On the Expectation-Maximization Algorithm for Rice-Rayleigh Mixtures With Application to Noise Parameter Estimation in Magnitude MR Datasets

Ranjan Maitra*

Abstract

Magnitude magnetic resonance (MR) images are noise-contaminated measurements of the true signal, and it is important to assess the noise in many applications. A recently introduced approach models the magnitude MR datum at each voxel in terms of a mixture of upto one Rayleigh and an *a priori* unspecified number of Rice components, all with a common noise parameter. The Expectation-Maximization (EM) algorithm was developed for parameter estimation, with the mixing component membership of each voxel as the missing observation. This paper revisits the EM algorithm by introducing more missing observations into the estimation problem such that the complete (observed and missing parts) dataset can be modeled in terms of a regular exponential family. Both the EM algorithm and variance estimation are then fairly straightforward without any need for potentially unstable numerical optimization methods. Compared to local neighborhood- and wavelet-based noise-parameter estimation methods, the new EM-based approach is seen to perform well not only on simulation datasets but also on physical phantom and clinical imaging data.

Keywords:Bayes Information Criterion, Integrated Completed Likelihood, local skewness, mixture model, Rayleigh density, Rice density, robust noise estimation, wavelets

1 Introduction

The noise parameter in an acquired MR image dataset quantifies the degradation in the signal from sources such as random currents in the system, from within the MR apparatus (Smith and Lange, 2000; Hennessy, 2000), or owing to variation within the magnetic field (Weishaupt et al., 2003). This parameter – denoted in this paper by its customary Greek letter σ – is the common standard deviation (SD) of the Gaussian distribution that models the noise-contaminated complex-valued realizations (Wang and Lei, 1994; Sijbers, 1998) arising from the Fourier reconstruction in k-space (Brown et al., 1982; Ljunggren, 1983; Tweig, 1983) that is at the heart of the acquired magnitude MR signal. Because the k-space data at each voxel are homogeneous spherical Gaussian-distributed, the magnitude MR signal has the Rayleigh density

$$\varrho(x;\sigma) = \frac{x}{\sigma^2}\exp\left(-\frac{x^2}{2\sigma^2}\right), \qquad x > 0 \tag{1}$$

at a background voxel (no true signal), and the Rice density (Rice, 1944, 1945)

$$\varrho(x;\sigma,\nu) = \frac{x}{\sigma^2}\exp\left(-\frac{x^2+\nu^2}{2\sigma^2}\right)\mathrm{I}_0\left(\frac{x\nu}{\sigma^2}\right), \quad x > 0 \tag{2}$$

at a foreground voxel with noise-uncontaminated true signal ν. Here $\mathrm{I}_0(\cdot)$ is the modified Bessel function of the first kind of zeroth order. The underlying physical characteristics of the tissue at each voxel relate to ν through the Bloch equation and user-controlled design parameters (Hinshaw and Lent, 1983) such as echo time (TE), repetition time (TR) and flip angle (ϕ).

This paper focuses on accurate estimation of σ, needed in many applications (see Sijbers et al., 2007, for a listing). For example, it can provide an assessment of image acquisition quality (McVeigh et al., 1985) and can guide improvements in scanner design and signal-noise ratio (SNR) characteristics, allowing for shorter image acquisition

*Department of Statistics, Iowa State University, Ames, IA, USA.

times and higher contrasts and resolutions (Bammer et al., 2005). The noise parameter is also needed in many reconstruction algorithms and applications, such as finding contours of the brain (Brummer et al., 1993), for synthetic MR imaging (Glad and Sebastiani, 1995; Maitra and Riddles, 2010), for MR image registration (Rohdea et al., 2005), segmentation (Zhang et al., 2001) or restoration (Ahmed, 2005; Pasquale et al., 2004) algorithms.

Current techniques estimate σ differently, based on whether one or many images are available. In the latter case, Sijbers et al. (1998) provided a method that is robust to structural image errors and artifacts (Wilde et al., 1997), but requires data on k-space in addition to magnitude images. The single image techniques, on the other hand, attempt to automatically extract (Sijbers et al., 2007) the background by thresholding a histogram of the magnitude image data, and then estimate σ using maximum likelihood (ML) estimation on these Rayleigh distributed voxels (1). This method is, however, inapplicable to estimating σ in images with little or no background. Aja-Fernández et al. (2009) tried to address this shortcoming by estimating the noise parameter under Gaussian assumptions – their method works well when the SNR is high but is biased under lower SNR (Rajan et al., 2010). A more versatile alternative (Maitra and Faden, 2009), that is also applicable to multiple-image scenarios, fits a mixture distribution of an *a priori* unknown number of Ricean[1] components with at most one Rayleigh density, all with common noise parameter σ, to the observed voxel-wise magnitude data, and then to estimate σ using ML through the Expectation-Maximization (EM) algorithm. Maitra and Faden (2009) recommended using the estimated variance of the estimated σ to gauge its stability and use that to select the number of Ricean components which, though ancillary to the problem, is needed to select the model and hence the estimated σ.

In contrast to the development of Maitra and Faden (2009), Rajan et al. (2010) proposed a somewhat more simplified approach to noise estimation. Specifically, they postulated that under the assumptions of an image being subdivided into several homogeneous segments, locally estimating the noise over each window region centered on each voxel would result in as many estimates as number of interior voxels. Because these windowed regions are typically smaller than the homogeneous image segments, the majority of these local estimates would be close to the true estimates of the global noise: thus, the mode of these estimates would provide a good estimate of the local noise. Two methods were used for the local estimation: in the first instance, the authors used time-consuming ML estimation methods in the spirit of Aja-Fernández et al. (2009). However, in the interests of computer time, the authors recommended using a more efficient but heuristic local estimation approach which calculates noise estimates based on Gaussian assumptions and follows this with a 10-term polynomial correction to account for the Ricean noise. While the exact derivations of these polynomial corrections are unclear, note that the authors have revised and released what they contend are more accurate coefficients (as per personal communication with the first author). We return to a more algorithmic detailing of these methods in Section 3.

Another approach to noise estimation was provided by Coupè et al. (2010) who used wavelets and proposed adapting for Rice data, the Gaussian-noise-based Median Absolute Deviation (MAD) estimator in the wavelet domain (Donoho and Johnstone, 1994; Donoho, 1995). Specifically, a correction to the noise estimator is made through the iterative SNR-estimation scheme of Koay and Basser (2006). This correction scheme needs the mean signal of the object as well as initial noise estimates, both of which are obtained using the first level of wavelet decomposition. More specific details on their methods are also provided in Section 3.

Maitra and Faden (2009) used the component membership of the magnitude data at each voxel as the missing information. However, their maximization (M-step) required numerical methods for optimization, providing a source for numerical instability. In Section 2 of this paper, we show that using the phase angle at each voxel as additional missing information simplifies calculations substantially so that the M-step updates are all in closed form. This approach also means that the likelihood of the complete observations is from the regular exponential family (REF) so that additional simplifications can be used in estimating the variance of σ. Results on a series of experiments of computer-generated and physical phantom data, and on clinical images in Section 3 show that $\hat{\sigma}$ is estimated under the new approach with lower variance, and that the Bayes Information Criterion (BIC) (Schwarz, 1978) recovers its traditional claim of being an excellent performer in estimating the number of components. The method is seen to be a strong competitor to the local estimation- and wavelet-based noise estimation methods of Rajan et al. (2010) and Coupè et al. (2010) for simulation experiments, and in addition outperforms both in the context of physical scanner-acquired data. We end the main part of this paper with some discussion (Section 4). An appendix provides some technical details and derivations.

[1]Note that some authors use the spelling "Rician": we follow others in using "Ricean" since the adjective is derived from the name of S. O. Rice.

