A Cautionary Analysis of STAPLE Using Direct Inference of Segmentation Truth

Koen Van Leemput[1,2] and Mert R. Sabuncu[1]

[1] A.A. Martinos Center for Biomedical Imaging, MGH, Harvard Medical School, USA
[2] Technical University of Denmark, Denmark

Abstract. In this paper we analyze the properties of the well-known segmentation fusion algorithm STAPLE, using a novel inference technique that analytically marginalizes out all model parameters. We demonstrate both theoretically and empirically that when the number of raters is large, or when consensus regions are included in the model, STAPLE devolves into thresholding the average of the input segmentations. We further show that when the number of raters is small, the STAPLE result may not be the optimal segmentation truth estimate, and its model parameter estimates might not reflect the individual raters' actual segmentation performance. Our experiments indicate that these intrinsic weaknesses are frequently exacerbated by the presence of undesirable global optima and convergence issues. Together these results cast doubt on the soundness and usefulness of typical STAPLE outcomes.

1 Introduction

Since its introduction a decade ago, the Simultaneous Truth and Performance Level Estimation (STAPLE) algorithm [1] has become an established technique for estimating the true underlying segmentation of a structure from multiple imperfect segmentations. Its applications range from combining manual delineations by different human expert raters, to fusing registration-based automatic segmentations in multi-atlas label fusion methods. The algorithm is based on an explicit probabilistic model of how an (unknown) true segmentation degrades into (observed) imperfect segmentations, allowing for different frequencies of segmentation errors by different raters. Starting from its basic formulation, it has since been extended in several directions, including accounting for spatially-varying rater performance [2,3], putting "error bars" on estimated rater performance measures [4], and modeling missing or repeat segmentations [5].

A detailed theoretical understanding of the results produced by STAPLE and its variants is hampered by the fact that the algorithm depends on a numerical optimization procedure for the parameters of its model. Here, we show that this optimization procedure can actually be avoided, since the parameters can be marginalized out analytically. This allows us to theoretically predict the behavior of STAPLE and its variants in several often-used scenarios – which we empirically verify – revealing several undesirable properties.

P. Golland et al. (Eds.): MICCAI 2014, Part I, LNCS 8673, pp. 398–406, 2014.

To the best of our knowledge, this is the first detailed analysis of the theoretical properties of STAPLE, although it has been criticized before in empirical studies (e.g., [6]).

2 Theoretical Analysis

2.1 The STAPLE Model

Let $\mathbf{d}^j = (d_1^j, \ldots, d_I^j)^T$ denote the segmentation of a structure by rater j, where $d_i^j \in \{0, 1\}$ is the one of two possible labels assigned to voxel i, and I is the total number of voxels. Given J raters, the collection of all segmentations is given by $\mathbf{D} = (\mathbf{d}^1 \ldots \mathbf{d}^J)$. Letting $\mathbf{t} = (t_1, \ldots, t_I)^T$ with $t_i \in \{0, 1\}$ denote the true underlying structure, the STAPLE algorithm is based on the following generative model for the observed segmentations:

$$p(\mathbf{t}|\boldsymbol{\pi}) = \prod_i \pi_{t_i} \quad \text{and} \quad p(\mathbf{D}|\mathbf{t}, \boldsymbol{\theta}) = \prod_j p(\mathbf{d}^j|\mathbf{t}, \boldsymbol{\theta}^j), \quad p(\mathbf{d}^j|\mathbf{t}, \boldsymbol{\theta}^j) = \prod_i \theta^j_{d_i^j, t_i}.$$

Here the vector $\boldsymbol{\pi} = (\pi_0, \pi_1)^T$, with $0 \leq \pi_t \leq 1$ and $\sum_t \pi_t = 1$, contains the expected frequencies of occurrence for each label t. Furthermore, $\theta^j_{d,t}$ denotes the probability that rater j assigns label d to a voxel if the true label is t, so that $\sum_d \theta^j_{d,t} = 1$. Finally, the vector $\boldsymbol{\theta}^j = ({\boldsymbol{\theta}_0^j}^T, {\boldsymbol{\theta}_1^j}^T)^T$ with $\boldsymbol{\theta}_t^j = (\theta^j_{0,t}, \theta^j_{1,t})^T$ collects the segmentation performance parameters of rater j, and the vector $\boldsymbol{\theta} = ({\boldsymbol{\theta}^1}^T, \ldots, {\boldsymbol{\theta}^J}^T)^T$ collects all performance parameters of all raters.

