, \tag{26}$$

$$\mathbf{\Sigma}_p = \frac{1}{N}\sum_{n=1}^{N}\Big((\tilde{\mathbf{D}}_n^p - \mu_{\tilde{\mathbf{T}}_n} - \mu_{\mathbf{b}_p})\mathbb{E}[\tau_n^p](\tilde{\mathbf{D}}_n^p - \mu_{\tilde{\mathbf{T}}_n} - \mu_{\mathbf{b}_p})^T + \mathbb{E}[\tau_n^p](\mathbf{\Sigma}_{\tilde{\mathbf{T}}_n} + \mathbf{\Sigma}_{\mathbf{b}_p})\Big). \tag{27}$$

Finally, finding the mode of $q^*(\nu_p)$ leads to the following equation when the likelihood is a Student's t distribution:

$$\sum_{n=1}^{N} -\psi\left(\frac{\nu_p}{2}\right) + \log\frac{\nu_p}{2} + 1 + \mathbb{E}[\log\tau_n^p] - \mathbb{E}[\tau_n^p] = 0, \tag{28}$$

with ψ being the digamma function. In practice, the ν_p are updated by solving the equation numerically. A similar approach could be implemented for optimizing the degree of freedom of the GDP distribution. However in practice, the numerical optimization is very unstable and we decided to set this parameter manually in the remainder of the paper.

3. Mixture of consensuses

3.1. Probabilistic framework

We also investigate the issue of dissensus rather than consensus among raters and propose a novel probabilistic framework that allows a mixture of consensuses to be estimated.

We now assume that the rater maps are derived not from a single map but from M consensus maps. We introduce for each rater a new binary latent variable $Z_{pm} \in \{0,1\}$, $\sum_m Z_{pm} = 1$, specifying from which consensus a rater map is generated. The associated component prior is given by the mixing coefficients π_m such that $p(Z_{pm} = 1) = \pi_m$. Moreover, we consider a simpler model than in the previous section, by removing the rater bias and assuming that the rater input probability maps are Gaussian distributed, i.e.:

$$p(\tilde{\mathbf{D}}^p|\tilde{\mathbf{T}}) = \prod_{m=1}^{M}\mathcal{N}(\tilde{\mathbf{D}}^p; \tilde{\mathbf{T}}_m, \mathbf{\Sigma}_p)^{Z_{pm}}. \tag{29}$$

The graphical model of the mixture of consensuses is presented in Fig. 4.

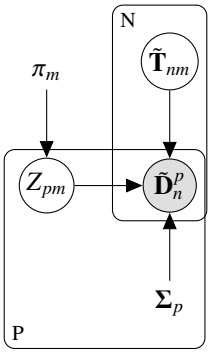


Fig. 4. Graphical model of the mixture of consensuses

3.2. Model inference

As for the robust probabilistic framework, we use variational inference to infer the consensus and model parameters. A naive solution would compute the posterior component probabilities, r_{pm} (a.k.a. the responsibilities), as a classical Gaussian mixture clustering problem with multivariate Gaussians of dimension N, thus leading to dubious results due to the curse of dimensionality (high dimension, few samples).

Instead, we propose to first reduce the dimension of each rater input map by applying a principal component analysis (PCA) and then to cluster the maps in this low-dimensional space. The resulting consensus maps are obtained by applying the inverse mapping from the component weights to the original space.

We assume again a mean field approximation implying that the approximation of the posterior factorizes as $q(U) = q(Z)q(\tilde{\mathbf{T}})$ with $U = \{Z, \tilde{\mathbf{T}}\}$. The optimal approximate distribution q_i^* maximizing the lower bound is given as before by Eq. 16. The following sections present the main results for each variational update; details of the derivations can be found in Appendix A.

3.2.1. Label posterior approximation

The variable Z indicates from which consensus each rater input map is generated. Eq. 16 applied to $q(Z_p)$ leads to a categorical distribution of parameters r_{pm}, with $r_{pm} = \rho_{pm}/\sum_m \rho_{pm}$ for $1 \leq m \leq M$, and:

$$\log\rho_{pm} = \log\pi_m + \sum_{n=1}^{N}\Big(-\frac{K}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{\Sigma}_p| - \frac{1}{2}\mathbb{E}[(\tilde{\mathbf{D}}_n^p - \tilde{\mathbf{T}}_{nm})^T\mathbf{\Sigma}_p^{-1}(\tilde{\mathbf{D}}_n^p - \tilde{\mathbf{T}}_{nm})]\Big). \tag{30}$$

3.2.2. Consensus posterior approximation

There is no longer a unique consensus but M consensuses to estimate. The approximate posterior distribution for each of them is a Gaussian distribution $\mathcal{N}(\tilde{\mathbf{T}}_{nm}; \mu_{\tilde{\mathbf{T}}_{nm}}, \mathbf{\Sigma}_{\tilde{\mathbf{T}}_{nm}})$, whose parameters are written below:

$$\mathbf{\Sigma}_{\tilde{\mathbf{T}}_{nm}} = \left[\sum_{p=1}^{P} r_{pm}\mathbf{\Sigma}_p^{-1}\right]^{-1}, \tag{31}$$

$$\mu_{\tilde{\mathbf{T}}_{nm}} = \mathbf{\Sigma}_{\tilde{\mathbf{T}}_{nm}} \sum_{p=1}^{P} r_{pm}\mathbf{\Sigma}_p^{-1}\tilde{\mathbf{D}}_n^p\,. \tag{32}$$

The raters contributions to the consensus m are now weighted by the responsibilities r_{pm}, i.e, the posterior probabilities of being generated from the consensus in question for each rater.

3.2.3. Update of the model parameters

The model parameters are the mixing coefficients π_m and the rater variance $\mathbf{\Sigma}_p$. The former is updated with the following formula:

$$\pi_m = \frac{\sum_{p=1}^{P} r_{pm}}{P}, \tag{33}$$

and the latter according to:

$$\mathbf{\Sigma}_p = \frac{1}{N}\Big(\sum_{n=1}^{N}\sum_{m=1}^{M} r_{pm}\left((\tilde{\mathbf{D}}_n^p - \mu_{\tilde{\mathbf{T}}_{nm}})(\tilde{\mathbf{D}}_n^p - \tilde{\mathbf{T}}_{nm})^T + \mathbf{\Sigma}_{\tilde{\mathbf{T}}_{nm}}\right)\Big). \tag{34}$$

The inference has been found experimentally to be very sensitive to the initial values. To increase its stability, we follow an incremental scheme inspired by variational boosting (Miller et al., 2017). We introduce one consensus map at a time and the distribution parameters of components included in the previous iterations are not updated. Initialization is performed at each iteration by summing the absolute value of the residuals $\text{res}_p = \sum_{n,m} |\tilde{\mathbf{D}}_n^p - \tilde{\mathbf{T}}_{nm}|$ and setting the responsibility for the new component to $\frac{\text{res}_p}{\sum_p \text{res}_p}$ for rater p. Other responsibilities are uniformly initialized such that $\sum_m r_{pm} = 1$. In practice, the algorithm is stopped when no rater is added to the newly introduced component after convergence. The sketch of the approach is summarized in Alg. 1.

Algorithm 1: Mixture of consensuses

Inputs:

- $\tilde{\mathbf{D}}$ `// raters continuous segmentation maps`
- M_{cons} `// maximum number of consensuses`

$\tilde{\mathbf{D}} = \text{PCA}(\tilde{\mathbf{D}})$ `// dimensionality reduction`
$m = 0$ `// current number of consensuses in the model`
while $m < M_{cons}$ **do**
  $m \longleftarrow m + 1$
  while *not converged* **do**
    for $1 \leq i \leq m$, $1 \leq p \leq P$ **do**
      Estimate r_{pi} and π_i from Eq. 30 and Eq. 33
      Estimate $\mathbf{\Sigma}_p$ from Eq 34
    end
    for $1 \leq n \leq N$, $i = m$ **do**
      `// distribution parameters of components already in the model at` $m-1$ `are not updated`
      Update $\mathbf{\Sigma}_{\tilde{\mathbf{T}}_{ni}}$ and $\mu_{\tilde{\mathbf{T}}_{ni}}$ from Eq. 31 and Eq. 32
    end
  end
  if $\pi_m < 10^{-10}$ **then**
    $m = M_{\text{cons}}$ `// stop when the new component is empty`
  end
end
$\mu_{\tilde{\mathbf{T}}} \longleftarrow \text{PCA}^{-1}(\mu_{\tilde{\mathbf{T}}})$ `// return to the original space`
return $\mu_{\tilde{\mathbf{T}}}$

4. Results

4.1. Material

We investigate our approach on prostate and nodule segmentations. Two types of experiments were conducted, depending on the nature of the segmentations used as input.

In the first case, we used binary segmentations drawn by medical experts as inputs. The binary masks were first transformed into probabilistic segmentations computed as the sigmoid of a Euclidean signed distance map, whose 0 level corresponds to the segmentation boundaries. The sigmoid function is defined as $\sigma(x) = 1/(1 + \exp(-\lambda_s x))$, where λ_s controls the slope of the transition between regions. Small lambda values are associated with increased uncertainty along the segmentation border.

In the second case, the inputs were continuous segmentations produced by several neural networks, trained beforehand by cross-validation on an independent training set. The consensus estimated between the neural networks was then compared to a surrogate ground truth defined as a majority vote of the medical experts' delineations. All networks used in this paper have a classical U-net architecture (Ronneberger et al., 2015).

The prostate dataset is a private collection of 40 MRI exams performed at 3 tesla (SIGNA$^{\text{TM}}$ Architect, GE Healthcare, Chicago, IL and MAGNETOM$^{\text{TM}}$ Skyra, Siemens Healthcare, Erlangen, Germany). All MRI protocols included 3D T2 weighted images with 0.5 mm to 1.0 mm slice thickness. The in-plane pixel size ranges from 0.4 mm to 0.8 mm. The dataset includes manual prostate delineations from 7 radiologists, whose levels of experience are dissimilar: three are considered as experts, two have an intermediary level, and the remaining two are junior radiologists with less experience. This dataset, with binary segmentations, will be denoted latter as ProstateBin.

Moreover, 5 neural networks were trained by 5-fold cross-validation on a subset of 98 3D T2 weighted images selected from the publicly available SPIE-AAPM-NCI PROSTATEx dataset (Litjens et al., 2014), and for which (Meyer et al., 2019) released ground truth segmentations made by an expert urolo-

gist. The performances of the networks were then evaluated on 7 unseen test scans extracted from the private dataset of 40 images described above. This set composed of 7 images and of the associated predictions of the 5 neural networks, will be referred to as ProstateNet.

The nodule dataset is the publicly available LIDC-IDRI database of lung CT scans (Armato III et al., 2011). It contains nodule delineations drawn by 4 radiologists. The raw CT images were re-sampled in a pre-processing step to obtain a common spatial resolution of 1 mm in all directions. A first set was constituted by considering the 20 largest nodules annotated by all radiologists. This set, containing 20 lesions and binary segmentations, will be denoted as NoduleBin in the remainder of the paper. The LIDC-IDRI dataset was furthermore separated into a training and a testing set. The former was used to train 9 neural networks by 9-fold cross validation. The networks were then evaluated on the 34 nodules of the test set having a 10mm minimum diameter. The set composed of the 34 test cases and the associated networks predictions will be referred to as NoduleNet in the remainder.

Tab. 3 summarizes the characteristics of the datasets used in the experiments. All results reported in this paper were obtained in 3D. The size of the inputs depends on the dataset. For the experiments on the nodule datasets, we used a cube of size $48 \times 48 \times 48$ centered at each nodule location. Computations were performed on the entire image for the prostate datasets. The typical image size in the prostate datasets was $160 \times 500 \times 500$.