2 Theory & Methods

2.1 The Rice-Rayleigh Mixture Distribution

Let $R_1, R_2, \ldots, R_n$ be the observed magnitude data at the n voxels in the MR image. Following Maitra and Faden (2009), each R_i is independently distributed according to the mixture distribution

$$R_i \sim \sum_{j=1}^{J} \pi_j \varrho(x; \sigma, \nu_j) \tag{3}$$

where π_j is the proportion of voxels with underlying signal ν_j and common noise parameter σ. We assume that ν_js are positive $j = 1, 2, \ldots, J-1$ while $\nu_J \geq 0$. All ν_js are distinct. If $\nu_J \equiv 0$, the Jth component density is given by (1), otherwise or for all other js, the density is given by (2). We refer to Maitra and Faden (2009) for discussion on possible interpretations of (3), noting that, as in Maitra and Faden (2009), our focus in this paper is exclusively on estimating σ given $\boldsymbol{R} = \{R_1, R_2, \ldots, R_n\}$ but $\boldsymbol{\pi} = \{\pi_j; j = 1, 2, \ldots, J\}$, $\boldsymbol{\nu} = \{\nu_j; j = 1, 2, \ldots, J\}$s and the number of components J are unknown nuisance parameters and need to be accounted for in the process.

2.2 The EM Algorithm for Parameter Estimation

At this point, we assume that J is given and fixed. Let us denote the full set of parameters by $\boldsymbol{\theta} = \{\boldsymbol{\nu}, \boldsymbol{\pi}, \sigma\}$. Note that since $\boldsymbol{\pi}$ has components on the $(J-1)$-dimensional simplex, we have $2J$ parameters that require to be estimated when $\nu_J > 0$ and $2J-1$ parameters when $\nu_J \equiv 0$. Direct parameter estimation can be computationally intractable even for small J, so Maitra and Faden (2009) provide an EM algorithm (Dempster et al., 1977) for ML estimation by augmenting the observed magnitude data $\boldsymbol{R}$ with unobserved labels $\boldsymbol{W} = \{W_{i,j}, i = 1, 2, \ldots, n; j = 1, 2, \ldots, J\}$ that correspond to each of the mixture components. $W_{i,j}$s are indicator variables, with $W_{i,j} = 1$ indicating that the ith observation has true signal ν_j. Then $\boldsymbol{W}$ and $\boldsymbol{R}$ together form the complete data, with complete log likelihood: $\ell(\boldsymbol{\theta}; \boldsymbol{R}, \boldsymbol{W}) = \sum_{i=1}^{n} \sum_{j=1}^{J} W_{i,j} \left[\log \pi_j + \log \varrho(R_i; \sigma, \nu_j)\right]$,which is not from the REF. Since $\boldsymbol{W}$ is not observed, it is estimated by its conditional expectation given $\boldsymbol{R}$ and the current iterated parameter estimates and maximized in the M-step. The maximization needs numerical optimization for which Maitra and Faden (2009) used L-BFGS-B (Byrd et al., 1995). This however brings in concerns on the stability and convergence (Zhu et al., 1994) of the optima at each M-step besides being considerably computationally expensive (note that each M-step iteration itself involves several iterative L-BFGS-B steps).

2.2.1 An Alternative Implementation of the EM Algorithm

In our alternative implementation, we augment the observed magnitude data not only with $\boldsymbol{W}$, but also with $\boldsymbol{\gamma} = \{\gamma_1, \gamma_2, \ldots, \gamma_n\}$ where γ_i is the acquired phase angle at each voxel. The complete data is then given by $\boldsymbol{Z} = (\boldsymbol{W}, \boldsymbol{R}, \boldsymbol{\gamma})$. Using a characterization of the Rice distribution, we note that R_i has the density (2) iff $R_i \cos\gamma_i$ and $R_i \sin\gamma_i$ are independently distributed as $N(\nu_j \cos\mu, \sigma^2)$ and $N(\nu_j \sin\mu, \sigma^2)$ for any given μ (in particular, for $\mu = 0$, which we use in this paper). The characterization also holds for the Rayleigh distribution, with $\nu_j = 0$. Note also that $R_i \cos\gamma_i$ and $R_i \sin\theta_i$ are akin to the real and imaginary parts of the complex MR data at each voxel, from which the magnitude observations are obtained. The complete data is then $\boldsymbol{Z} = (\boldsymbol{R}, \boldsymbol{W}, \boldsymbol{\gamma})$ with complete log likelihood, after ignoring terms not involving $\boldsymbol{\theta}$, given by

$$\ell(\boldsymbol{\theta}; \boldsymbol{R}, \boldsymbol{W}, \boldsymbol{\gamma}) = \sum_{i=1}^{n} \sum_{j=1}^{J} W_{i,j} \left[\log \pi_j - 2\log\sigma - \frac{R_i^2 \sin^2\gamma_i}{2\sigma^2} - \frac{(R_i \cos\gamma_i - \nu_j)^2}{2\sigma^2}\right],$$

which reduces to

$$\ell(\boldsymbol{\theta}; \boldsymbol{Z}) = -2n\log\sigma - \sum_{i=1}^{n} \sum_{j=1}^{J} W_{i,j} \left[\log \pi_j - \frac{R_i^2 + \nu_j^2}{2\sigma^2}\right] + \sigma^{-2} \sum_{i=1}^{n} \sum_{j=1}^{J} \nu_j R_i W_{i,j} \cos\gamma_i. \tag{4}$$

Note that (4) is a member of the REF. Further, observations on $W_{i,j}$ and γ_i being absent, we replace the terms in (4) involving them by their conditional expectation given the observed magnitude at the values of the current iterate. This constitutes the E-step of the algorithm.

E-Step calculations The conditional expectation of $W_{i,j}$ given R_i is as in Maitra and Faden (2009). Specifically, it is

$$w_{i,j}^{(t)} = \frac{\sum_{i=1}^{n} \pi_j^{(t-1)} \varrho(R_i; \sigma^{(t-1)}, \nu_j^{(t-1)})}{\sum_{i=1}^{n} \sum_{q=1}^{J} \pi_q^{(t-1)} \varrho(R_i; \sigma^{(t-1)}, \nu_q^{(t-1)})}, \tag{5}$$

where $\boldsymbol{\theta}^{(t)}$ is the parameter estimate at the tth EM iteration. (Note that since $W_{i,j}$ is an indicator variable, $w_{i,j}^{(t)}$ is also the conditional probability that $W_{i,j} = 1$ given R_i and the current parameter values $\boldsymbol{\theta}^{(t-1)}$.) Thus, we address terms involving the unknown $W_{i,j}$ in (4). To address terms involving $W_{i,j} \cos\gamma_i$ in (4), note that the term corresponding to $j = J$ drops out if $\nu_J = 0$, so it is enough to only consider the cases for which we have all positive ν_js. Then the conditional expectation of $W_{i,j} \cos\gamma_i$ given $\underline{\mathrm{R}}$ is equivalent to $E_{\boldsymbol{\theta}^{(t-1)}}(W_{ij} \cos\gamma_i \mid R_i)$, which can be obtained using $E_{\boldsymbol{\theta}^{(t-1)}}\left[E_{\boldsymbol{\theta}^{(t-1)}}(W_{ij} \cos\gamma_i \mid R_i, W_{i,j}) \mid R_i\right]$, from standard results on conditional expectations. This last reduces to $E_{\boldsymbol{\theta}^{(t-1)}}\left[W_{i,j} E_{\boldsymbol{\theta}^{(t-1)}}(\cos\gamma_i \mid R_i, W_{i,j}) \mid R_i\right]$ which is $w_{i,j}^{(t)} E_{\boldsymbol{\theta}^{(t-1)}}(\cos\gamma_i \mid R_i, W_{i,j} = 1)$, and which upon applying Theorem A.1 (see Appendix) yields the tth E-step update for $W_{i,j} \cos\gamma_i$ given R to be

$$w_{i,j;\cos}^{(t)} = w_{i,j}^{(t)} \frac{\mathrm{I}_1(\frac{R_i \nu_j}{\sigma^2})}{\mathrm{I}_0(\frac{R_i \nu_j}{\sigma^2})}, \tag{6}$$

where $\mathrm{I}_1(\cdot)$ is the modified Bessel function of the first kind of the first order.