Letting the vector $\boldsymbol{\omega} = (\boldsymbol{\theta}^T, \boldsymbol{\pi}^T)^T$ denote all the parameters of the model, a prior of the form

$$p(\boldsymbol{\omega}) = p(\boldsymbol{\pi})p(\boldsymbol{\theta}), \quad p(\boldsymbol{\pi}) \propto \pi_0^{\alpha_0}\pi_1^{\alpha_1}, \quad p(\boldsymbol{\theta}) \propto \prod_j \prod_t (\theta^j_{0,t})^{\alpha_{0,t}} (\theta^j_{1,t})^{\alpha_{1,t}}$$

is often used, where $\{\alpha_t, \alpha_{d,t}\}$ are hyperparameters whose values are assumed given. By selecting hyperparameter values $\alpha_{d,t} = 0, \forall d, t$ (implying that all values of $\theta^j_{d,t}$ are equally likely) and $\alpha_t = \rho \sum_j \sum_i [d_i^j = t], \rho \to \infty$ (effectively clamping the values of π_t to the average frequency of occurrence in the raters' segmentations), the original STAPLE model [1] is obtained. Alternatively, $\alpha_{d,t}$ can be set to specific positive values to encode an expectation of better-than-random segmentation performance [2,5]; and by setting $\alpha_t = 0, \forall t$ the expected frequencies of occurrence π_t can automatically be inferred from the data [5].

2.2 STAPLE Inference

Given the collection of available segmentations $\mathbf{D}$, the STAPLE algorithm first seeks the maximum a posteriori (MAP) estimate *of the model parameters*:

$$\widehat{\boldsymbol{\omega}} = \arg\max_{\boldsymbol{\omega}} p(\boldsymbol{\omega}|\mathbf{D}) = \arg\max_{\boldsymbol{\theta}, \boldsymbol{\pi}} \prod_i \left(\sum_{t_i} \left(\prod_j \theta^j_{d_i^j, t_i} \right) \pi_{t_i} \right) p(\boldsymbol{\theta})p(\boldsymbol{\pi}), \qquad (1)$$

using a dedicated expectation-maximization (EM) optimizer that exploits the specific structure of Eq. (1). Once a MAP parameter estimate $\widehat{\boldsymbol{\omega}}$ is found, it is

then used to infer the segmentation truth as: $\hat{\mathbf{t}}_{STAPLE} = \arg\max_{\mathbf{t}} p(\mathbf{t}|\mathbf{D}, \widehat{\boldsymbol{\omega}})$, which involves only voxel-wise binary decisions [1]. In some cases one is also interested in the performance parameters $\boldsymbol{\theta}_t^j$ of individual raters. In that scenario, the corresponding component $\hat{\boldsymbol{\theta}}_t^j$ is simply extracted from the high-dimensional parameter vector $\widehat{\boldsymbol{\omega}}$ and inspected [1]. In addition, "error bounds" around these values are sometimes computed by locally approximating the posterior $p(\boldsymbol{\omega}|\mathbf{D})$ using a Gaussian distribution and estimating its covariance structure [4].

2.3 STAPLE Inference as an Approximation

Given the raters' segmentations $\mathbf{D}$, the MAP estimate *of the segmentation truth* is given by $\hat{\mathbf{t}} = \arg\max_{\mathbf{t}} p(\mathbf{t}|\mathbf{D})$, which will generally be different from the STAPLE result $\hat{\mathbf{t}}_{STAPLE}$ obtained by maximizing $p(\mathbf{t}|\mathbf{D}, \widehat{\boldsymbol{\omega}})$. Furthermore, the distribution $p(\boldsymbol{\theta}_t^j|\mathbf{D})$ is a low-dimensional projection of a higher dimensional distribution $p(\boldsymbol{\omega}|\mathbf{D})$; its properties – including its maximum – cannot generally be inferred by simply ignoring all other components in $\boldsymbol{\omega}$.