4.2. *Robust probabilistic framework*

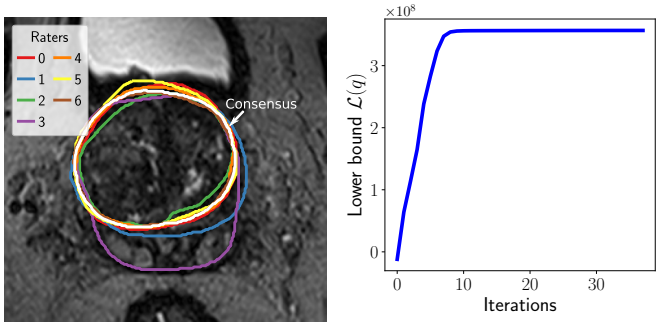


Fig. 5. Robust fusion of prostate binary segmentation masks using the Laplace distribution. (5a) Raters' manual delineations and estimated consensus shown on an axial T2 weighted image. (5b) Evolution of the lower bound.

4.2.1. *Qualitative analysis.*

First, we demonstrate the effectiveness of our proposed robust probabilistic model in taking into account the spatially varying performances of the raters. We consider the fusion of 7 binary prostate delineations from the ProstateBin dataset drawn by human experts into a single consensus using a framework based on the Laplace distribution fitted in 3D. The coefficient λ_s of the sigmoid function used to convert the input masks to probabilities was arbitrarily set to 5. The 7 raters segmentations and the estimated consensus are shown in Fig. 5a. During the inference, we maximize the lower bound, $\mathcal{L}(q)$, on the marginal log likelihood of the data. The evolution over the iterations is plotted in Fig. 5b.

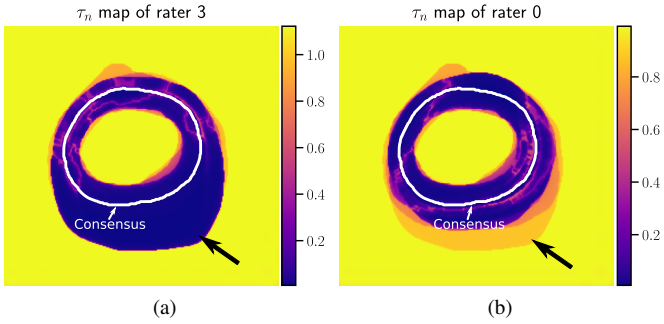


Fig. 6. The outlier, rater 3, exhibits locally poor performances linked to lower values of τ_n^p, in particular at the bottom of the image in the region indicated by the black arrow (6a). In contrast, rater 0 shows higher τ_n^p values in the same area (6b).

It can be seen that rater 3 seems to be an outlier with respect to the other raters at the bottom of the image, although they agree elsewhere. This local variation of the rater performance is successfully captured by the scale factor τ_n^p that spatially modulates the contribution of each rater to the consensus. In areas of poor rater performance, τ exhibits lower values which correspond to larger rater variance. Locally, raters with weak confidence will not contribute as much as others to the consensus. This is shown in Fig. 6a and 6b, where rater 3 has smaller τ_n values than rater 0 at the bottom of the image in the region highlighted by the black arrows.

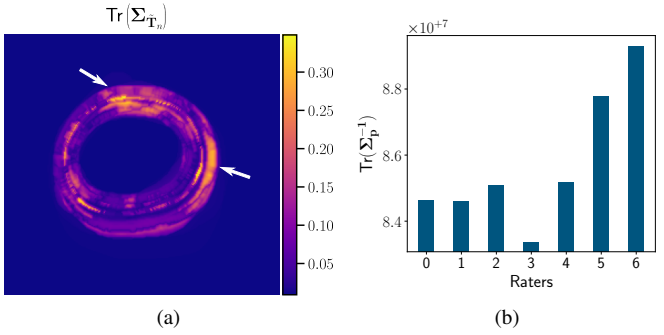


Fig. 7. Uncertainty map of the consensus (7a). Comparison of the raters precisions (7b).

The trace of the matrix $\mathbf{\Sigma}_{\tilde{\mathbf{T}}_n}$ represents the uncertainty associated with the consensus. Low trace values correspond to a high confidence in the consensus and are typically found in area where all raters agree, as shown in Fig. 7a. One can observe that the highest uncertainty is not located at the bottom of the image, where there is a disagreement between rater 3 and the others, but in the image regions indicated by the white arrows. This somewhat counter-intuitive result is explained by the fact that the consensus uncertainty is estimated as a combination of the raters' precisions, weighted by the scale factor τ_n^p, as shown in Eq. 17. Rater 3 is considered by the model to present poorer performances in comparison to the others, which corresponds

Table 3. Characteristics of the datasets used for the experiments. (MV: majority vote.)

	ProstateBin	ProstateNet	NoduleBin	NoduleNet
# of cases	40	7	20	34
# of experts	7	5	4	9
Expert category	Radiologists	Neural networks	Radiologists	Neural networks
Segmentation type	Binary	Continuous	Binary	Continuous
Surrogate ground truth	NA	MV of 7 radiologists	NA	MV of 4 radiologists

to low scale factor and precision values, as shown in Fig. 6a and Fig. 7b. Thus, rater 3 is barely taken into account for the consensus uncertainty estimation, which relies much more on the other experts. One can note that, in the Gaussian baseline model, there are no scale variables. The consensus uncertainty is a simple combination of the raters' precisions and is thus constant within the image. In particular, regions of disagreement between raters have the same level of uncertainty as regions where all raters agree. Therefore, our robust approach leads to a more realistic estimate of the consensus uncertainty, by allowing variations in the image depending on the level of agreement between raters.

The possibility of localizing visually, in a convenient manner, the most unreliable regions of the consensus is an advantage of our model in comparison to the Gaussian baseline model, but also to the classical binary STAPLE algorithm, which does not provide any estimate of the uncertainty associated with the consensus.

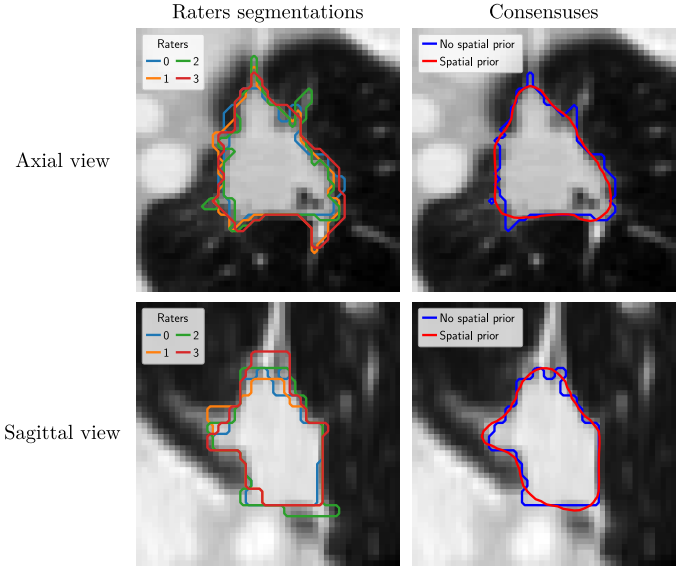


Fig. 8. Impact of the spatial prior on the smoothness of the consensus map obtained with the Laplace distribution on a nodule segmentation case on CT scan from the NoduleBin dataset.

One contribution of our work is the introduction of a spatial regularization prior over the consensus map. In the discrete setting, the spatial consistency of the consensus was enforced with an MRF prior, for example in (Warfield et al., 2004). In our continuous approach, spatial correlations between voxels are taken into account by the definition of a GLSP prior over the consensus map. The key parameters are the spacing, s, between the basis function centers, the standard deviations (or radii), r, of the Gaussian functions and the position of the origin basis function. Together, they influence the level of regularization of the consensus map, large spacing and radii being associated with smoother outputs. Fig. 8 compares the consensuses, obtained for a nodule of the NoduleBin dataset with or without spatial regularization, in a model where the input rater maps are assumed to follow a Laplace distribution. For the model fitted with spatial regularization, the spacing, s, was set to 4 and the radius was equal to 12. The influence of the prior is clearly visible with far smoother contours.

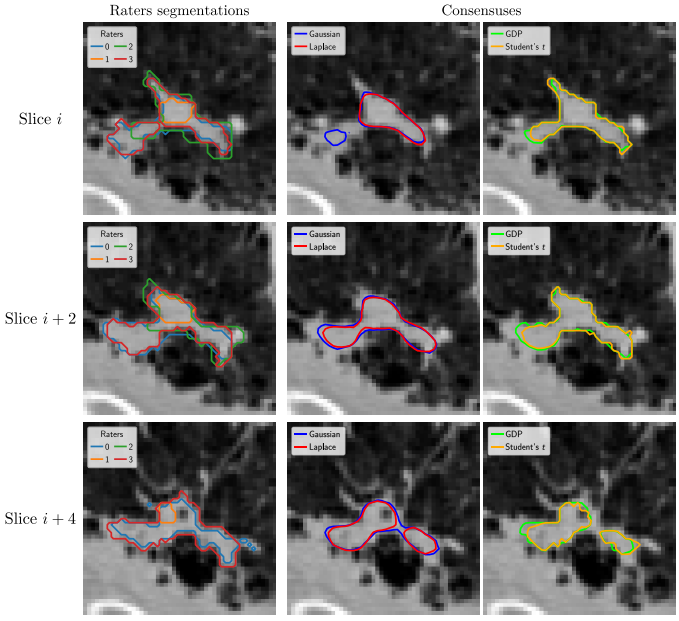


Fig. 9. Comparison of the heavy-tailed distributions on a nodule segmentation case extracted from the NoduleBin dataset.

We provide a visual comparison between the heavy-tailed distributions and the Gaussian reference in Fig. 9 on a nodule segmentation example from the NoduleBin dataset. The models are fitted in 3D with same spatial regularization parameters for all distributions. The inputs are the four radiologists' binary segmentations transformed to probabilities, using as before $\lambda_s = 5$ for the sigmoid function. The four manual delin-

eations are given in the first column and the associated consensuses in the second one. It can be seen that the Student's t and GDP distributions give similar results. Both have an additional degree of freedom in comparison with the Laplace and Gaussian, which allows the shape of the tail of the distribution to be adapted to the data. For this case, the mean degree of freedom ν_p between raters is 0.3 after convergence for the Student's t. It was manually set to 2 for the GDP. These values lead to heavier tails than the Laplace and Gaussian, which could explain the similar results.

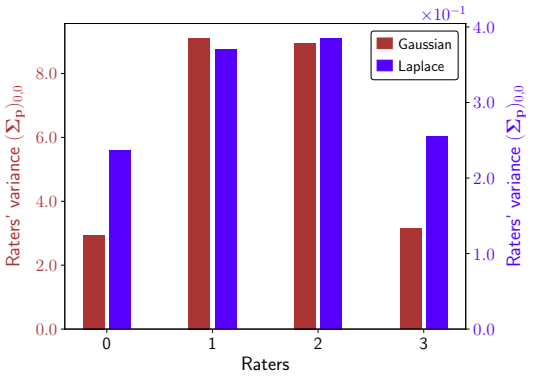


Fig. 10. Raters variances $(\mathbf{\Sigma}_p)_{0,0}$ corresponding to the foreground region for the models based on a Gaussian and Laplace distributions.