M-Step calculations The M-step maximizes the conditional expectation of (4) given $\boldsymbol{R}$, historically denoted as the Q function, and evaluated at the current estimated values of the parameters $\boldsymbol{\theta}^{(t-1)}$. Setting the first partial derivatives of this Q-function with respect to the parameters yields the following M-step updates at the tth iteration:

$$\begin{aligned}
\nu_j^{(t)} &= \frac{\sum_{i=1}^{n} R_i w_{i,j;\cos}^{(t)}}{\sum_{i=1}^{n} w_{i,j}^{(t)}}, j = 1, 2, \ldots, J \\
\pi_j^{(t)} &= \frac{\sum_{i=1}^{n} w_{i,j}^{(t)}}{\sum_{j=1}^{J} \sum_{i=1}^{n} w_{i,j}^{(t)}}, j = 1, 2, \ldots, J \\
\sigma^{2(t)} &= \frac{1}{2n} \sum_{i=1}^{n} \left[R_i^2 - 2R_i \sum_{j=1}^{J} w_{i,j,\cos}^{(t)} \nu_j^{(t)} + \sum_{j=1}^{J} w_{i,j,\cos}^{(t)} \nu_j^{2(t)} \right],
\end{aligned}$$

which are all of closed form and easily calculated at every iteration. Note also that in the above, if the J-th component is Rayleigh-distributed, the corresponding $\nu_J^{(t)}$ is always set at zero. Also, the closed-form nature of the M-step updates points to substantial computational savings in contrast with the tedious numerical optimization in M-step of Maitra and Faden (2009).

The EM algorithm starts from some initialized values of the parameters and alternates the E- and M-steps till convergence which is declared when there is very little relative increase in the observed log likelihood. In our implementation, we have followed Maitra and Faden (2009) in addressing separately the cases for when ν_J is positive or zero, *i.e.*, for the case when the Jth component is Rice- and Rayleigh-distributed, respectively. Implementation for both cases is similar, but for the additional restriction that $\nu_J \equiv 0$ in the latter case. Once the EM-converged estimates are obtained, the likelihood of (3) is evaluated separately for the cases $\nu_J \equiv 0$ and $\hat{\nu}_J > 0$: the case with the higher value, along with the corresponding parameters $\hat{\boldsymbol{\theta}} = \{\hat{\sigma}, \hat{\boldsymbol{\nu}}, \hat{\boldsymbol{\pi}}\}$, are the parameter MLEs for given J. For that J, $\hat{\sigma} \equiv \hat{\sigma}^{(J)}$, is the MLE of the common noise parameter of the image given a J-component Rice-Rayleigh mixture distribution.

2.2.2 Initialization

The EM algorithm proceeds from initializing parameter values and converges to a (local) maximum in the vicinity of its initialization. Thus, the initial values can have tremendous consequences on its performance. In the Rice mixture setup, Maitra and Faden (2009) have provided a computationally expensive deterministic approach that built on and adapted the multi-stage initializer of Maitra (2009). In the case of Gaussian mixtures however, Maitra and Melnykov (2010) have shown that randomly-chosen initializations done using the *em-EM* approach of Biernacki et al. (2003)

and its *Rnd-EM* variant (Maitra, 2009) perform the best. We therefore propose and adopt a hybrid version of both *em-EM* and *Rnd-EM*. Specifically, we choose M randomly chosen sets of initial values, and use these values to the EM algorithm for a small number of iterations (m) or until convergence, whichever is sooner. The observed log likelihood is evaluated at each of these M sets of final values and the EM algorithm is initialized till convergence from the highest of these final values. The basic philosophy is that we will start the algorithm on trial runs at several randomly chosen initial values and prune away all that do not show maximum promise in a short number of steps. When $m = 1$, we have the *Rnd-EM* algorithm, while we have the *em-EM* algorithm when we have a $m \to \infty$ and lax convergence criterion in the trial runs. In our experiments in this paper, we report results done using $m = 5$ and $M = 500 + 50J$, though other choices in the ballpark did not provide vastly different results.

Intelligent choice of initializing candidates One issue that arises in the context of stochastic initialization methods is the way in which the J candidate starting points are chosen. Traditionally, these have been chosen by sampling randomly and without replacement J observations from the data, and proceeding with the above. This is, however, a fairly wasteful strategy, because it prefers more candidates from larger homogeneous components. Such candidates rarely reflect the true composition of the data and usually are discarded as showing less promise compared to the rest. This sort of uniform random selection of points is usually quite problematic in image datasets which have large proportions of voxels from a similar component (*eg* background), so that a simple random sample of J points has a high chance of having more than one observation from the same component. We therefore propose a more intelligent way of choosing the candidate initializers.

Our proposal also randomly chooses initial values from observations in the dataset, but iteratively allows a greater probability of inclusion for those observations that are farther apart from those initializing values already included in an earlier step. Our specific approach is as follows.

1. For the mixture model involving a Rayleigh component, we first set $\mu_J = 0$, otherwise set μ_J as a randomly chosen R_i (with uniform probability of selection $1/n$). At this point, assign all observations to this class (call the assignment W_i, for $i = 1, 2, \ldots, n$). Thus, $W_i \equiv J$.

2. Remove the effect of μ_{W_i} from each observations, *i.e.*, let $X_i = R_i - \mu_{W_i}$. Use $\hat{\sigma}^2 = \sum_{i=1}^{n} X_i^2/2n$ as the current preliminary initial estimate of σ. For Rayleigh-distributed data (when $W_i = J$, and $\mu_J = 0$, the above is the ML estimate for σ, but for other cases, this is a preliminary, ad-hoc estimate.

3. For $j = 1$, Set $\mu_j = X_l$, where l is chosen to be l with probability p_l proportional to $1 - \exp\{-X_l^2/2\hat{\sigma}^2\}$, and $l \in \{1, 2, \ldots, n\}$. (This sampling strategy indicates that observations that are farther away from the current selected μs have a greater chance of being picked as the mean). Update each W_i to be the j for which R_i is closest to μ_j. Go back to Step 2.

4. Repeat Step 3 for $j = 2, 3, \ldots, J - 1$.

5. Let $\pi_j = \#\{W_i = j\}/n, j = 1, 2, \ldots, J$. The initializing candidate is thus $\{\hat{\sigma}, (\pi_j, \mu_j); j = 1, 2, \ldots, n\}$.

2.2.3 Determining convergence

A reviewer has very kindly asked us to clarify how convergence is decided. This is an important issue in many iterative algorithms (Altman et al., 2003) with ramifications in several cases: in our experiments, we have used bounds on the relative change in log likelihood as our criterion. Specifically, we declare that convergence is reached when the relative increase in log likelihood is no more than a pre-determined ϵ, set in our experiments to be 10^{-4}.