The crux of this paper is that the STAPLE inference of both the segmentation truth and performance parameters will only be a good approximation when the distribution $p(\mathbf{t}|\mathbf{D})$ is strongly peaked around its optimal value $\hat{\mathbf{t}}$. To understand this, it is instructive to write out the parameter posterior as:

$$p(\boldsymbol{\omega}|\mathbf{D}) = \sum_{\mathbf{t}} p(\boldsymbol{\omega}|\mathbf{t}, \mathbf{D}) p(\mathbf{t}|\mathbf{D}) \tag{2}$$

with

$$p(\boldsymbol{\omega}|\mathbf{t}, \mathbf{D}) = \frac{p(\mathbf{t}, \mathbf{D}|\boldsymbol{\omega}) p(\boldsymbol{\omega})}{p(\mathbf{t}, \mathbf{D})} = p(\boldsymbol{\pi}|\mathbf{t}) \prod_j \prod_t p(\boldsymbol{\theta}_t^j|\mathbf{t}, \mathbf{D}), \tag{3}$$

where $p(\boldsymbol{\pi}|\mathbf{t})$ and $p(\boldsymbol{\theta}_t^j|\mathbf{t}, \mathbf{D})$ are beta distributions. The last step in Eq. (3) is based on the fact that the normalizer

$$p(\mathbf{t}, \mathbf{D}) = \int_{\boldsymbol{\omega}} p(\mathbf{t}, \mathbf{D}|\boldsymbol{\omega}) p(\boldsymbol{\omega}) \mathrm{d}\boldsymbol{\omega} = \int_{\boldsymbol{\theta}} p(\mathbf{D}|\mathbf{t}, \boldsymbol{\theta}) p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \int_{\boldsymbol{\pi}} p(\mathbf{t}|\boldsymbol{\pi}) p(\boldsymbol{\pi}) \mathrm{d}\boldsymbol{\pi} \quad \propto$$

$$\mathrm{B}(N_0 + \alpha_0 + 1, N_1 + \alpha_1 + 1) \left(\prod_j \prod_t \mathrm{B}\big(N_{0,t}^j + \alpha_{0,t} + 1, N_{1,t}^j + \alpha_{1,t} + 1\big) \right) \tag{4}$$

involves a marginalization over the model parameters that is given in analytical form. Here $\mathrm{B}(\cdot, \cdot)$ denotes the beta function, $N_{d,t}^j$ the number of voxels assigned to label d by rater j when the truth label in $\mathbf{t}$ is t, and N_t the total number of voxels where the truth label is t.

Referring to Eq. (2), the posterior $p(\boldsymbol{\omega}|\mathbf{D})$ is obtained by summing conditional posteriors $p(\boldsymbol{\omega}|\mathbf{t}, \mathbf{D})$, one for each possible $\mathbf{t}$ and weighed according to how probable that $\mathbf{t}$ is. When $p(\mathbf{t}|\mathbf{D})$ is strongly peaked around $\hat{\mathbf{t}}$, the resulting summation will be dominated by the contribution of $\hat{\mathbf{t}}$ only: $p(\boldsymbol{\omega}|\mathbf{D}) \simeq p(\boldsymbol{\omega}|\hat{\mathbf{t}}, \mathbf{D})$. As Eq. (3) shows, this distribution factorizes across the individual parameter components, so that the location of the maximum of $p(\boldsymbol{\theta}_t^j|\mathbf{D})$ can indeed be obtained by extracting the corresponding component from the joint maximum location $\hat{\boldsymbol{\omega}}$. Furthermore, since the individual factors are (narrow) beta distributions, their product will be strongly peaked around $\hat{\boldsymbol{\omega}}$, so that in turn $p(\mathbf{t}|\mathbf{D}) = \int_{\boldsymbol{\omega}} p(\mathbf{t}|\mathbf{D}, \boldsymbol{\omega}) p(\boldsymbol{\omega}|\mathbf{D}) \mathrm{d}\boldsymbol{\omega} \simeq p(\mathbf{t}|\mathbf{D}, \widehat{\boldsymbol{\omega}})$ and therefore $\hat{\mathbf{t}}_{STAPLE} \simeq \hat{\mathbf{t}}$.