The possibility of locally varying rater contributions to the consensus for the robust model leads to rater performance estimates different from those obtained with a global estimation, as for the Gaussian baseline. In Fig. 10, we compare the variances $(\mathbf{\Sigma}_p)_{0,0}$ corresponding to the foreground region obtained with a Gaussian or Laplace distributions. $(\mathbf{\Sigma}_p)_{0,0}$ and $(\mathbf{\Sigma}_p)_{1,1}$ can be considered the counterparts of the sensitivity and the specificity estimated in the binary setting by the STAPLE algorithm. In our framework, a large variance corresponds to poor rater performance. One can observe that the ranking between raters is close between the two distributions. However, the orders of magnitude are different with smaller variances for the robust approach. This can be explained by the fact that the experts agree in most of the image regions. The discrepancies, which contribute to poor rater performances, lie only on a small narrow band along the nodule boundary. The local estimation of the performances allows this to be taken into account by the robust approach, which leads logically to smaller variances.

The objective of variational inference is the maximization of a lower bound over the data marginal likelihood. This lower bound can be computed to monitor the model convergence but also to provide a criteria for model selection (Blei et al., 2017). Fig. 11 compares the lower bound values reached after convergence and the inference time for the different distributions on the 20 nodules from the NoduleBin dataset. The Student's t seems to lead to the highest lower bound values. This distribution has, with the GDP, an additional degree of freedom allowing the shape of the distribution to be modified and better fitted to the data. Because of numerical instabilities, this parameter is fixed manually for the GDP, whereas it is learnt automatically for the Student's t in a data-driven way. This could explain the higher lower bound values reached by the Student's t.

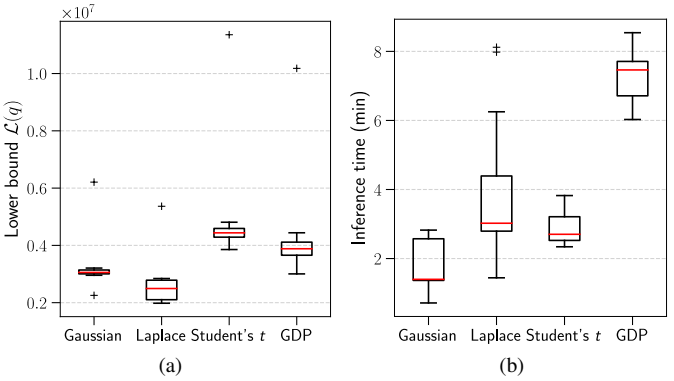


Fig. 11. Lower bound values reached after convergence (11a) and computation times obtained on the 20 nodules from NoduleBin dataset (11b).

Regarding the computational times, the Gaussian baseline model seems to be faster, but with fewer parameters and variables to estimate. For the GDP, the expectations involving the scale factor z are evaluated with Lentz's algorithm as shown in Appendix A, which logically leads to longer computation times.

Fig. 12 provides a visual comparison between our robust approach based on a Laplace distribution and models proposed in previous works. In particular, we compare our model with the original STAPLE algorithm introduced in (Warfield et al., 2002), which does not include any spatial regularization of the consensus. We also compare it to two extensions of STAPLE for continuous inputs, namely, to PSTAPLE, introduced in (Akhondi-Asl and Warfield, 2013), which uses an MRF as regularization prior, and to the continuous STAPLE algorithm, proposed in (Warfield et al., 2008), from which our approach was developed. The comparison is performed on a nodule segmentation case from the NoduleBin dataset. The inputs are therefore the delineations drawn by the radiologists and transformed to continuous maps, except for the STAPLE algorithm which handles binary inputs. For the MRF prior of PSTAPLE, a 4-connectivity neighborhood is considered, and β, the MRF hyperparameter, is set to 2. Regarding the GLSP prior used in our model, the parameters are the same as used previously.

One can observe that STAPLE and PSTAPLE lead to similar results, which could be expected as the latter is a direct extension of the former for probabilistic inputs. The effect of the MRF prior can be noted, with slightly smoother contours for PSTAPLE. The continuous STAPLE and the robust approach based on a Laplace likelihood also produce similar maps. However, our approach includes a spatial regularization prior which logically leads to smoother outputs. Moreover, the hyperparameter of the spatial prior is learnt automatically in our approach, which is an advantage in comparison with the MRF prior. Our approach is also more robust with respect to the outlier rater 1, in particular compared to the STAPLE algorithm.

4.2.2. Quantitative analysis.

The main difficulty when assessing the performances of data fusion algorithms is the absence of an unequivocal ground truth,

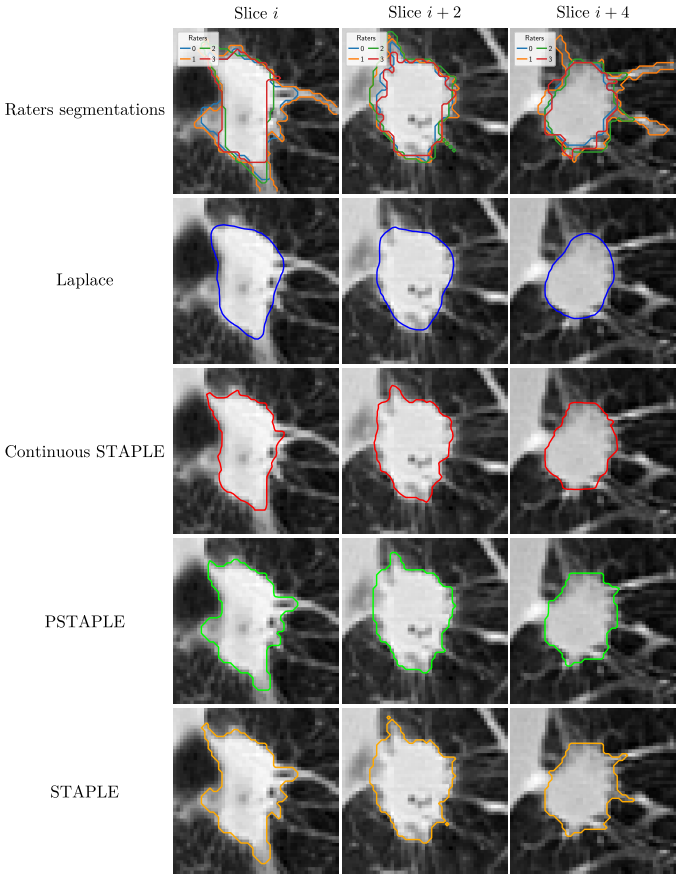


Fig. 12. Comparison of the robust model using a Laplace distribution with approaches proposed in previous works on a nodule segmentation case from the NoduleBin dataset.

which prevents any accurate quantitative comparison. This is particularly true in the medical imaging domain, where the inter-rater variability can be large.

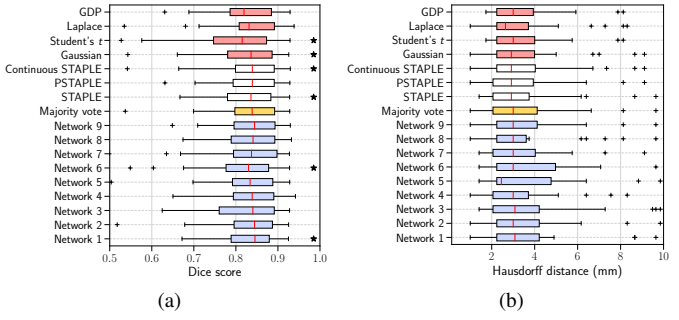


Fig. 13. Dice score (13a) and Hausdorff distance (13b) distributions over the NoduleNet dataset. Distributions marked with a ⋆ are found to be significantly different from the majority voting baseline with the Wilcoxon signed-rank test at significance level 0.05.

In this section, we provide a quantitative comparison framework between our robust probabilistic approaches and methods proposed in previous works, including the most simple one, i.e., majority voting. We now consider probabilistic segmentations generated by several neural networks trained by cross-validation and tested on the NoduleNet and ProstateNet

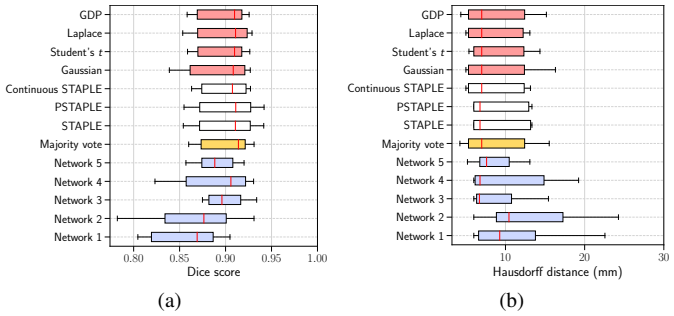


Fig. 14. Dice score (14a) and Hausdorff distance (14b) distributions over the ProstateNet dataset.

datasets, as detailed in section 4.1. The data fusion approaches are used to estimate a consensus between the predictions made by the different neural networks. Therefore, in contrast to the previous section, the inputs are already continuous. They need to be binarized only for the STAPLE and majority voting algorithms.

The consensuses are compared to a surrogate ground truth defined as a majority vote of the human raters' segmentations of the test set. We emphasise that, while this reference overcomes the absence of real ground truth, it is also a limitation of the comparison. We use the Dice score and the Hausdorff distance as performance metrics for the comparison. The former is a region-based metric and the latter a distance-based metric. Evaluating the performance of segmentation algorithms is a difficult task (Fenster and Chiu, 2005), and defining proper metrics remains an open challenge. Therefore, the Dice score and the Hausdorff distance may themselves be a limitation of the comparison and this should be kept in mind when analysing the results.

Regarding the lung nodules, the NoduleNet dataset contains 34 lesions, all of diameter greater than 10 mm. Smaller nodules were excluded from the analysis. The ensemble of raters is composed of 9 neural networks, whose performances on the test set in terms of Dice score and Hausdorff distance are presented in Fig. 13. The Wilcoxon signed-rank test was used to compare distributions with the majority voting baseline. For the prostate dataset, 5 neural networks were used to produce probabilistic segmentations of a test set of 7 images. Thus, the prostate and nodule datasets allow us to perform a comparison between the data fusion approaches on two different structures of interest, but also with a different number of rater input maps. Fig. 14 shows the Dice scores and Hausdorff distances distributions for the prostate. Due to the small sample size, differences between distributions and the majority voting baseline were not tested.

First, we can observe that the neural networks of the NoduleNet dataset have more homogenous performances than those of the ProstateNet dataset. The latter were trained with a much smaller number of cases, which may explain the larger discrepancies.

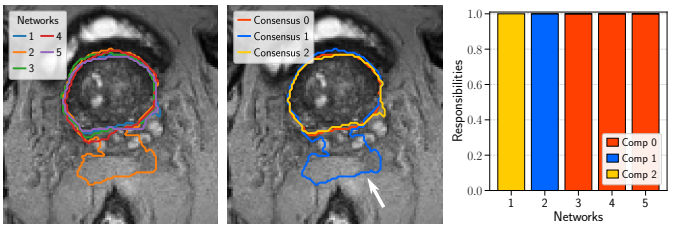
Second, the differences between methods are small and almost never statistically significant. This is also visible in Tab. 4 and 5, which give the mean Dice score and mean Hausdorff distance for each method on the nodule and prostate datasets,

Table 4. Mean Dice scores and Hausdorff distances computed between the consensuses estimated with different methods from several neural network outputs, and the reference defined as a majority vote of experts on the NoduleNet dataset.