2.2.4 Variance of the estimate

As with other ML-based parameter estimation methods, (at the very least, approximate) variance estimates can be obtained readily. Maitra and Faden (2009) provide a very tedious implementation (involving Hessians and the like) of the Louis (1982) approach to calculating the observed information $\mathcal{I}_{\boldsymbol{R}}$. Our suggested implementation of the E-M algorithm in this paper, however, provides us with an additional payoff. This is because as mentioned earlier, the complete log likelihood (4) is a member of the REF, so that another simplification may be availed of. Specifically, for independent identically distributed observations from (3), letting ∇q_i be the gradient vector of the expected complete loglikelihood at the ith observation $q_i \equiv q_i(\boldsymbol{\theta}^{(J)}; R_i)$, the information matrix $\mathcal{I}(\boldsymbol{\theta}^{(J)})$ of the obtained EM estimates

is easily estimated through its empirical counterpart $(\nabla q_1 \vdots \nabla q_2 \vdots \ldots \vdots \nabla q_n)(\nabla q_1 \vdots \nabla q_2 \vdots \ldots \vdots \nabla q_n)'|_{\boldsymbol{\theta}^{(J)}=\hat{\boldsymbol{\theta}}^{(J)}}$ — see pp. 64–66 of McLachlan and Peel (2000) or pp. 114–5 of McLachlan and Krishnan (2008). In our case,

$$q_i(\boldsymbol{\theta}; R_i) = -2\log\sigma + \sum_{j=1}^{J}\left[w_{i,j}^{(t)}\log\pi_j - \frac{R_i^2 + w_{i,j}^{(t)}\nu_j^2}{2\sigma^2} + \frac{\nu_j R_i w_{i,j;\cos}^{(t)}}{\sigma^2}\right]$$

so that the gradient vector ∇q_i is obtained from the partial derivatives of q_i with respect to the elements of the parameter vector as follows:

$$\begin{aligned}
\frac{\partial q_i}{\partial \nu_j} &= \frac{R_i w_{i,j,\cos}^{(t)} - 2w_{i,j}^{(t)}\nu_j}{\sigma^2}, \quad j = 1, 2, \ldots, J \\
\frac{\partial q_i}{\partial \pi_j} &= \frac{w_{i,j}^{(t)}}{\pi_j} - \frac{w_{i,J}^{(t)}}{\pi_J}, \qquad j = 1, 2, \ldots, J-1 \\
\frac{\partial q_i}{\partial \sigma} &= -\frac{2}{\sigma} + \frac{R_i^2 - \sum_{j=1}^{J}(2R_i w_{i,j,\cos}^{(t)}\nu_j - w_{i,j}^{(t)}\nu_j^2)}{\sigma^3}
\end{aligned}$$

which are evaluated at the converged values of $\hat{\boldsymbol{\theta}}$ and the corresponding $w_{i,j}^{(t)}$s and $w_{i,j;\cos}^{(t)}$s. Once the empirical information matrix is obtained, it is inverted to provide an estimate of the variance-covariance matrix of $\hat{\boldsymbol{\theta}}$. The square root of the diagonal entry corresponding to σ provides us with the standard error for our estimate $\hat{\sigma}^{(J)}$.

2.3 Choosing the optimal number of components

There are a number of approaches (McLachlan and Peel, 2000; Fraley and Raftery, 2002) to choosing the most appropriate number of components in finite mixture models and model-based clustering. We develop and study two sets of approaches as applied to our problem.

2.3.1 Penalty-based approaches

There are several penalty-based approaches in the literature, but in this application, we have only investigated two of the most promising ones.

Bayes Information Criterion Perhaps the most popular penalized approach, the Bayes Information Criterion (BIC) (Schwarz, 1978) finds the optimal number (J_{opt}) of groups (from a range $J \in \{1, 2, \ldots, J_{\max}\}$) minimizing the negative log likelihood of the J-component model augmented by adding a penalty that is equal to n times the logarithm of the number of parameters in that model. The BIC is a popular choice and has some very desirable properties for Gaussian mixtures (Keribin, 2000). We use BIC to estimate the optimal J, which we denote as J_{opt}. Under these circumstances, the estimated σ in the J_{opt}-component mixture model is what we henceforth refer to as our BIC-estimate of σ.

The Integrated Completed Likelihood The Integrated Completed Likelihood (ICL) (Biernacki et al., 2000) uses the complete loglikelihood with the missing observations replaced by their conditional modes given the observed data. This value is penalized using similar approaches as in the case of penalized likelihood: the penalty as for BIC above is one popular and often well-performing choice. We use ICL with a BIC penalty in this paper: the estimated σ of the corresponding J_{opt}-component model is then referred to as the ICL-estimated σ.

2.3.2 Variability-based approaches

Maitra and Faden (2009) also proposed an alternative approach for their implementation of the EM for Rice-Rayleigh mixtures, based on the standard error $SE_{\hat{\sigma}^{(J)}}$. The basic idea here was that with increasing J, the model is initially more adequately specified, leading to a decrease in the uncertainty in the parameter estimate. This pattern changes course, however, beyond the true J because there is again more uncertainty in $\hat{\sigma}^{(J)}$ because (at least some of the) new allocations in the E-step are assigned in error. Thus Maitra and Faden (2009) suggested looking for the first J after which $SE_{\hat{\sigma}^{(J)}}$ rises. They demonstrated good performance of this approach in many of their experiments: however,

they also reported shortcomings in certain cases. One of the reasons may be that there are several parameters that are also estimated along with $\hat{\sigma}$ and the variability in those parameters should also give us some idea of the stability and precision of the estimates in the J-component model. We therefore look into a comprehensive evaluation of these parameters by investigating the variability in the expected complete log likelihood $Q(\cdot) = \sum_{i=1}^{n} q_i(\cdot)$ of the parameters upon convergence.

Note that calculation of the variance in the $Q(\cdot)$ requires some careful consideration. This is because the usual first-order delta method is inapplicable since $\nabla Q(\hat{\boldsymbol{\theta}}) = 0$ (as a consequence of $Q(\cdot)$ being maximized with respect to the parameters in the M-step). Therefore a higher- (second-) order delta method is employed. To do so, note that by standard results, $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically distributed as a normally distributed zero-mean random vector $\boldsymbol{\Gamma}$ with dispersion matrix given by the inverse of the information matrix ($\mathcal{I}^{-1}(\hat{\boldsymbol{\theta}})$, which can be estimated using the development of Section 2.2.4. Then by application of the second-order delta method, $n(Q(\hat{\boldsymbol{\theta}}) - Q(\boldsymbol{\theta}))$ is asymptotically distributed as $-\frac{1}{2}\boldsymbol{\Gamma}'\mathcal{H}_Q\boldsymbol{\Gamma}$ where $\mathcal{H}_Q$ is the Hessian of Q. Then, writing $\mathcal{I}^{-1}(\hat{\boldsymbol{\theta}})$ as $\boldsymbol{\Sigma}$. and $\boldsymbol{\Sigma}^{-\frac{1}{2}}$ as the square root of the nonnegative definite matrix $\boldsymbol{\Sigma}$, $\boldsymbol{\Gamma}'\mathcal{H}_Q\boldsymbol{\Gamma} = \boldsymbol{Z}\boldsymbol{\Sigma}^{\frac{1}{2}}\mathcal{H}_Q\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{Z} = \sum_{i=1}^{2J-1}\sum_{j=1}^{2J-1}\lambda_{ij}Z_iZ_j$, where $\boldsymbol{Z}$ is a standard normally distributed random vector and λ_{ij} are the elements of the matrix $\boldsymbol{\Sigma}^{\frac{1}{2}}\mathcal{H}_Q\boldsymbol{\Sigma}^{\frac{1}{2}}$. The variance of $Q(\hat{\boldsymbol{\theta}})$ is therefore given by $\frac{1}{2n}\sum_{i=1}^{2J-1}\sum_{j=1}^{i}\lambda_{ij}^2$.