2.4 Two Cases Where STAPLE Inference Will Be Accurate

There are two common scenarios, analyzed below, where $p(\mathbf{t}|\mathbf{D})$ is sharply peaked around $\hat{\mathbf{t}}$ and where STAPLE therefore provides accurate inference. However, as we shall see, both scenarios will also render the STAPLE results akin to simply thresholding the average segmentation map $\bar{\mathbf{d}} = \sum_j \mathbf{d}^j/J$ – similar to majority voting which thresholds $\bar{\mathbf{d}}$ at level 0.5.

We start by writing the conditional posterior distribution of the segmentation truth label in a single voxel. Because of Eq. (4), we have that

$$p(t_i|\mathbf{D}, \mathbf{t}_{\setminus i}) \propto \left(\prod_j \frac{N^j_{d_i^j, t_{i\setminus i}} + \alpha_{d_i^j, t_i} + 1}{\sum_d (N^j_{d, t_{i\setminus i}} + \alpha_{d,t_i} + 1)} \right) (N_{t_{i\setminus i}} + \alpha_t + 1), \tag{5}$$

where $\mathbf{t}_{\setminus i}$ denotes the truth labels in all voxels except voxel i, and $N^j_{d,t_{\setminus i}}$ and $N_{t_{\setminus i}}$ are the corresponding voxel counts.

Many Raters: When the number of raters is very large ($J \gg 0$), we have that, around the optimum $\hat{\mathbf{t}}$, the log of the ratio of conditional posterior probabilities in a voxel i behaves approximately as a simple linear function of the fraction of raters that assigned voxel i to foreground, denoted by $f_i = \sum_j d_i^j/J$:

$$\log \left(\frac{p(t_i = 1|\mathbf{D}, \hat{\mathbf{t}}_{\setminus i})}{p(t_i = 0|\mathbf{D}, \hat{\mathbf{t}}_{\setminus i})} \right) \simeq s_{\hat{\mathbf{t}}} \cdot f_i + o_{\hat{\mathbf{t}}}, \tag{6}$$

with slope $s_{\hat{\mathbf{t}}} = J(\bar{c}_{0,0} + \bar{c}_{1,1} - \bar{c}_{1,0} - \bar{c}_{0,1})$ and offset $o_{\hat{\mathbf{t}}} = J(\bar{c}_{0,1} - \bar{c}_{0,0}) + (c_1 - c_0)$, where $\bar{c}_{d,t} = \frac{\sum_j c^j_{d,t}}{J}$, $c^j_{d,t} = \log \left(\frac{\hat{N}^j_{d,t} + \alpha_{d,t} + 1}{\sum_d (\hat{N}^j_{d,t} + \alpha_{d,t} + 1)} \right)$, and $c_t = \log \left(\hat{N}_t + \alpha_t + 1 \right)$. This is because of the large number of summations involved when $J \gg 0$ (law of large numbers), and because $\hat{N}^j_{d,t_{\setminus i}} \simeq \hat{N}^j_{d,t}$ and $\hat{N}_{t_{\setminus i}} \simeq \hat{N}_t$. When the foreground fraction f_i exceeds a certain threshold, the "log odds" of Eq. (6) becomes positive and the voxel is assigned to label $t_i = 1$, *independent of which raters exactly labeled the voxel as fore- or background*. Furthermore, the same threshold applies to all voxels, since the slope $s_{\hat{\mathbf{t}}}$ and offset $o_{\hat{\mathbf{t}}}$ are independent of i. Note that the slope $s_{\hat{\mathbf{t}}}$ depends directly on J, so that the joint posterior $p(\mathbf{t}|\mathbf{D})$ will be strongly peaked when $J \gg 0$ – therefore $\hat{\mathbf{t}}_{STAPLE}$ can be expected to correspond to a thresholded average segmentation map $\bar{\mathbf{d}}$ (although not necessarily at the threshold level 0.5 used by majority voting).