	Dice score	Hd (mm)	Hd 95% (mm)
Majority vote	**0.83** (±0.08)	4.05 (±3.3)	2.09 (±2.29)
STAPLE	**0.83** (±0.07)	4.01 (±3.36)	2.14 (±2.3)
PSTAPLE	**0.83** (±0.07)	3.9 (±3.23)	2.07 (±2.25)
Continuous STAPLE	**0.83** (±0.08)	3.91 (±2.97)	2.19 (±2.38)
Gaussian	0.82 (±0.08)	3.86 (±2.97)	2.18 (±2.35)
Laplace	**0.83** (±0.08)	**3.51** (±2.27)	1.81 (±1.3)
Student's t	0.79 (±0.17)	3.56 (±2.08)	1.81 (±1.02)
GDP	0.81 (±0.09)	3.61 (±2.08)	**1.76** (±1.05)

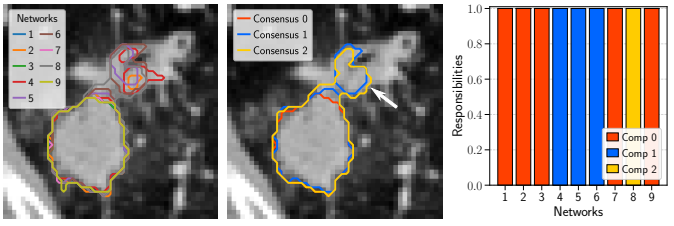
Table 5. Mean Dice scores and Hausdorff distances computed between the consensuses estimated with different methods from several neural network outputs, and the reference defined as a majority vote of experts on the ProstateNet dataset.

	Dice score	Hd (mm)	Hd 95% (mm)
Majority vote	**0.9** (±0.03)	8.9 (±4.49)	4.32 (±1.6)
STAPLE	**0.9** (±0.04)	9.21 (±3.77)	4.68 (±2.27)
PSTAPLE	**0.9** (±0.04)	9.14 (±3.69)	4.66 (±2.24)
Continuous STAPLE	**0.9** (±0.03)	8.65 (±3.82)	4.26 (±1.54)
Gaussian	0.89 (±0.04)	9.11 (±4.57)	4.79 (±2.46)
Laplace	**0.9** (±0.03)	**8.6** (±3.75)	**4.22** (±1.5)
Student's t	**0.9** (±0.03)	9.06 (±3.81)	4.23 (±1.55)
GDP	**0.9** (±0.03)	8.86 (±4.39)	4.27 (±1.61)



(a) (b) (c)

Fig. 15. Mixture of consensuses on a prostate segmentation example from the ProstateNet dataset. Input probabilistic segmentations produced by neural networks (15a). Estimated consensuses (15b). Responsibilities with 3 relevant components (15c).



(a) (b) (c)

Fig. 16. Mixture of consensuses for a lung nodule from the NoduleNet dataset. Input probabilistic segmentations produced by neural networks (16a). Estimated consensuses (16b). Responsibilities with 3 relevant components (16c).

respectively. In particular, the simple majority voting approach already gives good results, even better than those produced by the more complex STAPLE algorithm. Regarding our framework, better Dice scores seem to be obtained with a Gaussian distribution than with a Student's t or GDP likelihoods. In contrast, the latter two lead to smaller Hausdorff distances. The model based on a Laplace distribution appears to be the most complete, as it produces balanced results between the region- and the distance-based metrics. In particular, it leads to the largest Dice scores and smallest Hausdorff distances on both datasets.

Although the differences are not statistically significant, these experiments show that our robust probabilistic framework achieves state-of-the-art results and even seems to lead to slightly better performances when the model uses a Laplace distribution.

4.3. Mixture of consensuses

In this last result section, we provide examples of mixtures of consensuses in Fig. 15 and 16. The inputs are the probabilistic segmentations produced by the neural networks trained by cross-validation. The mixture model is fitted on two examples extracted from the ProstateNet and NoduleNet datasets.

Fig 15b and 16b show the consensuses obtained after convergence. In both cases, three relevant contours are found. Without the mixture approach, only one consensus corresponding to the first component would have been obtained, and the regions indicated by arrows would have been ignored.

The responsibilities are presented in Fig 15c and 16c. They indicate from which consensus each network segmentation map was generated. Thus, this method provides a novel way to cluster raters depending on their segmentations for a given image.

We now explore the idea of clustering raters over a batch of images, in particular over the 34 and 7 images of the NoduleNet and ProstateNet datasets. For each image of these two test sets, mixtures of consensuses were estimated and the networks were assigned to the consensus corresponding to their highest responsibility. This leads to a first clustering of the raters at the image level. Results are then aggregated over the whole test sets using hierarchical clustering with a complete-linkage approach, based on the following distance: $d(x, y) = N - N_{xy}$, where x and y denotes two raters, N is the number of segmentation cases in the dataset and N_{xy} is the number of segmentation cases where rater x and rater y are assigned to the same consensus. At each step, the two clusters having the most consensuses in common are combined. Results are presented in Fig. 17. It shows, for example, that the network 6 is assigned to the same consensus as networks 4 and 5 in at least 41.2% of the nodule segmentation cases.

Although the networks seem to have similar performances on the NoduleNet dataset, as shown in Fig. 13, this approach allows two main clusters to be extracted. The group composed of networks 4, 5 and 6 appears to have a significantly different behavior than the others, as they only share 32.4% of the consensuses on the whole dataset. This difference can be visually assessed in Fig. 16a, where networks 4, 5 and 6 lead to a larger segmented region than the others.

The differences between networks are smaller on the ProstateNet dataset. Networks 3, 4 and 5 are always assigned

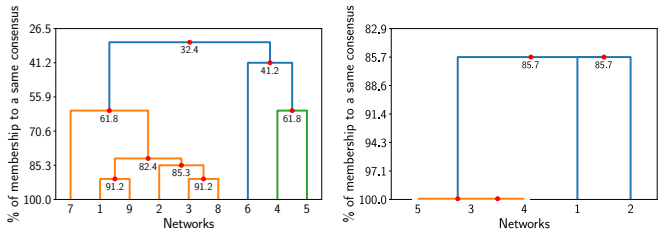


Fig. 17. Complete-linkage clustering of the networks based on the percentage of membership of the same consensus for the NoduleNet (17a) and ProstateNet (17b) datasets.

to the same consensuses. In contrast, networks 1 and 2 are isolated in 14.3% of the cases. According to the results presented in Fig. 14, they seem to exhibit poorer performances than the others. This difference appears to be confirmed by the clustering approach.

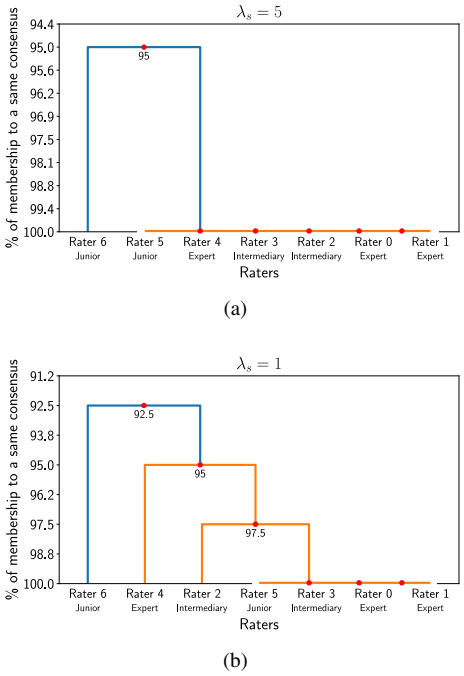


Fig. 18. Complete-linkage clustering of the human raters on the ProstateBin dataset, based on the percentage of membership of the same consensus, for two values of the coefficient λ_s of the sigmoid function used to convert the binary segmentation masks to continuous maps.

Finally, we study the application of the mixture of consensuses for the clustering of raters for whom only binary segmentations are available. In particular, we apply the approach to the segmentations drawn by the 7 radiologists on the images of the ProstateBin dataset. For each of the 40 images, a mixture of consensuses model is fitted after converting the binary segmentation masks to probabilities. We study the influence of the coefficient λ_s of the sigmoid function by presenting results for $\lambda_s = 5$, which assumes sharp transitions between image regions, and for $\lambda_s = 1$, which corresponds to a scenario with more uncertainty along the segmentation boundaries. Dendograms for the two λ_s values are shown in Fig. 18.

First, one can observe that the number of clusters increases with smaller λ_s values. Moreover, for $\lambda_S = 5$, 6 raters out of 7 are always grouped together over the 40 images. Rater 6 is the only one to be isolated, and only on 2 images of the dataset. This may be related to the rater lack of experience, as he/she is one of the two junior radiologists of the panel. For $\lambda_s = 1$, two other raters are extracted by the clustering approach. One has an intermediary level of experience, but the other one is considered as an expert. This result demonstrates the applicability of our approach for studying the inter-rater variability.

5. Discussion

Estimating a consensus between raters is an important task in the medical imaging domain. Our work focuses on the specific problem of fusing continuous segmentation maps. It addresses three major limitations of approaches proposed in previous works, namely the estimation of the rater bias, the regularization of the consensus map and the local assessment of the rater performances.

Comparison with state-of-the-art methods showed the effectiveness of our approach. However, a limitation of the study is the definition of the ground truth used to evaluate and compare the data fusion methods. The use of majority voting in this paper is arbitrary, and the resulting surrogate reference may actually be a flawed estimate of the real ground truth. This limitation is not specific to our paper. It is a general problem when comparing segmentation algorithms. Yet, (Lampert et al., 2016) showed that the approach used to form the ground truth highly influences the ranking between algorithms. This problem is particularly important for medical imaging, because of the difficulty in collecting high-quality ground truths and because of the inter-rater variability. Even when the ground truth is available, for instance in the presence of numerical or physical phantoms, the metric used to assess the performances may impact the result (Fenster and Chiu, 2005; Taha and Hanbury, 2015). How to properly evaluate the quality of segmentations remains an open issue and an interesting challenge for future work.

One contribution of our work is the introduction of a spatial prior to regularize the consensus map. In this paper, the spatial regularity was enforced using a GLSP prior, but there are alternatives for the regularization of continuous fields. One possibility is, for example, to define a prior penalizing the total variations in the consensus map (Babacan et al., 2008).

Although specifically designed for continuous inputs, our data fusion approach can handle binary segmentations, once they are converted to probabilities. In this paper, we used a transformation based on a Euclidean distance map and the sigmoid function. Varying the value of the parameter λ_s of the sigmoid leads to different consensus estimates by allowing various levels of uncertainty to be simulated. It is an advantage in comparison to the discrete data fusion methods, which neglect uncertainty by always assuming sharp transitions between image regions. However, one limitation of the approach followed in this paper is that the coefficient is independent of the location, leading to equal levels of uncertainty along the segmentation

boundaries in all image regions. It could be improved by varying the slope of the transitions depending on the location and contrast in the image, allowing, for example, more uncertainty to be assumed in areas where raters disagree.

Our approach provides a statistical framework for assessing the performances of the raters. In particular, the mixture of consensuses model is a novel approach to study the inter-rater variability, cluster raters and detect outliers. The approach, inspired by variational boosting, allows the appropriate number of consensuses to be estimated in a data-driven way. It requires a reduction of dimension, performed in this paper by PCA. This method maximizes the variance of the data projected in the latent space, which is an attractive property when the objective is to identify patterns among raters. However, other reduction techniques could be used, and their investigation represents an avenue for future work. In addition, approaches developed for high-dimensional data clustering and based on the introduction of regularization constraints over the class covariance matrices could also be investigated (Bouveyron et al., 2007).

In practice, the mixture of consensuses was tested on datasets with a number of raters equal to 5, 7 and 9. It is likely that the search for multiple consensuses needs a number of raters at least equal to 4 or 5 to be relevant. Our approach based on variational boosting was driven by the need to obtain a set of consensus segmentations that is not too dependent on the initial parameters of the Gaussian mixture. Yet, it would be important in future work to further explore the robustness of the consensuses with respect to additional parameters (number of PCA components, use of alternative dimension reduction techniques...).