It remains to calculate $\mathcal{H}_Q$. To do so, we report only the non-vanishing second-order partial derivatives:

$$
\begin{aligned}
\frac{\partial^2 Q}{\partial \nu_j^2} &= -\frac{1}{\sigma^2}\sum_{i=1}^{n} w_{i,j}^{(t)}, \quad j = 1, 2, \ldots, J \\
\frac{\partial^2 Q}{\partial \pi_j^2} &= -\sum_{i=1}^{n}\left(\frac{w_{ij}^{(t)}}{\pi_j^2} + \frac{w_{iJ}^{(t)}}{\pi_J^2}\right), \quad j = 1, 2, \ldots, J-1 \\
\frac{\partial^2 Q}{\partial \sigma^2} &= \frac{2}{\sigma^2} - \frac{3}{\sigma^4}\sum_{i=1}^{n}\left[R_i^2 + \sum_{j=1}^{J} w_{ij}^{(t)}\nu_j^2 - 2\sum_{j=1}^{J} w_{ij;\cos}^{(t)} R_i\nu_j\right] \\
\frac{\partial^2 Q}{\partial \nu_j \partial \sigma} &= -\frac{2}{\sigma^3}\sum_{i=1}^{n}\left[w_{ij;\cos}^{(t)} R_i - w_{ij}^{(t)}\nu_j\right], j = 1, 2, \ldots, J \\
\frac{\partial^2 Q}{\partial \pi_j \partial \pi_{j'}} &= -\sum_{i=1}^{n}\frac{w_{iJ}^{(t)}}{\pi_J^2}, \quad j = 1, 2, \ldots, J-1.
\end{aligned}
$$

In the above, $w_{ij}^{(t)}$ and $w_{ij;\cos}^{(t)}$ are the converged E-step quantities. Once the variance of Q is calculated for each J, it is used to calculate that J_{opt} after which the variability increases. We refer to the estimated σ of the so-chosen J_{opt}-component model as the Q-variability-based estimate of σ. The σ estimated using the variability in $\hat{\sigma}$ of Maitra and Faden (2009) is referred to as the $\hat{\sigma}$-variability-based estimate of $\hat{\sigma}$.

2.4 Sampling from the image cube

Since the EM algorithm converges slowly and may not be feasible to apply to the entire dataset, we follow Maitra and Faden (2009) in taking a coarse sub-grid of voxels and using that to obtain our BIC-estimated value of σ.

3 Performance Evaluations

The methodologies proposed in this paper were evaluated on realistic computer-generated and physical phantom datasets where the true σ could be set or accurately estimated. We also illustrated performance on a few clinical datasets. We examined performance of our algorithm in obtaining the BIC-estimated, the ICL-estimated, the $\hat{\sigma}$- and Q-variability-based estimated $\hat{\sigma}$. Additionally we compared performance with the local estimation methods in Rajan et al. (2010) and the robust wavelet-based noise estimation methods of Coupè et al. (2010). We now discuss these comparison sets of methods in some detail.

3.1 Comparison Methods for Noise Estimation

3.1.1 Local noise estimation approaches

Rajan et al. (2010) proposed building on the noise estimation approach of Aja-Fernández et al. (2009) which was only applicable to the case for high-SNR images in which case the Rice distribution is well-approximated by a Gaussian distribution. Rajan et al. (2010)'s suggested methodology proposes drawing a local windowed region around each voxel. Thus, there are as many windowed regions as interior voxels in the image. Under the assumption that there are large segments in the image, a large number of these windowed regions have a constant signal. The voxels in each of these regions can be assumed to be independent identically distributed realizations from the Rice distribution: thus the noise parameter can be estimated using standard likelihood methods(as in, say, Sijbers and den Dekker, 2004). An alternative but heuristic approach also provided by Rajan et al. (2010) obviates the need for computationally expensive likelihood maximization by estimating the noise parameter in each region using a local skewness-based estimator. Here, the variance (s^2) in each windowed region is calculated and scaled by a "correction factor" which depends on the estimated skewness (γ) of the distribution in that region. The correction factor is a ninth-order polynomial expression of the form $\Psi(\gamma) = \sum_{i=0}^{9} \psi_i \gamma^i$ where ψs are as in Table 1. These coefficients are from a lookup table

Table 1: Corrected coefficients in the ninth-order ten-term polynomial correction factor of Rajan et al. (2010). Decimal-points accuracy is as provided by the authors.

ψ_0	1.0007570413
ψ_1	2.8981188340
ψ_2	-72.9432278777
ψ_3	1162.679213636
ψ_4	-9838.85598962208
ψ_5	47813.9607638493
ψ_6	-137448.5785417688
ψ_7	230670.4056296062
ψ_8	-208666.38136498138
ψ_9	78562.5551923769

created by the authors: note however, that as per personal communication with the corresponding author in Rajan et al. (2010), the original published coefficients are incorrect and have since been replaced in favor of the ones in Table 1. The square root of the corrected variance ($\sqrt{s^2\Psi(\gamma)}$) is then taken to be the local noise parameter estimate for the windowed region. The mode of these windowed region estimates is then postulated to be the global estimate of the Ricean noise parameter. We denote the estimate obtained using the local MLEs and local skewness as $\hat{\sigma}_{\ell ML}$ and $\hat{\sigma}_{\ell sk}$ respectively.

Rajan et al. (2010) do not provide much guidance on the size of the local windowing regions. In personal communication, they suggest determining the size of the region based on dimension, slice thickness (for three-dimensional images) and pixel resolution. Specifically, for two-dimensional images having more than 256×256 pixels, they suggest using (two-dimensional) 9×9-sized window regions while for other two-dimensional images, 7×7-sized window regions may be used. For three-dimensional images with thick slices (defined to be greater than 2mm) they suggest using only two-dimensional window regions as above. For images with thinner slices, they suggest using a $7\times7\times3$-window regions for images having resolution greater than 256×256 pixels (in the axial plane), and $5\times5\times3$-sized regions otherwise. I have followed these specifications in all the experimental evaluations.

The issue of mode selection to obtain $\hat{\sigma}_{\ell ML}$ or $\hat{\sigma}_{\ell sk}$ is also not discussed in Rajan et al. (2010). Inspection of their Matlab code reveals that the authors propose selecting the mode of the local estimates by rounding every local estimate to their closest integer, and choosing the most frequently-occurring non-zero integer estimates as the final estimates. Our experiments also follow this recipe.

3.1.2 Robust Ricean noise estimation via wavelets

Donoho and Johnstone (1994) and Donoho (1995) provide a robust scheme for noise estimation in Gaussian data using

the (noise) HHH sub-band of the decomposed wavelet coefficients of a three-dimensional image. Their MAD estimate is given by $\hat{\sigma}_0 = median(\mid y_i \mid (/0.6745$ with y_i being the wavelet coefficients of the HHH sub-band. Coupè et al. (2010) propose restricting the coefficients in the above estimator to the object portion. This is done by first segmenting the corresponding LLL sub-band into the background and the object using a k-means algorithm (MacQueen, 1967; Hartigan and Wong, 1979) with $k = 2$ clusters. (Note that imperfect initialization is not an issue here because univariate data are being grouped into two clusters, so we start the k-means algorithm with the minimum and maximum values of the data as the starting means for the two clusters.) Thus voxels in the LLL sub-band are grouped into object and background. A further step removes the object voxels with high local gradients (*i.e.*, we remove all object voxels having local gradient in magnitude higher than the median). The remaining voxels in the corresponding HHH sub-band are used to obtain the MAD noise parameter estimate under Gaussian assumptions.