Large Consensus Regions: Even if the number of raters J is small, Eq. (6) will still be a good approximation when large values for the hyperparameters $\alpha_{d,t}$ are used, as these effectively make the different $c^j_{d,t}$ similar across all raters. This will happen when large "consensus regions" are included in the analysis, i.e., image regions where all raters agree on the same label, since the net effect of such voxels will be to act as large hyperparameters $\alpha_{0,0} \gg 0$ (for background areas) and $\alpha_{1,1} \gg 0$ (for foreground areas) on the remaining, non-consensus voxels. In that specific scenario, $c_{0,1}$ and $c_{1,0}$ will attain large negative values, yielding a large slope $s_{\hat{\mathbf{t}}}$ indicative of a sharply peaked $p(\mathbf{t}|\mathbf{D})$, so that again $\hat{\mathbf{t}}_{STAPLE}$ can be expected to be a thresholded map $\bar{\mathbf{d}}$ in this case as well.

2.5 Direct Inference of Segmentation Truth

In addition to providing theoretical insight, Eq. (5) also suggests a new way of *directly* inferring the segmentation truth using discrete optimization – without estimating continuous model parameters first. In particular, starting from some initial labeling, the MAP segmentation truth $\hat{\mathbf{t}} = \arg\max_{\mathbf{t}} p(\mathbf{t}|\mathbf{D})$ can be estimated by visiting each voxel in turn, assigning it to the label that maximizes Eq. (5), and repeating this procedure until no voxels change labels. A similar procedure can also be used to generate Monte Carlo samples from $p(\mathbf{t}|\mathbf{D})$: by repeatedly visiting each voxel, in random order, and assigning label t_i with probability $p(t_i|\mathbf{D}, \mathbf{t}_{\setminus i})$, a large number S of samples $\{\mathbf{t}^{(s)}\}_{s=1}^{S}$ of the segmentation truth can be obtained (so-called Gibbs sampling). Such samples can then be used to assess the full posterior distribution of specific (combinations of) performance parameters, e.g., $p(\boldsymbol{\theta}_t^j|\mathbf{D}) = \sum_{\mathbf{t}} p(\boldsymbol{\theta}_t^j|\mathbf{t}, \mathbf{D}) p(\mathbf{t}|\mathbf{D}) \simeq \frac{1}{S} \sum_{s=1}^{S} p(\boldsymbol{\theta}_t^j|\mathbf{t}^{(s)}, \mathbf{D})$.

3 Experiments

In order to verify our theoretical analysis, we performed experiments in the context of multi-atlas label fusion, in which a manually annotated brain MR scan of each of 39 subjects was non-linearly warped to the remaining 38 subjects as described in [7]. These warps were applied to the manual segmentations of 10 brain structures (cerebral white matter, cerebral cortex, lateral ventricle, thalamus, caudate, putamen, pallidum, hippocampus, and amygdala in the left hemisphere, as well as brain stem), which were subsequently used as input to a binary STAPLE set-up (treating each structure in turn as foreground and the remaining voxels as background).

We studied three variations of STAPLE. In the original "Basic" variant [1], all voxels in the image are considered; a flat prior on the segmentation performance parameters is used (i.e., $\alpha_{d,t} = 0$); and a pre-computed spatial prior $\boldsymbol{\pi}$ is clamped to the average relative size of the foreground/background in the input segmentations. The "Restricted" variant is identical except that all voxels in which all raters agreed on the same label ("consensus areas") are excluded from the analysis [8]. Finally, the "Advanced" variant also discards all consensus areas, but encourages better-than-random segmentation performance by setting $\alpha_{0,0} = \alpha_{1,1} = 2$ and $\alpha_{0,1} = \alpha_{1,0} = 0$ [5,2]; in addition $\alpha_0 = \alpha_1 = 0$ so that the spatial prior $\boldsymbol{\pi}$ is automatically estimated from the data [5].