Moreover, the mixture of consensuses was only tested with a simplified model assuming a Gaussian likelihood and no rater bias. This model could be extended by adding a bias for the raters and replacing the Gaussian with a robust distribution. However, in contrast to the classical robust model, the rater bias would not be directly related to over- or under-segmentation anymore, because of the projection into the latent space. Similarly, it would not be possible to connect in a straightforward manner the variations in the scale factor to specific locations in the image, making the model less interpretable. Furthermore, the mixture of consensuses with a Gaussian distribution is already more robust than the Gaussian model with a unique component, in particular by allowing outliers to be isolated. We note that a related approach for outlier detection was proposed by (Commowick and Warfield, 2009). However, it is purely based on a statistical comparison of the raters' biases and variances and does not allow several consensuses to be generated.

Another interesting topic of research is the evaluation of the intra-rater variability, which reflects the consistency of a rater when segmenting the same image several times. This could be assessed using our model, by fusing the different segmentations of an image produced by a rater and sharing the variance $\mathbf{\Sigma}_p$ between the input maps. After convergence, this parameter would give an estimate of the intra-rater variability.

Moreover, the experiments in this paper were designed such that the raters performances were evaluated independently from one image to another. Yet, it is reasonable to assume that part of a rater's performance does not depend on a given image. For example, errors related to a lack of experience may be repeated over a whole set of images. In order to take this observation into account, one possibility would be to add a prior over the rater performance parameters, and then learn the prior hyperparameters using several segmentation cases. One can assume that this strategy would lead to a more robust estimate of the raters' performances.

This approach could also be followed to constrain the scale factor to take more uniform values between the raters. Indeed, we can see on Fig. 6 that, although raters 0 and 3 agree in the corner of the image, their τ_n^p values are not equal. This does not mean that they do not contribute equally to the consensus in these image regions, as each rater contribution also depends on the rater variance. However, more uniform scale factor values could be obtained by the introduction of a prior and sharing its parameters between the raters.

6. Conclusion

Consensus estimation between raters is an important but difficult problem. The main challenge is to assess the performance of each rater and the associated uncertainty properly. Many approaches have been proposed to address this challenge for discrete inputs. In contrast, the continuous setting has received less attention.

In this paper, we focused on this latter case and proposed a novel robust Bayesian framework for the fusion of continuous segmentation maps based on heavy-tailed distributions. A major contribution of our work is the local assessment of the raters performances, which were only estimated globally in previous approaches. These locally varying performances are made possible by writing the heavy-tailed distributions as Gaussian scale mixtures. Moreover, the spatial consistency of the consensus is enforced by the introduction of a regularization prior. We propose a convenient inference framework based on variational calculus that allows the model variables and parameters to be estimated in a data-driven way.

Consensuses obtained with the heavy-tailed distributions were visually compared and this qualitative comparison demonstrated that the distributions lead to different segmentation results. A quantitative comparison with methods proposed in previous works was performed using probabilistic segmentations generated by neural networks. We showed that our approaches achieved state-of-the-art results. In particular, the model fitted with a Laplace distribution led to slightly better performances, both for the region- and distance-based metrics.

This paper also explores the novel concept of mixtures of consensuses. Unlike classical approaches, several consensuses can be obtained, which highlight the potential presence of several patterns among raters. This model also provides a novel way to cluster raters, allowing outliers to be extracted.

Several ideas to extend our framework were developed in the discussion. In particular, applying our framework to several segmentations generated by a rater on the same image to study the intra-rater variability seems to be a promising research avenue for future work.

In conclusion, we believe our method may be a useful tool to estimate a consensus between several segmentation maps, and the approach could be of interest in other fields of application where data fusion is required.

Acknowledgments

This work was partially funded by the French government, through the UCA[JEDI] and 3IA Côte d'Azur "Investments in the Future" projects managed by the National Research Agency (ANR) with the reference numbers ANR-15-IDEX-01 and ANR-19-P3IA-0002 and supported by the Inria Sophia Antipolis - Méditerranée "NEF" computation cluster. Data were partially extracted from the Clinical Data Warehouse of the Greater Paris University Hospitals (Assistance Publique – Hôpitaux de Paris).

Appendix A. Variational updates.

Derivations of the variational update formula are given in this appendix, for the robust probabilistic framework of section 2 that uses heavy-tailed distributions, and for the mixture of consensuses model of section 3.

Appendix A.1. Robust probabilistic framework.

The log joint probability of the heavy-tailed probabilistic model $p(\tilde{\mathbf{D}}, \tilde{\mathbf{T}}, \mathbf{b}, \mathbf{W}, S)$ factorizes as $p(\tilde{\mathbf{D}}|\tilde{\mathbf{T}}, \mathbf{b}, S)p(\mathbf{b})p(\tilde{\mathbf{T}}|\mathbf{W})p(\mathbf{W})p(S)$.

Appendix A.1.1. Update of $q(\tilde{\mathbf{T}})$.

Eq. 16 applied to the consensus posterior approximation gives:

$$\begin{aligned}\log q^*(\tilde{\mathbf{T}}) &= \mathbb{E}[\log p(\tilde{\mathbf{D}}|\tilde{\mathbf{T}}, \mathbf{b}, S) + \log p(\tilde{\mathbf{T}}|\mathbf{W})] + cst\,,\\ &= \mathbb{E}[\log p(\tilde{\mathbf{D}}|\tilde{\mathbf{T}}, \mathbf{b}, \tau) + \log p(\tilde{\mathbf{T}}|\mathbf{W})] + cst\,.\end{aligned} \tag{A.1}$$

$p(\tilde{\mathbf{D}}|\tilde{\mathbf{T}}, \mathbf{b}, \tau)$ is a Gaussian distribution according to the scale mixture representation. Discarding the terms independent of $\tilde{\mathbf{T}}_n$, $\mathbb{E}[\log p(\tilde{\mathbf{D}}_n|\tilde{\mathbf{T}}_n, \mathbf{b}, \tau_n)]$ is written:

$$\begin{aligned}\mathbb{E}[\log p(\tilde{\mathbf{D}}_n|\tilde{\mathbf{T}}_n, \mathbf{b}, \tau_n)] = &-\frac{1}{2}\tilde{\mathbf{T}}_n^T\left(\sum_{p=1}^{P}\boldsymbol{\Sigma}_p^{-1}\mathbb{E}[\tau_n^p]\right)\tilde{\mathbf{T}}_n\\ &+ \tilde{\mathbf{T}}_n^T\left(\sum_{p=1}^{P}\boldsymbol{\Sigma}_p^{-1}\mathbb{E}[\tau_n^p](\tilde{\mathbf{D}}_n^p - \mathbb{E}[\mathbf{b}_p])\right) + cst\,.\end{aligned} \tag{A.2}$$

The second term in Eq. A.1 is due to the spatial regularization and can be expressed as follows:

$$\mathbb{E}[\log p(\tilde{\mathbf{T}}_n|\mathbf{W})] = -\frac{1}{2}\tilde{\mathbf{T}}_n^T\boldsymbol{\Sigma}_T^{-1}\mathbf{I}_K\tilde{\mathbf{T}}_n + \tilde{\mathbf{T}}_n^T\boldsymbol{\Sigma}_T^{-1}\mathbb{E}[\mathbf{W}]\boldsymbol{\Phi}_n + cst\,. \tag{A.3}$$

After regrouping and identifying the quadratic and linear terms in $\tilde{\mathbf{T}}_n$, we recognize a Gaussian distribution of parameters given by Eqs. 17 and 18.

Appendix A.1.2. Update of $q(\mathbf{b})$.

Following the same approach, we have for the rater bias:

$$\log q^*(\mathbf{b}) = \mathbb{E}[\log p(\tilde{\mathbf{D}}|\tilde{\mathbf{T}}, \mathbf{b}, \tau) + \log p(\mathbf{b}|\beta)] + cst\,. \tag{A.4}$$

Considering rater p, the first term of Eq. A.4 gives:

$$\begin{aligned}\mathbb{E}[\log p(\tilde{\mathbf{D}}^p|\tilde{\mathbf{T}}, \mathbf{b}_p, \tau^p)] = &-\frac{1}{2}\mathbf{b}_p^T\left(\sum_{n=1}^{N}\boldsymbol{\Sigma}_p^{-1}\mathbb{E}[\tau_n^p]\right)\mathbf{b}_p\\ &+ \mathbf{b}_p^T\left(\sum_{n=1}^{N}\boldsymbol{\Sigma}_p^{-1}\mathbb{E}[\tau_n^p](\tilde{\mathbf{D}}_n^p - \mathbb{E}[\tilde{\mathbf{T}}_n])\right) + cst\,,\end{aligned} \tag{A.5}$$

and the second term can be written as $\mathbb{E}[\log p(\mathbf{b}_p|\beta)] = -\frac{\beta}{2}\mathbf{b}_p^T\mathbf{b}_p + cst$. Combining the two and rearranging leads to the Gaussian distribution described by Eqs. 19 and 20.

Appendix A.1.3. Update of $q(\tau)$.

We now present the derivations for the posterior approximation of the scale factor τ. Discarding the terms independent of τ_n^p, Eq. 16 gives:

$$\log q^*(\tau_n^p) = \mathbb{E}[\log p(\tilde{\mathbf{D}}_n^p|\tilde{\mathbf{T}}, \mathbf{b}_p, \tau_n^p) + \log p(\tau_n^p)] + cst\,. \tag{A.6}$$

The results for the different distributions are reported below.

Student's t distribution. The prior over the scale factor follows a Gamma distribution. Eq. A.6 can then be re-written as follows:

$$\begin{aligned}\log q^*(\tau_n^p) = &\left(\frac{K+\nu_p}{2} - 1\right)\log\tau_n^p\\ &- \frac{1}{2}\tau_n^p\left(\mathbb{E}[(\tilde{\mathbf{D}}_n^p - \tilde{\mathbf{T}}_n - \mathbf{b}_p)^T\boldsymbol{\Sigma}_p^{-1}(\tilde{\mathbf{D}}_n^p - \tilde{\mathbf{T}}_n - \mathbf{b}_p)] + \nu_p\right) + cst\,.\end{aligned} \tag{A.7}$$

We recognize a Gamma distribution of parameters $\frac{K+\nu_p}{2}$ and $\frac{\nu_p+E}{2}$ as given in Tab. 2, with $E = \mathbb{E}[(\tilde{\mathbf{D}}_n^p - \tilde{\mathbf{T}}_n - \mathbf{b}_p)^T\boldsymbol{\Sigma}_p^{-1}(\tilde{\mathbf{D}}_n^p - \tilde{\mathbf{T}}_n - \mathbf{b}_p)]$.

Laplace distribution. In this case, the prior is defined as an inverse Gamma distribution of parameters $\frac{K+1}{2}$ and $\frac{1}{8}$. Eq. A.6 leads to:

$$\begin{aligned}\log q^*(\tau_n^p) &= \frac{K}{2}\log\tau_n^p - \frac{\tau_n^p}{2}E - \frac{K+3}{2}\log\tau_n^p - \frac{1}{8\tau_n^p} + cst\,,\\ &= -\frac{3}{2}\log\tau_n^p - \frac{1}{2}\left(\frac{E}{\tau_n^p}\left(\tau_n^p - \frac{1}{\sqrt{4E}}\right)^2\right) + cst\,,\end{aligned} \tag{A.8}$$

where E is defined above. Thus, the scale factor posterior approximation is an inverse Gaussian distribution whose parameters are given in Tab. 2.