In order to obtain the noise parameter estimators under general Ricean assumptions, Coupè et al. (2010) propose scaling the MAD estimate obtained above by the square root of Koay and Basser (2006)'s correction factor $\zeta(\theta) = 2 + \theta^2 - \frac{\pi}{8} \exp\left(-\frac{\theta^2}{2}\right) \left[(2+\theta^2)\mathrm{I}_0(\theta^2/4) + \theta^2 \mathrm{I}_1(\theta^2/4)\right]^2$ where θ is the estimated SNR of the image. θ itself is estimated iteratively, following Koay and Basser (2006), to satisfy

$$\theta = \sqrt{\zeta(\theta)(1 + \bar{m}_o^2/\hat{\sigma}_0^2) - 2}. \tag{7}$$

The discerning reader may note that equation (7) differs somewhat from equation (9) in Coupè et al. (2010) – the latter is suspected to have typographical errors. We state equation (7) to match equation (11) of Koay and Basser (2006). where $\hat{\sigma}_0$ is as above and $\bar{m}_o$ is the mean signal of the object, obtained by segmenting, via k-means ($k = 2$), the image intensities, and restricting attention to the higher of the two means obtained upon termination.

Coupè et al. (2010) developed the noise estimation procedure for three-dimensional images. In our view, the methodology applies readily to two dimensions: we have used this applied version for our two-dimensional experiments. Further, note that equation (7) assumes that $\zeta(\theta)(1 + \bar{m}_o^2/\hat{\sigma}_0^2) - 2$ is positive. In our (two-dimensional experiments), we have not found this to always hold. Finally, in our experiments, we have used two wavelet filters: the Haar wavelet and the Daubechies (1992)' orthonormal compactly supported wavelet of length $L = 8$ least asymmetric family. Noise parameter estimates obtained using these two wavelets are denoted by $\hat{\sigma}_{\text{haar}}$ and $\hat{\sigma}_{\text{la8}}$, respectively.

3.2 Experimental Setup and Results

The scope of our experiments and other details mirrored Maitra and Faden (2009) to allow for ready comparison. For brevity, we refer to that paper for details on the setup, providing only a summary of the setup here.

3.2.1 Computer-generated Phantom Data

This set of evaluations again followed Maitra and Faden (2009) in using the Brainweb interface of Cocosco et al. (1997) to obtain noiseless computer-generated three-dimensional images, each of dimension $180 \times 216 \times 180$ pixels, and with three different intensity nonuniformity (INU) proportions of none (0%), modest (20%) and substantial (40%) bias fields. To each background and foreground pixel of the noiseless image, we added independent realizations from the Rayleigh and Rice distributions respectively, with common $\sigma \in \{5, 10, 20, 30, 50\}$. These 5 values were chosen to provide substantial to very low average SNRs of 5.33, 2.66, 1.33, 0.89 and 0.53, respectively. Also, we replicated 50 datasets at each setting to account for simulation variability in our experiments.

Figure 1 provides a graphical display of the distribution of the relative absolute errors in estimating σ using each method. Performance of each method relative to different field INU proportions and σ-values is quantitatively summarized in Table 2. From the figure and the table, we see that among the mixture-based methods using the EM algorithm of this paper, the BIC is the best performer regardless of the presence and strength of the bias field (INU proportion). The ICL estimation methods are also quite competitive but the other EM-based methods have a more varied performance: in particular, there are cases where they are quite off-the-mark. We note that the BIC-based method of this paper does better than the BIC-based method of Maitra and Faden (2009) at all settings, perhaps because the improved stability of our estimation methodology has resulted in more accurate calculation of the criterion. Thus, it appears that upon using the alternative approach to EM proposed in this paper, BIC is able to recover its general tag as a good competitor in mixture model selection, and should be the recommended estimation method over all choices of INU and (true) σ-values.

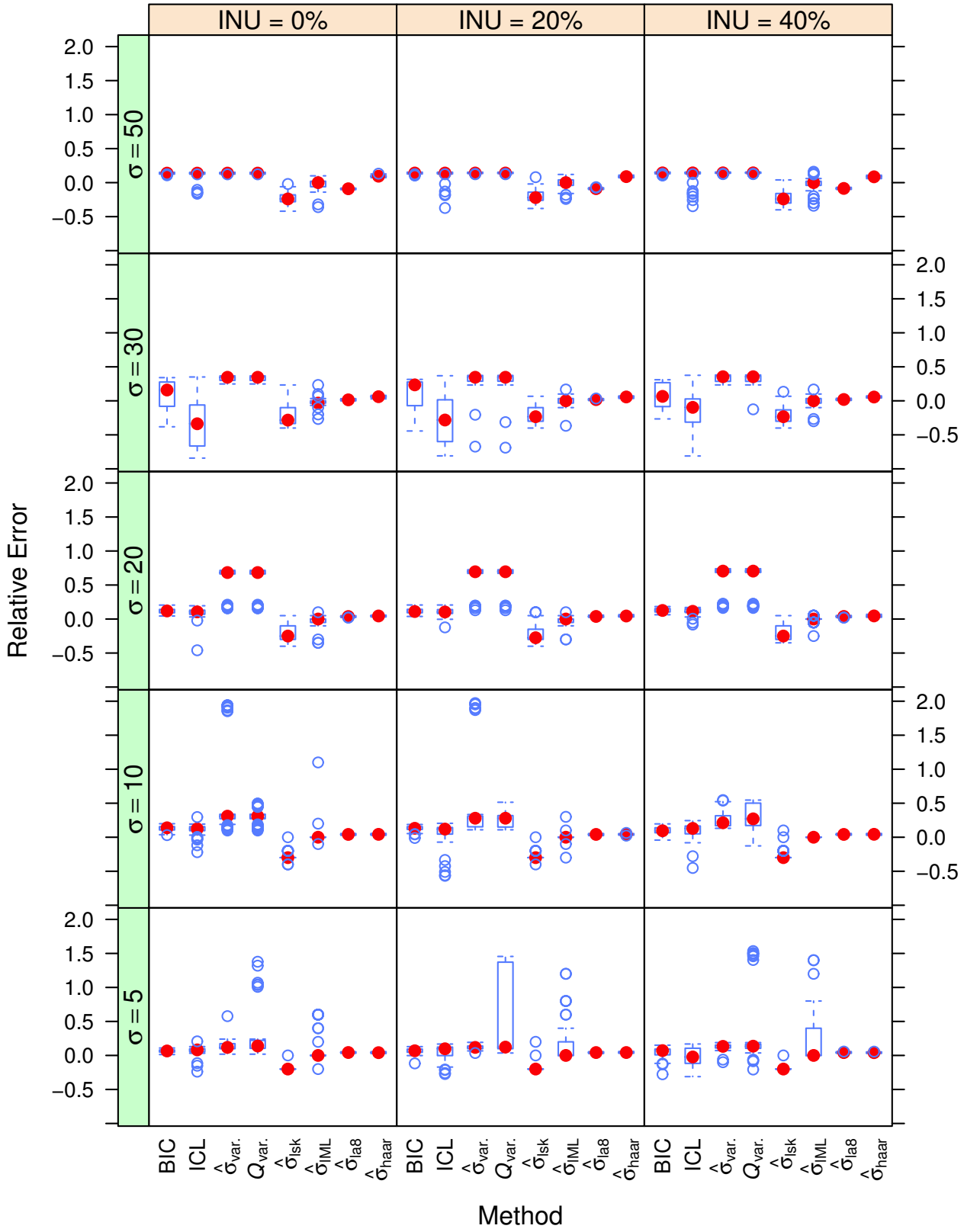

Figure 1: Relative errors in the estimated noise parameters for the Brainweb-simulated data.

Table 2: Summary measures of relative errors of each of the noise parameter estimation methods. Reported measures are of mean relative error (**Bias**) and the root mean squared relative error (**RMS**).