As is common in the literature, each variant was initialized with high sensitivity and specificity parameters (we used $\theta_{0,0}^j = \theta_{1,1}^j = 0.99$); for the "Advanced" variant π_1 was initialized as 0.5. To reduce the number of experiments, we only studied a random subset of 20 subjects from the available 39. We ran experiments in 2D, selecting, for each experiment, the coronal slice in which the structure being studied was the largest in the co-registered 3D volumes. To quantify the influence of the number of raters, each experiment was run with the segmentations restricted to the first 5, 15, and 37 of the available 38 ones.

4 Results

Since our analysis predicts thresholding behavior in certain scenarios but not the applicable threshold level (which itself depends on the found solution), we conducted the following experiment. For each STAPLE result, we thresholded $\bar{\mathbf{d}}$ at varying levels $(l - 1/2)/J, l = 1, \ldots, J$ and recorded the level yielding the highest Dice score with $\hat{\mathbf{t}}_{STAPLE}$. The resulting threshold levels and corresponding Dice scores, averaged across all 20 subjects and all 10 structures, are shown in the table below (standard deviations are in parentheses):

raters	Basic		Restricted		Advanced	
	threshold	Dice	threshold	Dice	threshold	Dice
5	0.30 (±0.00)	1.00 (±0.01)	0.41 (±0.04)	0.89 (±0.09)	0.45 (±0.05)	0.91 (±0.08)
15	0.17 (±0.03)	0.99 (±0.01)	0.46 (±0.05)	0.98 (±0.02)	0.46 (±0.06)	0.98 (±0.01)
37	0.14 (±0.06)	1.00 (±0.01)	0.44 (±0.14)	0.99 (±0.01)	0.43 (±0.15)	0.99 (±0.01)

For the Basic variant, which includes large consensus regions, the thresholding behavior of STAPLE is apparent for all number of raters. The threshold level, which corresponds to the point of zero-crossing $-o_{\hat{\mathbf{t}}}/s_{\hat{\mathbf{t}}}$ of the line of Eq. (6), is clearly below the majority voting 0.5 level because the size of the background is very large, which through $c_{1,0}$ increases $s_{\hat{\mathbf{t}}}$ but not $o_{\hat{\mathbf{t}}}$. The threshold level also decreases as more raters are added – thereby gradually yielding the *union* of all segmentations – because the fixed spatial prior favors background, making $c_1 < c_0$ and therefore rendering $-o_{\hat{\mathbf{t}}}/s_{\hat{\mathbf{t}}}$ dependent on J.

Both the Restricted and Advanced variants, which only consider non-consensus voxels, clearly exhibit the predicted thresholding behavior for $J \gg 0$, with Dice scores around 0.99 for both methods when $J = 37$, and around 0.98 when $J = 15$. However, for the Restricted case these numbers mask a more complex underlying phenomenon: As can be seen from Eq. (1), when $p(\boldsymbol{\pi})$ clamps $\boldsymbol{\pi}$ to $(0.5, 0.5)^T$ and $p(\boldsymbol{\theta}) \propto 1$, $p(\boldsymbol{\omega}|\mathbf{D})$ is invariant to swaps of the type $\theta^j_{d,0} \leftrightarrow \theta^j_{d,1}$, which corresponds to interchanging the role of the background and foreground label. Since the spatial prior was very close to 0.5 across all experiments (mean value 0.46, standard deviation 0.06), the posteriors $p(\boldsymbol{\omega}|\mathbf{D})$ were typically bimodal (cf. Fig (1)). In this variant a spatial prior different from 0.5 is the only factor discerning between the two modes, but in more than 20% of the cases for 15 and 37 raters, finding the global optimum would have yielded a solution similar to thresholding $\bar{\mathbf{d}}$ *and subsequently inverting the labels.* The fact that this is not apparent from the table above is because STAPLE got trapped in the wrong solution in all these cases (cf. Fig. (1)). When the number of raters was 5, STAPLE failed to locate the global maximum in 55% of the cases; the spatial prior encouraged the wrong solution in 46% of the cases.