GDP distribution. The prior over the scale factor is also an inverse Gamma distribution, but of parameters $\frac{K+1}{2}$ and $\frac{(z_n^p)^2}{2}$. Thus, we have:

$$\begin{aligned}\log q^*(\tau_n^p) &= \frac{K}{2}\log\tau_n^p - \frac{\tau_n^p E}{2} - \frac{K+3}{2}\log\tau_n^p - \frac{\mathbb{E}[(z_n^p)^2]}{2\tau_n^p} + cst\,,\\ &= -\frac{3}{2}\log\tau_n^p - \frac{1}{2}\left(\frac{E}{\tau_n^p}\left(\tau_n^p - \sqrt{\frac{\mathbb{E}[(z_n^p)^2]}{E}}\right)^2\right) + cst\,.\end{aligned} \tag{A.9}$$

Therefore, Eq. 16 again yields an inverse Gaussian distribution with the parameters given in Tab. 2.

Appendix A.1.4. Update of $q(z)$.

This section gives the derivations for the additional scale factor z which appears when the generalized double Pareto distribution is written as a Laplace scale mixture. Eq. 16 applied to $q(z_n^p)$ gives:

$$\begin{aligned}\log q^*(z_n^p) &= (K+1)\log z_n^p - \mathcal{T}_n^p\frac{(z_n^p)^2}{2} + (\nu_p - 1)\log z_n^p\\ &\quad - \nu_p z_n^p + cst\,,\\ &= (K+\nu_p)\log z_n^p - \frac{1}{2}\left((z_n^p)^2\mathcal{T}_n^p + 2\nu_p z_n^p\right) + cst\,,\end{aligned} \tag{A.10}$$

where $\mathcal{T}_n^p = \mathbb{E}\left[\frac{1}{\tau_n^p}\right]$. The normalization constant of Eq. 21 can be obtained by integration of Eq. A.10. Let $J^+(p,q,r)$ be the following integral:

$$J^+(p,q,r) = \int_0^\infty x^p \exp(qx - rx^2)dx\,, \tag{A.11}$$

with $p \geq 0$, $-\infty < q < \infty$ and $r > 0$. It can then be shown (Neville, 2013) that:

$$J^+(p,q,r) = (2r)^{-\frac{p+1}{2}}\Gamma(p+1)\exp\left(\frac{q^2}{8r}\right)D_{-p-1}\left(-\frac{q}{\sqrt{2r}}\right), \tag{A.12}$$

where D_ν is the parabolic cylinder function of order $\nu \in \mathbb{R}$. From Eq. A.10, we have that $q^*(z_n^p) \propto (z_n^p)^{K+\nu_p}\exp\left(-\frac{(z_n^p)^2}{2}\mathcal{T}_n^p - \nu_p z_n^p\right)$. Therefore, using Eq. A.12 with $p = K+\nu_p$, $q = -\nu_p$ and $r = \frac{\mathcal{T}_n^p}{2}$, we get the density of Eq. 21. In practice, we only need $\mathbb{E}[z_n^p]$ and $\mathbb{E}[(z_n^p)^2]$ to perform the inference. These expectations can be computed using the same approach. For $\mathbb{E}[z_n^p]$, we use Eq. A.12 with $p = K+\nu_p+1$, $q = -\nu_p$, and $r = \frac{\mathcal{T}_n^p}{2}$ which gives:

$$\mathbb{E}[z_n^p] = \frac{(K+\nu_p+1)D_{-K-\nu_p-2}\left(\frac{\nu_p}{\sqrt{\mathcal{T}_n^p}}\right)}{\sqrt{\mathcal{T}_n^p}D_{-K-\nu_p-1}\left(\frac{\nu_p}{\sqrt{\mathcal{T}_n^p}}\right)}\,. \tag{A.13}$$

Likewise, we have $p = K+\nu_p+2$ with same q and r for $\mathbb{E}[(z_n^p)^2]$, which yields:

$$\mathbb{E}[(z_n^p)^2] = \frac{(K+\nu_p+1)(K+\nu_p+2)D_{-K-\nu_p-3}\left(\frac{\nu_p}{\sqrt{\mathcal{T}_n^p}}\right)}{\mathcal{T}_n^p D_{-K-\nu_p-1}\left(\frac{\nu_p}{\sqrt{\mathcal{T}_n^p}}\right)}\,. \tag{A.14}$$

The function $R_\nu(x)$ defined as $R_\nu(x) = \frac{D_{-\nu-2}(x)}{D_{-\nu-1}}(x)$ leads to underflow problems for large x or ν. Therefore, we follow (Neville, 2013) and compute the ratio using Lentz's algorithm, which is based on the continued fraction representation of the function.

Appendix A.1.5. Update of $q(\mathbf{W})$.

Eq. 16 applied to $\mathbf{W}_k$ gives:

$$\begin{aligned}\log q^*(\mathbf{W}_k) &= \mathbb{E}[\log p(\tilde{\mathbf{T}}_k|\mathbf{W}_k) + \log p(\mathbf{W}_k)] + cst\,,\\ &= -\frac{1}{2}\mathbf{W}_k^T\left(\mathbf{\Sigma}_T^{-1}\sum_{n=1}^N \mathbf{\Phi}_n^T\mathbf{\Phi}_n\right)\mathbf{W}_k\\ &\quad + \mathbf{W}_k^T\left(\mathbf{\Sigma}_T^{-1}\sum_{n=1}^N \mathbf{\Phi}_n\mathbb{E}[\tilde{\mathbf{T}}_{nk}]\right) - \frac{\alpha}{2}\mathbf{W}_k^T\mathbf{W}_k + cst\,.\end{aligned} \tag{A.15}$$

Regrouping the quadratic and linear terms in $\mathbf{W}_k$, we obtain a Gaussian distribution whose parameters are given by Eqs. 22 and 23.

Appendix A.1.6. Update of the remaining parameters.

The update formulas for α, $\mathbf{\Sigma}_T$, $\mathbf{\Sigma}_p$, β and ν_p are obtained by considering these parameters as variables and assuming that their posterior approximation $q(.)$ is a Dirac distribution. The mode of the posterior distribution approximation is found by maximizing Eq. 16, which leads to Eqs. 24–28.

Appendix A.2. Mixture of consensuses.

The model log joint probability $p(\tilde{\mathbf{D}}, \tilde{\mathbf{T}}, Z)$ factorizes as $p(\tilde{\mathbf{D}}|\tilde{\mathbf{T}}, Z)p(Z)$.

Appendix A.2.1. Update of $q(Z)$.

Eq. 16 applied to the variable Z leads to:

$$\log q^*(Z) = \mathbb{E}[\log p(\tilde{\mathbf{D}}|\tilde{\mathbf{T}}, Z) + \log p(Z)] + cst\,. \tag{A.16}$$

The first term can be developed to give:

$$\begin{aligned}\mathbb{E}[\log p(\tilde{\mathbf{D}}|\tilde{\mathbf{T}}, Z)] = \sum_{p=1}^P\sum_{m=1}^M z_{pm}\Bigg(\sum_{n=1}^N -\frac{K}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{\Sigma}_p|\\ - \frac{1}{2}\mathbb{E}[(\tilde{\mathbf{D}}_n^p - \tilde{\mathbf{T}}_{nm})^T\mathbf{\Sigma}_p^{-1}(\tilde{\mathbf{D}}_n^p - \tilde{\mathbf{T}}_{nm})]\Bigg),\end{aligned} \tag{A.17}$$

and the second term is equal to $\sum_{p=1}^P\sum_{m=1}^M z_{pm}\log\pi_m$. Summing the two expectations and taking the exponential, we get $q^*(Z) \propto \prod_p\prod_m \rho_{pm}^{Z_{pm}}$, where the expression of ρ_{pm} is given by Eq. 30. We finally obtain a product of categorical distributions with parameters r_{pm}, such that $q^*(Z) = \prod_p\prod_m r_{pm}^{Z_{pm}}$.

Appendix A.2.2. Update of $q(\tilde{\mathbf{T}})$.

Applying Eq. 16 to the consensus posterior approximation and discarding the terms independent with respect to the mth map leads to:

$$\begin{aligned}\log q^*(\tilde{\mathbf{T}}_{nm}) = -\frac{1}{2}\tilde{\mathbf{T}}_{nm}^T\left(\sum_{p=1}^{P} r_{pm}\boldsymbol{\Sigma}_p^{-1}\right)\tilde{\mathbf{T}}_{nm} \\ + \tilde{\mathbf{T}}_{nm}^T\left(\sum_{p=1}^{P} r_{pm}\boldsymbol{\Sigma}_p^{-1}\tilde{\mathbf{D}}_n^p\right) + cst\,.\end{aligned} \tag{A.18}$$

We recognize a Gaussian distribution whose parameters are given by Eq. 31 and Eq. 32.

Appendix A.2.3. Update of the model parameters.

The update formulas for $\boldsymbol{\Sigma}_p$ and π_m are obtained by considering them as variables whose approximate posterior is a Dirac distribution. We find the mode of each distribution by maximizing Eq. 16 and using the fact that $\sum_m \pi_m = 1$ for the mixing coefficients, which leads to Eqs. 33 and 34.

Appendix B. Lower bound.

In this paper, we propose a variational inference scheme to estimate the posterior approximations. It is based on the maximization of a lower bound $\mathcal{L}(q)$ over the data marginal log likelihood. The lower bound can be computed and is used in practice as a stopping criterion, except when the framework is based on a GDP distribution, because of the long computation time.

Appendix B.1. Robust probabilistic framework.

We can re-write Eq. 15 as follows:

$$\begin{aligned}\mathcal{L}(q) = \mathbb{E}[\log p(\tilde{\mathbf{D}}|\tilde{\mathbf{T}},\mathbf{b},S)] + \mathbb{E}[\log p(S)] + \mathbb{E}[\log p(\mathbf{b})] \\ + \mathbb{E}[\log p(\tilde{\mathbf{T}}|\mathbf{W})] + \mathbb{E}[\log p(\mathbf{W})] - \mathbb{E}[\log q(S)] \\ - \mathbb{E}[\log q(\mathbf{b})] - \mathbb{E}[\log q(\tilde{\mathbf{T}})] - \mathbb{E}[\log q(\mathbf{W})]\,.\end{aligned} \tag{B.1}$$

The values of the different expectations are reported bellow.

Appendix B.1.1. Expectations involving the scale factors.

We first focus on the expectations involving the scale factors τ and z. $\mathbb{E}[\log p(\tau_n^p)]$ is given in Tab. B.6 and $\mathbb{E}[\log q(\tau_n^p)]$ is given in Tab. B.7.

For the GDP likelihood, there is the additional latent variable z. Expectations involving this scale factor are given below.

$$\begin{aligned}\mathbb{E}[\log p(z_n^p|\nu_p)] = (\nu_p - 1)\mathbb{E}[\log z_n^p] + \nu_p\log\nu_p \\ - \log\Gamma(\nu_p) - \nu_p\mathbb{E}[z_n^p]\,.\end{aligned} \tag{B.8}$$

$$\begin{aligned}\mathbb{E}[\log q(z_n^p)] = \frac{K+\nu_p+1}{2}\log\mathcal{T}_n^p + (K+\nu_p)\mathbb{E}[\log z_n^p] \\ - \nu_p\mathbb{E}[z_n^p] - \frac{\mathcal{T}_n^p}{2}\mathbb{E}[(z_n^p)^2] - \log\Gamma(K+\nu_p+1) \\ - \frac{\nu_p^2}{4\mathcal{T}_n^p} - \log D_{-K-\nu_p-1}\left(\frac{\nu_p}{\sqrt{\mathcal{T}_n^p}}\right),\end{aligned} \tag{B.9}$$

where $\mathcal{T}_n^p = \mathbb{E}\left[\frac{1}{\tau_n^p}\right]$. Appendix A explains how to compute $\mathbb{E}[z_n^p]$ and $\mathbb{E}[(z_n^p)^2]$. The expectation $\mathbb{E}[\log z_n^p]$ vanishes when computing $\mathbb{E}[\log p(\tau_n^p)] + \mathbb{E}[\log p(z_n^p)] - \mathbb{E}[\log q(z_n^p)]$ with Eq. B.4, B.8 and B.9 and it does not need to be evaluated in practice. Other expectations are given in Appendix C.