		INU Proportion = 0%		**INU Proportion = 20%**		**INU Proportion = 40%**	
True	**Method**	**Bias**	**RMS**	**Bias**	**RMS**	**Bias**	**RMS**
$\sigma = 5$	BIC	0.064	0.067	0.066	0.077	0.040	0.093
	ICL	0.068	0.097	0.040	0.129	-0.022	0.138
	$\hat{\sigma}_{\text{var.}}$	0.141	0.164	0.120	0.124	0.128	0.138
	$Q-$ var.	0.263	0.425	0.567	0.839	0.300	0.568
	$\hat{\sigma}_{\ell sk}$	-0.196	0.198	-0.188	0.198	-0.196	0.198
	$\hat{\sigma}_{\ell ML}$	0.088	0.226	0.160	0.353	0.248	0.458
	$\hat{\sigma}_{\text{la8}}$	0.043	0.043	0.042	0.042	0.042	0.042
	$\hat{\sigma}_{\text{haar}}$	0.042	0.042	0.044	0.045	0.041	0.042
$\sigma = 10$	BIC	0.132	0.138	0.127	0.133	0.097	0.111
	ICL	0.109	0.137	0.052	0.194	0.086	0.150
	$\hat{\sigma}_{\text{var.}}$	0.623	0.927	0.462	0.755	0.276	0.306
	$Q-$ var.	0.307	0.327	0.264	0.293	0.306	0.347
	$\hat{\sigma}_{\ell sk}$	-0.282	0.288	-0.282	0.287	-0.266	0.277
	$\hat{\sigma}_{\ell ML}$	0.024	0.159	0.000	0.063	0.000	0.000
	$\hat{\sigma}_{\text{la8}}$	0.041	0.042	0.041	0.041	0.042	0.043
	$\hat{\sigma}_{\text{haar}}$	0.043	0.043	0.043	0.043	0.043	0.043
$\sigma = 20$	BIC	0.116	0.122	0.118	0.124	0.128	0.132
	ICL	0.094	0.130	0.102	0.117	0.109	0.121
	$\hat{\sigma}_{\text{var.}}$	0.590	0.624	0.628	0.654	0.609	0.643
	$Q-$ var.	0.590	0.624	0.628	0.654	0.609	0.643
	$\hat{\sigma}_{\ell sk}$	-0.208	0.245	-0.218	0.262	-0.205	0.245
	$\hat{\sigma}_{\ell ML}$	-0.024	0.091	-0.020	0.071	-0.006	0.047
	$\hat{\sigma}_{\text{la8}}$	0.036	0.037	0.038	0.039	0.037	0.038
	$\hat{\sigma}_{\text{haar}}$	0.044	0.045	0.045	0.046	0.045	0.046
$\sigma = 30$	BIC	0.087	0.224	0.102	0.232	0.086	0.190
	ICL	-0.332	0.474	-0.240	0.446	-0.166	0.367
	$\hat{\sigma}_{\text{var.}}$	0.330	0.332	0.296	0.337	0.333	0.336
	$Q-$ var.	0.330	0.332	0.293	0.340	0.325	0.335
	$\hat{\sigma}_{\ell sk}$	-0.213	0.263	-0.187	0.226	-0.211	0.244
	$\hat{\sigma}_{\ell ML}$	-0.018	0.069	-0.012	0.071	-0.004	0.073
	$\hat{\sigma}_{\text{la8}}$	0.016	0.017	0.018	0.019	0.020	0.022
	$\hat{\sigma}_{\text{haar}}$	0.059	0.060	0.056	0.057	0.056	0.057
$\sigma = 50$	BIC	0.139	0.139	0.141	0.141	0.143	0.143
	ICL	0.118	0.140	0.110	0.150	0.099	0.155
	$\hat{\sigma}_{\text{var.}}$	0.140	0.140	0.142	0.142	0.145	0.145
	$Q-$ var.	0.140	0.140	0.142	0.142	0.145	0.145
	$\hat{\sigma}_{\ell sk}$	-0.228	0.240	-0.206	0.226	-0.224	0.245
	$\hat{\sigma}_{\ell ML}$	-0.023	0.087	-0.016	0.077	-0.028	0.107
	$\hat{\sigma}_{\text{la8}}$	-0.091	0.091	-0.088	0.088	-0.085	0.085
	$\hat{\sigma}_{\text{haar}}$	0.096	0.097	0.088	0.089	0.085	0.086

The wavelet-based methods perform very creditably, on the whole. Additionally, there is very little difference between the choice of the wavelets: both $\hat{\sigma}_{\mathrm{haar}}$ and $\hat{\sigma}_{\mathrm{la8}}$ perform similarly and very well. Indeed they are the best performers in several settings, outperforming sometimes even the EM-based methods.

The performance of the local-estimation-based methodology of Rajan et al. (2010) is more mixed. Figure 1 and Table 2 both indicate that the performance of the local-MLE-based methodology is substantially better than that of the heuristic local-skewness-estimation-based approaches. While still generally competitive, the local-MLE-based methodology is a bit worse than the wavelet-based or the BIC-approaches. We note that our evaluations here have been based following exactly the recipe of Rajan et al. (2010) – this includes the rather simplistic way of selecting the mode of the local estimates. It would be of interest to evaluate performance using more sophisticated mode-selection approaches.

The results of our experiments indicate good performance for our EM methods: indeed, our EM approach using BIC is very competitive with the wavelet-based and local-MLE-based methods. We now evaluate performance of our methodology on the two-dimensional physical phantom datasets introduced in Maitra and Faden (2009).

3.2.2 Physical Phantom Data

Our next set of evaluations was on the two-dimensional physical phantom of Maitra and Faden (2009) scanned in a Siemens 3T Magnetom Trio Scanner using a 12-channel head array coil with a gradient echo (GRE) sequence with echo time (TE) of 10ms, relaxation time (TR) 180ms and flip angle of 7°. The field-of-view (FOV) of the scanned phantom 256mm$\times$256mm and the images were acquired at a resolution of 1mm$\times$1mm. (For a display of the twelve magnitude images obtained, see Figure 3 of Maitra and Faden (2009).) Background regions were carefully drawn by visual inspection on these phantom datasets and σ was estimated for each channel. These twelve σs formed the "ground truth" for this experiment. Finally, we used an offset of $m = 4$ pixels, reducing the sample size considered in our estimation to be $n = 4096$ pixels. Table 3 provides the results of our experiments which are also displayed in Figure 2.

Table 3: Results on estimating σ for the phantom dataset using the mixture-modeling-based methods introduced in this paper as well as Rajan et al. (2010)'s local-noise-estimation-based and Coupè et al. (2010)'s wavelet-based noise estimation methods. Missing results are for cases where equation (7) returned a complex value in the course of the iterations.

Rep.	**True**	**BIC**	**ICL**	$\hat{\sigma}$**-var.**	Q**-var.**	$\hat{\sigma}_{\ell sk}$	$\hat{\sigma}_{\ell ML}$	$\hat{\sigma}_{\mathrm{la8}}$	$\hat{\sigma}_{\mathrm{haar}}$
1	1.19	1.10	0.97	1.18	13.63	1	1	2.02	2.26
2	1.43	1.21	1.10	1.51	1.21	1	1	2.75	3.68
3	1.00	0.94	0.73	0.95	0.95	1	1	3.37	-
4	0.70	0.68	0.38	0.65	0.81	1	1	1.11	0.53
5	0.82	0.82	0.97	0.88	0.87	1	1	0.64	3.43
6	0.60	0.63	0.74	0.78	0.58	1	1	0.18	-
7	1.25	1.24	1.31	1.41	1.24	1	1	2.39	2.12
8	1.32	1.40	1.47	2.08	1.43	1	1	-	-
9	0.93	1.04	0.94	1.21	0.94	1	1	-	-
10	0.70	0.73	0.73	0.82	0.69	1	1	-	-
11	0.90	0.85	0.94	0.98	0.94	1	1	1.68	-
12	0.69	0.71	0.50	0.71	0.90	1	1	1.34	0.15

Overall, it appears that the BIC estimates are closest to the "ground truth", with mean relative errors ($(\hat{\sigma} - \sigma)/\sigma$) of -0.6%, while those for the ICL, $\hat{\sigma}$-variability, and Q-variability estimates were -6.1%, 13.5% and 90.4%, respectively. (The high mean relative errors for the two variability-based estimates are each caused by one anomalous estimate: note that the median relative errors are on the order of -0.28%, 2.65%, 8.56% and 2.65% respectively.) Once again, therefore, the BIC-based estimation method is the best performer: indeed, it vastly outperforms the summaries in Maitra and Faden (2009).