The Advanced variant discerns between the two modes by encouraging performance parameters that are better-than-random, and the resulting models were found to always identify the correct solution. For all 15 and 37 rater cases, for which $p(\boldsymbol{\omega}|\mathbf{D})$ is strongly peaked, STAPLE also successfully located this solution. When initializing the proposed discrete optimizer at $\hat{\mathbf{t}}_{STAPLE}$ to see if further improvements in $\log p(\mathbf{t}|\mathbf{D})$ could be achieved, only modest improvements were obtained (0.150 and 1.225 on average for 37 and 15 raters, respectively,

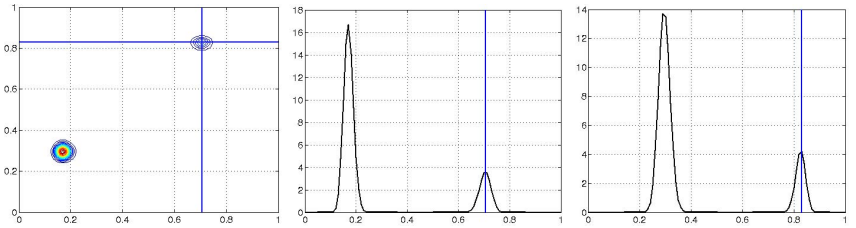

Fig. 1. Example of the bimodal parameter posterior often seen with the Restricted variant (37 raters, left lateral ventricle). The left shows a contour plot of $p(\theta^j_{1,1}, \theta^j_{0,0}|\mathbf{D})$ for one rater j, computed using the proposed Gibbs sampler; the two other plots show $p(\theta^j_{1,1}|\mathbf{D})$ and $p(\theta^j_{0,0}|\mathbf{D})$. The blue lines indicate the location of the (suboptimal) STAPLE parameter estimate – the global optimum would swap the fore- and background.

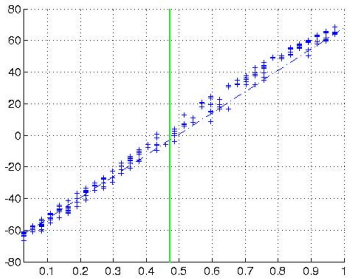

Fig. 2. Plot of the linear behavior of the "log odds": each cross corresponds to a voxel plotted against its foreground fraction (Advanced variant, 37 raters, left hippocampus). Voxels above the 0 level are assigned to foreground; a very similar result (Dice overlap 0.989) can be obtained by thresholding $\bar{\mathbf{d}}$ at level 0.472 (indicated by the thick green line).

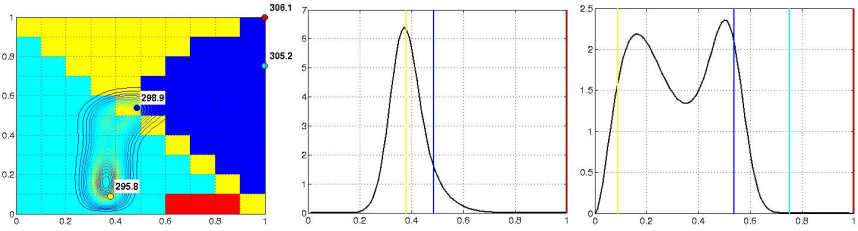

Fig. 3. Example of the broad, complex parameter posteriors often seen when only 5 raters are used (Advanced variant, left hippocampus). The plots are similar to those shown in Fig. 1, except that local optima arrived at when re-initializing STAPLE differently are shown in different colors. Tiles with the same color in the left plot indicate initializations in a 10×10 grid that arrive at the same optimum (indicated with a color-filled circle with the value of $-\log p(\mathbf{D}|\mathbf{t})p(\mathbf{t})$ indicated). Note that the (presumably global) yellow optimum is not found by STAPLE (shown in blue), and that the local optima in the high-dimensional parameter space do not generally correspond to local optima in the lower-dimensional marginal distributions.