Appendix B.1.2. Remaining expectations.

$$\begin{aligned}\mathbb{E}[\log p(\tilde{\mathbf{D}}|\tilde{\mathbf{T}},\tau,\mathbf{b})] = \sum_{n=1}^{N}\sum_{p=1}^{P}\Bigg(-\frac{1}{2}\log|\boldsymbol{\Sigma}_p| + \frac{K}{2}\mathbb{E}[\log\tau_n^p] \\ - \frac{1}{2}\Big[(\tilde{\mathbf{D}}_n^p - \mu_{\tilde{\mathbf{T}}_n} - \mu_{\mathbf{b}_p})^T\mathbb{E}[\tau_n^p]\boldsymbol{\Sigma}_p^{-1}(\tilde{\mathbf{D}}_n^p - \mu_{\tilde{\mathbf{T}}_n} - \mu_{\mathbf{b}_p}) \\ + \mathbb{E}[\tau_n^p]\Big(\operatorname{Tr}(\boldsymbol{\Sigma}_p^{-1}\boldsymbol{\Sigma}_{\mathbf{b}_p}) + \operatorname{Tr}(\boldsymbol{\Sigma}_p^{-1}\boldsymbol{\Sigma}_{\tilde{\mathbf{T}}_n})\Big)\Big]\Bigg) + cst\,.\end{aligned} \tag{B.10}$$

$$\begin{aligned}\mathbb{E}[\log p(\tilde{\mathbf{T}}|\mathbf{W})] = \sum_{n=1}^{N}\Bigg(-\frac{K}{2}\log\boldsymbol{\Sigma}_T \\ - \frac{1}{2}\Big[(\mu_{\tilde{\mathbf{T}}_n} - \mu_{\mathbf{W}}\boldsymbol{\Phi}_n)^T\boldsymbol{\Sigma}_T^{-1}\mathbf{I}_K(\mu_{\tilde{\mathbf{T}}_n} - \mu_{\mathbf{W}}\boldsymbol{\Phi}_n) \\ + \boldsymbol{\Sigma}_T^{-1}\operatorname{Tr}(\boldsymbol{\Sigma}_{\tilde{\mathbf{T}}_n}) + \boldsymbol{\Sigma}_T^{-1}\sum_{k=1}^{K}\operatorname{Tr}(\boldsymbol{\Phi}_n\boldsymbol{\Phi}_n^T\boldsymbol{\Sigma}_{\mathbf{W}_k})\Big]\Bigg) + cst\,.\end{aligned} \tag{B.11}$$

$$\mathbb{E}[\log p(\mathbf{b})] = \sum_{p=1}^{P}\frac{K}{2}\log\beta - \frac{\beta}{2}\left[\mu_{\mathbf{b}_p}^T\mu_{\mathbf{b}_p} + \operatorname{Tr}(\boldsymbol{\Sigma}_{\mathbf{b}_p})\right] + cst\,. \tag{B.12}$$

$$\mathbb{E}[\log p(\mathbf{W})] = \sum_{k=1}^{K}\frac{L}{2}\log\alpha - \frac{\alpha}{2}\left[\mu_{\mathbf{W}_k}^T\mu_{\mathbf{W}_k} + \operatorname{Tr}(\boldsymbol{\Sigma}_{\mathbf{W}_k})\right] + cst\,. \tag{B.13}$$

Finally, $\mathbb{E}[\log q(\tilde{\mathbf{T}})]$, $\mathbb{E}[\log q(\mathbf{b})]$ and $\mathbb{E}[\log q(\mathbf{W})]$ are given by:

$$\mathbb{E}[\log q(\tilde{\mathbf{T}})] = \sum_{n=1}^{N} -\frac{1}{2}\log|\boldsymbol{\Sigma}_{\tilde{\mathbf{T}}_n}| + cst\,, \tag{B.14}$$

$$\mathbb{E}[\log q(\mathbf{b}_p)] = \sum_{p=1}^{P} -\frac{1}{2}\log|\boldsymbol{\Sigma}_{\mathbf{b}_p}| + cst\,, \tag{B.15}$$

$$\mathbb{E}[\log q(\mathbf{W})] = \sum_{k=1}^{K} -\frac{1}{2}\log|\boldsymbol{\Sigma}_{\mathbf{W}_k}| + cst\,. \tag{B.16}$$

Appendix B.2. Mixture of consensuses.

The lower bound for the mixture of consensuses model can be written as follows:

$$\begin{aligned}\mathcal{L}(q) = \mathbb{E}[\log p(\tilde{\mathbf{D}}|\tilde{\mathbf{T}},Z)] + \mathbb{E}[\log p(Z)] \\ - \mathbb{E}[\log q(\tilde{\mathbf{T}})] - \mathbb{E}[\log q(Z)]\,.\end{aligned} \tag{B.17}$$

Table B.6. Formula giving $\mathbb{E}[\log p(\tau_n^p)]$ for the three heavy-tailed likelihoods. The values of the constants for the Laplace and GDP distributions are given by $C_{\mathbf{L}} = -\frac{K+1}{2}\log 8 - \log\Gamma\left(\frac{K+1}{2}\right)$ and $C_{\mathbf{GDP}} = -\frac{K+1}{2}\log 2 - \log\Gamma\left(\frac{K+1}{2}\right)$, respectively.

Likelihood	$\mathbb{E}[\log p(\tau_n^p)]$	
Student's t	$-\log\Gamma\left(\frac{\nu_p}{2}\right) + \frac{\nu_p}{2}\log\frac{\nu_p}{2} + \left(\frac{\nu_p}{2}-1\right)\mathbb{E}[\log\tau_n^p] - \frac{\nu_p}{2}\mathbb{E}[\tau_n^p]$	(B.2)
Laplace	$-\frac{K+3}{2}\mathbb{E}[\log\tau_n^p] - \frac{1}{8}\mathbb{E}\left[\frac{1}{\tau_n^p}\right] + C_{\mathrm{L}}$	(B.3)
GDP	$(K+1)\mathbb{E}[\log z_n^p] - \frac{K+3}{2}\mathbb{E}[\log\tau_n^p] - \frac{1}{2}\mathbb{E}[(z_n^p)^2]\mathbb{E}\left[\frac{1}{\tau_n^p}\right] + C_{\mathrm{GDP}}$	(B.4)

Table B.7. Formula giving $\mathbb{E}[\log q(\tau_n^p)]$ for the three heavy-tailed likelihoods. a_{np} and b_{np} are given in Tab. 2.

Likelihood	$\mathbb{E}[\log q(\tau_n^p)]$	
Student's t	$-\log\Gamma(a_{np}) + a_{np}\log b_{np} + (a_{np}-1)\mathbb{E}[\log\tau_n^p] - b_{np}\mathbb{E}[\tau_n^p]$	(B.5)
Laplace	$-\frac{3}{2}\mathbb{E}[\log\tau_n^p] - \frac{1}{2}\log(8\pi) - \frac{1}{2}$	(B.6)
GDP	$\frac{1}{2}\log\frac{\mathbb{E}[(z_n^p)^2]}{2\pi} - \frac{3}{2}\mathbb{E}[\log\tau_n^p] - \frac{1}{2}$	(B.7)

Developing each term, we obtain:

$$\mathbb{E}[\log p(\tilde{\mathbf{D}}|\tilde{\mathbf{T}}, Z)] = \sum_{n=1}^{N}\sum_{m=1}^{M}\sum_{p=1}^{P} r_{pm}\Big[-\frac{K}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{\Sigma}_p| - \frac{1}{2}\Big((\tilde{\mathbf{D}}_n^p - \mu_{\tilde{\mathbf{T}}_{nm}})^T\mathbf{\Sigma}_p^{-1}(\tilde{\mathbf{D}}_n^p - \mu_{\tilde{\mathbf{T}}_{nm}}) + \mathrm{Tr}(\mathbf{\Sigma}_p^{-1}\mathbf{\Sigma}_{\tilde{\mathbf{T}}_{nm}})\Big)\Big], \tag{B.18}$$

$$\mathbb{E}[\log p(Z)] = \sum_{p=1}^{P}\sum_{m=1}^{M} r_{pm}\log\pi_m\,, \tag{B.19}$$

$$\mathbb{E}[\log q(\tilde{\mathbf{T}})] = \sum_{n=1}^{N}\sum_{m=1}^{M} -\frac{1}{2}\log|\mathbf{\Sigma}_{\tilde{\mathbf{T}}_{nm}}| + cst\,, \tag{B.20}$$

$$\mathbb{E}[\log q(Z)] = \sum_{p=1}^{P}\sum_{m=1}^{M} r_{pm}\log r_{pm}\,. \tag{B.21}$$

Appendix C. Additional expectations.

In this last section, we gather together some useful expectations involved in the variational updates or in the evaluation of the lower bound.

Appendix C.1. Robust probabilistic model.

$$\mathbb{E}[\mathbf{b}_p] = \mu_{\mathbf{b}_p}\,. \tag{C.1}$$

$$\mathbb{E}[\tilde{\mathbf{T}}_n] = \mu_{\tilde{\mathbf{T}}_n}\,. \tag{C.2}$$

$$\mathbb{E}[\tilde{\mathbf{T}}_{nk}] = \mu_{\tilde{\mathbf{T}}_{nk}}\,. \tag{C.3}$$

$$\mathbb{E}[\mathbf{W}_k] = \mu_{\mathbf{W}_k}\,. \tag{C.4}$$

$$\mathbb{E}[(\tilde{\mathbf{D}}_n^p - \tilde{\mathbf{T}}_n - \mathbf{b}_p)^T\mathbf{\Sigma}_p^{-1}(\tilde{\mathbf{D}}_n^p - \tilde{\mathbf{T}}_n - \mathbf{b}_p)] = (\tilde{\mathbf{D}}_n^p - \mu_{\tilde{\mathbf{T}}_n} - \mu_{\mathbf{b}_p})^T\mathbf{\Sigma}_p^{-1}(\tilde{\mathbf{D}}_n^p - \mu_{\tilde{\mathbf{T}}_n} - \mu_{\mathbf{b}_p}) + \mathrm{Tr}(\mathbf{\Sigma}_p^{-1}\mathbf{\Sigma}_{\mathbf{b}_p}) + \mathrm{Tr}(\mathbf{\Sigma}_p^{-1}\mathbf{\Sigma}_{\tilde{\mathbf{T}}_n})\,. \tag{C.5}$$

Moreover, $\mathbb{E}[\mathbf{W}]$ corresponds to the gathering of the expectations $\mathbb{E}[\mathbf{W}_k]$ given above in a matrix of size $K \times L$.