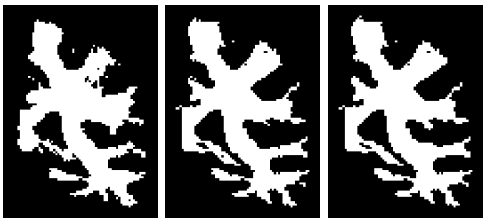

Fig. 4. When the number of raters is small, STAPLE may not yield the optimal segmentation truth estimate, even when it finds globally optimal parameters (as is the case in this example). From left to right: majority voting, STAPLE, and proposed discrete optimizer (Advanced variant, 5 raters, left cerebral white matter).

corresponding to 1.162 and 3.404 ratios in probabilities), which is in line with our theoretical predictions. Fig. (2) shows the "log odds" for each voxel under consideration for a 37-rater case, plotted vs. its foreground fraction f_i; the match with the predicted line of Eq. (6) is clear. When only 5 raters were used, much larger improvements in $\log p(\mathbf{t}|\mathbf{D})$ could be obtained with the discrete optimizer: 11.850 on average, which corresponds to solutions that are over $100,000$ times more likely. Some of this is attributable to the fact that $p(\boldsymbol{\omega}|\mathbf{D})$ is often a complex distribution that makes STAPLE susceptible to getting trapped in local optima: For each case we repeatedly re-ran STAPLE using a 10×10 parameter grid for initialization of $(\theta^j_{1,1}, \theta^j_{0,0})$ (cf. Fig. (3)), and found that in 23% a better optimum could be located this way. However, even when the correct parameter estimate $\hat{\boldsymbol{\omega}}$ was found, the broad distribution $p(\boldsymbol{\omega}|\mathbf{D})$ typically makes the STAPLE solution $\hat{\mathbf{t}}_{STAPLE}$ amendable to further improvement (Fig. (4)). Note in Fig. (3) also the difficulty in interpreting the value of individual components of $\hat{\boldsymbol{\omega}}$ in these cases.

5 Discussion

In this paper we have analyzed the theoretical properties of the STAPLE algorithm, revealing several fundamental shortcomings that cast doubt on the soundness and usefulness of results obtained with this method. We note that, although we only considered the binary STAPLE case here, the obtained results readily translate to cases with more than two labels.

Acknowledgments. This research was supported by NIH NCRR (P41-RR14075), NIBIB (R01EB013565, K25EB013649), and a BrightFocus grant (AHAF-A2012333).

References

1. Warfield, S.K., et al.: Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE TMI 23(7), 903–921 (2004)

2. Commowick, O., et al.: Estimating a reference standard segmentation with spatially varying performance parameters: Local MAP STAPLE. IEEE TMI 31(8), 1593–1606 (2012)
3. Asman, A.J., Landman, B.A.: Formulating spatially varying performance in the statistical fusion framework. IEEE TMI 31(6), 1326–1336 (2012)
4. Commowick, O., Warfield, S.K.: Estimation of inferential uncertainty in assessing expert segmentation performance from STAPLE. IEEE TMI 29(3), 771–780 (2010)
5. Landman, B., et al.: Robust statistical fusion of image labels. IEEE TMI 31(2), 512–522 (2012)
6. Langerak, T.R., et al.: Label fusion in atlas-based segmentation using a selective and iterative method for performance level estimation (SIMPLE). IEEE TMI 29(12), 2000–2008 (2010)
7. Sabuncu, M.R., et al.: A generative model for image segmentation based on label fusion. IEEE TMI 29(10), 1714–1729 (2010)
8. Rohlfing, T., Russakoff, D.B., Maurer, C.R.: Expectation maximization strategies for multi-atlas multi-label segmentation. In: Taylor, C.J., Noble, J.A. (eds.) IPMI 2003. LNCS, vol. 2732, pp. 210–221. Springer, Heidelberg (2003)