Regarding the scale factor τ, we need the expectations $\mathbb{E}[\tau_n^p]$ and $\mathbb{E}[\log\tau_n^p]$ for the Student's t distribution. The latter is notably involved in the estimation of the degree of freedom. There are given by:

$$\mathbb{E}[\tau_n^p] = \frac{a_{np}}{b_{np}}\,, \tag{C.6}$$

$$\mathbb{E}[\log\tau_n^p] = \psi(a_{np}) - b_{np}\,, \tag{C.7}$$

where a_{np} and b_{np} are the parameters of the Gamma distribution described in Tab. 2 and ψ is the digamma function. For the Laplace and GDP distributions, we have to evaluate $\mathbb{E}[\tau_n^p]$ and $\mathbb{E}\left[\frac{1}{\tau_n^p}\right]$. They can be written as follows:

$$\mathbb{E}[\tau_n^p] = \mu_{np}\,, \tag{C.8}$$

$$\mathbb{E}\left[\frac{1}{\tau_n^p}\right] = \frac{1}{\mu_{np}} + \frac{1}{\lambda_{np}}\,, \tag{C.9}$$

where μ_{np} and λ_{np} are the parameters of the inverse Gaussian distributions given in Tab. 2. Moreover, a third expectation,

$\mathbb{E}[\log \tau_n^p]$, appears in some terms of the lower bound. In contrast to the Gaussian case, it does not have a closed-form formula for the Laplace or GDP distributions. However, this is not a problem in practice as it vanishes when gathering the different parts, in particular after summation of Eqs. B.3, B.6 and B.10 for the Laplace distribution, and summation of Eqs. B.4, B.7 and B.10 for the GDP distribution.

Appendix C.2. Mixture of consensuses.

$$\mathbb{E}[(\tilde{\mathbf{D}}_n^p - \tilde{\mathbf{T}}_{nm})^T \boldsymbol{\Sigma}_p^{-1} (\tilde{\mathbf{D}}_n^p - \tilde{\mathbf{T}}_{nm})] = (\tilde{\mathbf{D}}_n^p - \mu_{\tilde{\mathbf{T}}_{nm}})^T \boldsymbol{\Sigma}_p^{-1} (\tilde{\mathbf{D}}_n^p - \mu_{\tilde{\mathbf{T}}_{nm}}) + \mathrm{Tr}(\boldsymbol{\Sigma}_p^{-1} \boldsymbol{\Sigma}_{\tilde{\mathbf{T}}_{nm}}) \,. \tag{C.10}$$

References

Akhondi-Asl, A., Hoyte, L., Lockhart, M.E., Warfield, S.K., 2014. A Logarithmic Opinion Pool Based STAPLE Algorithm for the Fusion of Segmentations With Associated Reliability Weights. IEEE Transactions on Medical Imaging 33, 1997–2009.

Akhondi-Asl, A., Warfield, S.K., 2013. Simultaneous Truth and Performance Level Estimation Through Fusion of Probabilistic Segmentations. IEEE Transactions on Medical Imaging 32, 1840–1852.

Archambeau, C., Verleysen, M., 2007. Robust Bayesian clustering. Neural Networks 20, 129–138.

Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., et al., 2011. The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans. Medical Physics 38, 915–931.

Arslan, O., 2004. Family of multivariate generalized t distributions. Journal of Multivariate Analysis 89, 329–337.

Asman, A.J., Landman, B.A., 2011. Robust Statistical Label Fusion Through Consensus Level, Labeler Accuracy, and Truth Estimation (COLLATE). IEEE Transactions on Medical Imaging 30, 1779–1794.

Asman, A.J., Landman, B.A., 2012. Formulating Spatially Varying Performance in the Statistical Fusion Framework. IEEE Transactions on Medical Imaging 31, 1326–1336.

Asman, A.J., Landman, B.A., 2013. Non-local statistical label fusion for multi-atlas segmentation. Medical Image Analysis 17, 194 – 208.

Audelan, B., Hamzaoui, D., Montagne, S., Renard-Penna, R., Delingette, H., 2020. Robust Fusion of Probability Maps, in: Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, Springer International Publishing, Cham. pp. 259–268.

Babacan, S.D., Molina, R., Katsaggelos, A.K., 2008. Parameter Estimation in TV Image Restoration Using Variational Distribution Approximation. IEEE Transactions on Image Processing 17, 326–339.

Blei, D.M., Kucukelbir, A., McAuliffe, J.D., 2017. Variational Inference: A Review for Statisticians. Journal of the American Statistical Association 112, 859–877.

Bouveyron, C., Girard, S., Schmid, C., 2007. High-dimensional data clustering. Computational Statistics & Data Analysis 52, 502–519.

Commowick, O., Akhondi-Asl, A., Warfield, S.K., 2012. Estimating A Reference Standard Segmentation With Spatially Varying Performance Parameters: Local MAP STAPLE. IEEE Transactions on Medical Imaging 31, 1593–1606.

Commowick, O., Warfield, S.K., 2009. A Continuous STAPLE for Scalar, Vector, and Tensor Images: An Application to DTI Analysis. IEEE Transactions on Medical Imaging 28, 838–846.

Commowick, O., Warfield, S.K., 2010. Incorporating Priors on Expert Performance Parameters for Segmentation Validation and Label Fusion: A Maximum a Posteriori STAPLE, in: Proceedings of the 13th International Conference on Medical Image Computing and Computer-Assisted Intervention: Part III, Springer-Verlag, Berlin, Heidelberg. p. 25–32.

Fenster, A., Chiu, B., 2005. Evaluation of Segmentation algorithms for Medical Imaging, in: 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, pp. 7186–7189.

Giri, R., 2016. Bayesian sparse signal recovery using scale mixtures with applications to speech. Ph.D. thesis. UC San Diego.

Gómez, E., Gomez-Viilegas, M., Marín, J., 1998. A multivariate generalization of the power exponential family of distributions. Communications in Statistics - Theory and Methods 27, 589–600.

Gómez-Sánchez-Manzano, E., Gómez-Villegas, M.A., Marín, J.M., 2008. Multivariate Exponential Power Distributions as Mixtures of Normal Distributions with Bayesian Applications. Communications in Statistics - Theory and Methods 37, 972–985.

Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: a tutorial (with comments by M. Clyde, David Draper and E. I. George, and a rejoinder by the authors. Statistical Science 14, 382 – 417.

Joskowicz, L., Cohen, D., Caplan, N., Sosna, J., 2019. Inter-observer variability of manual contour delineation of structures in CT. European Radiology 29, 1391–1399.

Kocak, B., Durmaz, E.S., Kaya, O.K., Ates, E., Kilickesmez, O., 2019. Reliability of Single-Slice–Based 2D CT Texture Analysis of Renal Masses: Influence of Intra- and Interobserver Manual Segmentation Variability on Radiomic Feature Reproducibility. American Journal of Roentgenology 213, 377–383.

Lampert, T.A., Stumpf, A., Gançarski, P., 2016. An Empirical Study Into Annotator Agreement, Ground Truth Estimation, and Algorithm Evaluation. IEEE Transactions on Image Processing 25, 2557–2572.

Landman, B.A., Asman, A.J., Scoggins, A.G., Bogovic, J.A., Xing, F., Prince, J.L., 2012. Robust Statistical Fusion of Image Labels. IEEE Transactions on Medical Imaging 31, 512–522.

Langerak, T.R., van der Heide, U.A., Kotte, A.N.T.J., et al., 2010. Label Fusion in Atlas-Based Segmentation Using a Selective and Iterative Method for Performance Level Estimation (SIMPLE). IEEE Transactions on Medical Imaging 29, 2000–2008.

Litjens, G., Debats, O., Barentsz, J., Karssemeijer, N., Huisman, H., 2014. Computer-Aided Detection of Prostate Cancer in MRI. IEEE Transactions on Medical Imaging 33, 1083–1092.

Liu, X., Montillo, A., Tan, E.T., Schenck, J.F., 2013. iSTAPLE: improved label fusion for segmentation by combining STAPLE with image intensity, in: Ourselin, S., Haynor, D.R. (Eds.), Medical Imaging 2013: Image Processing, International Society for Optics and Photonics. SPIE. pp. 727 – 732.

McDonald, J.B., Newey, W.K., 1988. Partially Adaptive Estimation of Regression Models via the Generalized t Distribution. Econometric Theory 4, 428–457.

Menze, B.H., Jakab, A., Bauer, S., et al., 2015. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE Transactions on Medical Imaging 34, 1993–2024.

Meyer, A., Rakr, M., Schindele, D., Blaschke, S., Schostak, M., Fedorov, A., Hansen, C., 2019. Towards Patient-Individual PI-Rads v2 Sector Map: Cnn for Automatic Segmentation of Prostatic Zones From T2-Weighted MRI, in: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), pp. 696–700.

Miller, A.C., Foti, N.J., Adams, R.P., 2017. Variational Boosting: Iteratively Refining Posterior Approximations, in: Precup, D., Teh, Y.W. (Eds.), Proceedings of the 34th International Conference on Machine Learning, PMLR, International Convention Centre, Sydney, Australia. pp. 2420–2429.

Neville, S.E., 2013. Elaborate distribution semiparametric regression via mean field variational Bayes. Ph.D. thesis. University of Wollongong.

Park, T., Casella, G., 2008. The Bayesian Lasso. Journal of the American Statistical Association 103, 681–686.

Pascal, F., Bombrun, L., Tourneret, J., Berthoumieu, Y., 2013. Parameter Estimation For Multivariate Generalized Gaussian Distributions. IEEE Transactions on Signal Processing 61, 5960–5971.

Petersen, R.P., Truong, P.T., Kader, H.A., Berthelet, E., Lee, J.C., et al., 2007. Target Volume Delineation for Partial Breast Radiotherapy Planning: Clinical Characteristics Associated with Low Interobserver Concordance. International Journal of Radiation Oncology*Biology*Physics 69, 41–48.

Pohl, K.M., Fisher, J., Bouix, S., Shenton, M., McCarley, R.W., Grimson, W.E.L., Kikinis, R., Wells, W.M., 2007. Using the logarithm of odds to define a vector space on probabilistic atlases. Medical Image Analysis 11, 465–477. URL: https://www.sciencedirect.com/science/article/pii/S1361841507000576, doi:https://doi.org/10.1016/j.media.2007.06.003. special Issue on the Ninth International Conference on Medical Image Computing and Computer-Assisted Interventions - MICCAI 2006.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation, in: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer International Publishing, Cham. pp. 234–241.

Sabuncu, M.R., Yeo, B.T.T., Van Leemput, K., Fischl, B., Golland, P., 2010. A Generative Model for Image Segmentation Based on Label Fusion. IEEE Transactions on Medical Imaging 29, 1714–1729.

Sadeghigol, Z., Kahaei, M.H., Haddadi, F., 2016. Model based variational Bayesian compressive sensing using heavy tailed sparse prior. Signal Processing: Image Communication 41, 158–167.

Taha, A.A., Hanbury, A., 2015. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Medical Imaging 15, 29.

Tang, Y., Gao, R., Lee, H.H., Han, S., Chen, Y., et al., 2021. High-resolution 3d abdominal segmentation with random patch network fusion. Medical Image Analysis 69, 101894.

Wang, Y., Zhou, Y., Shen, W., Park, S., Fishman, E.K., Yuille, A.L., 2019. Abdominal multi-organ segmentation with organ-attention networks and statistical fusion. Medical Image Analysis 55, 88–102.

Warfield, S.K., Zou, K.H., Wells, W.M., 2002. Validation of Image Segmentation and Expert Quality with an Expectation-Maximization Algorithm, in: Dohi, T., Kikinis, R. (Eds.), Medical Image Computing and Computer-Assisted Intervention — MICCAI 2002, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 298–306.

Warfield, S.K., Zou, K.H., Wells, W.M., 2004. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Transactions on Medical Imaging 23, 903–921.

Warfield, S.K., Zou, K.H., Wells, W.M., 2008. Validation of image segmentation by estimating rater bias and variance. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 366, 2361–2375.

Xing, F., Prince, J.L., Landman, B.A., 2016. Investigation of Bias in Continuous Medical Image Label Fusion. PLOS ONE 11, 1–15.