The local-noise-estimation-based methods of Rajan et al. (2010) all report the value of unity. This may well be a consequence of the rounding introduced by them in the calculation of the mode and needs further investigation by comparing with more sophisticated mode-finding approaches. The performance of the wavelet-based methods is spotty even where the algorithm is able to provide estimates. Thus, the results of these experiments indicates that our

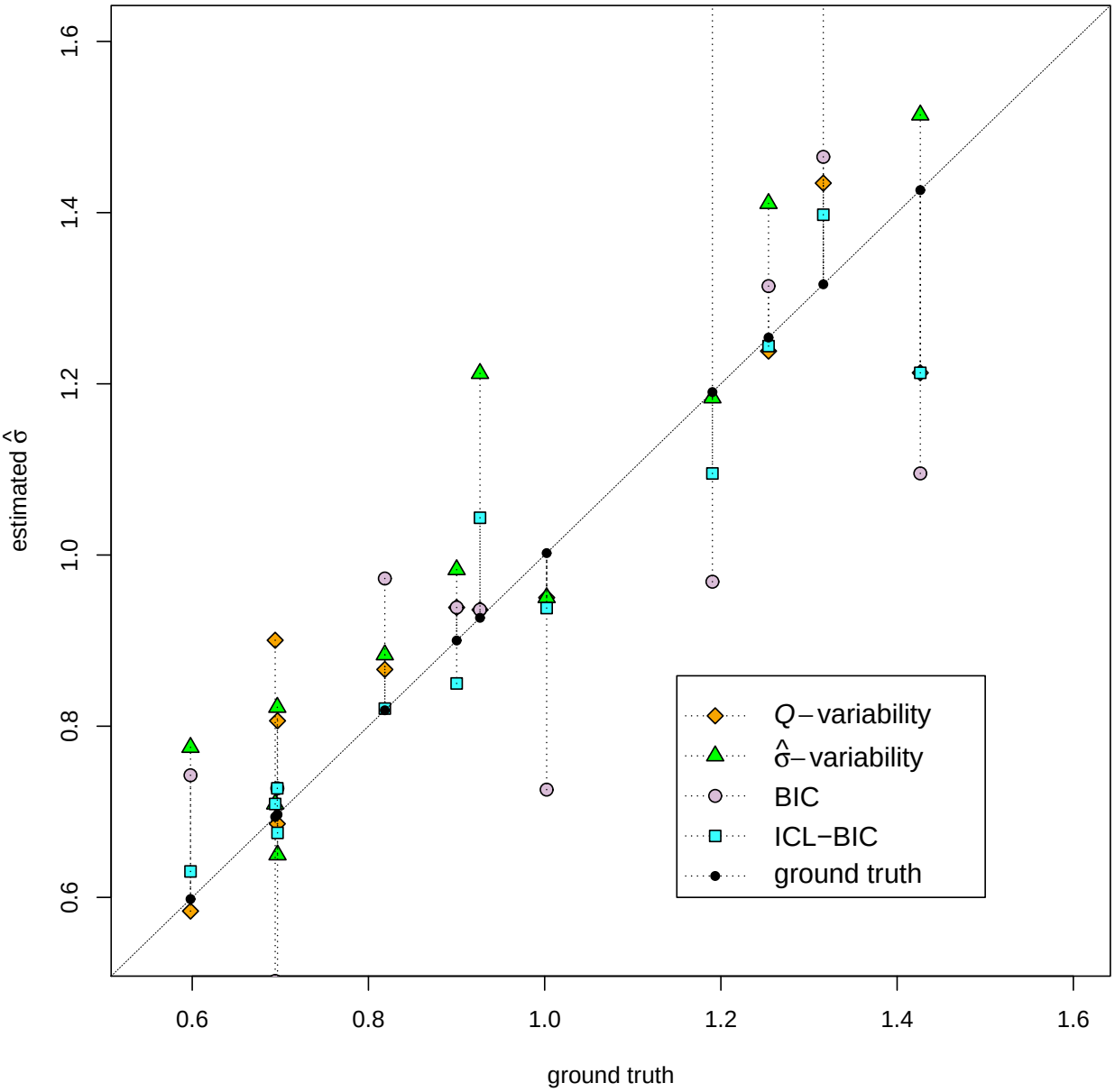

Figure 2: Estimates of σ obtained using the BIC-, ICL- and the $\hat{\sigma}$- and Q-variability-based methods plotted against the "ground truth". For greater clarity of presentation, we have not plotted estimates obtained using the methods of Rajan et al. (2010) or Coupè et al. (2010).

EM-based approach used in conjunction with BIC is the best performer. We now report performance on experiments done using four clinical datasets.

3.2.3 Application to Clinical Datasets

Our noise parameter estimation methodology was also evaluated on the four clinical magnitude MR datasets of Maitra and Faden (2009) that were all obtained using a GE 1.5T Signa scanner. Three of these datasets were from a registered set of ρ-, T_1- and T_2-weighted images on a healthy normal male volunteer using a spin-echo imaging sequence and acquired at a resolution of 1.15mm$\times$1.15mm$\times$7.25mm in a FOV set to be 294mm$\times$294mm$\times$145mm. Each of these

datasets had background regions which were carefully demarcated by an expert, from where the "ground truth" σ was estimated. The fourth dataset was on a MR breast scan on a female with suspected malignant lesion, acquired under TE/TR/flip angle settings of 2.54/4.98/12°. Image resolution was 0.8929mm×0.8929mm×1.25mm, with FOV at 400mm×400mm×220mm. A 187.5mm×117.8mm×220mm was cropped to exclude large non-breast regions of chest, air and so on – the resultant image had 210×132×176 voxels. The absence of background voxels for this dataset makes it difficult to compute the "ground truth" for comparison.

Table 4 summarizes estimates obtained using the different methods for estimating σ. For the $\boldsymbol{\rho}$-weighted MR

Table 4: Estimated σs on clinical datasets obtained using the EM-based, local-noise-estimation-based and wavelet-based approaches, along with their "ground truth" estimates (where available).

Dataset	**ground truth**	**BIC**	**ICL**	$\hat{\sigma}$**-var.**	Q**-var.**	$\hat{\sigma}_{\ell sk}$	$\hat{\sigma}_{\ell ML}$	$\hat{\sigma}_{\text{la8}}$	$\hat{\sigma}_{\text{haar}}$
$\boldsymbol{\rho}$-weighted	0.994	0.955	0.955	1.281	1.082	2	3	3.892	3.822
$\mathbf{T}_1$-weighted	0.833	0.921	0.921	1.251	0.983	2	3	3.40	3.277
$\mathbf{T}_2$-weighted	0.824	0.806	0.806	1.872	1.062	2	3	4.383	4.373
Breast	–	6.085	6.085	7.315	6.887	2	2	4.302	6.00

dataset, it appears that all methods, and especially the BIC, ICL and Q-variability estimation methods do a better job at estimating σ than in Maitra and Faden (2009). For the T_1- and T_2-weighted datasets, the BIC- and the ICL estimates are also quite competitive, though they are marginally worse than the estimates in Maitra and Faden (2009). The BIC- and ICL- estimates also reported smaller values for the estimates for the breast data than in Maitra and Faden (2009). The wavelet-based estimation methodology does surprisingly poorly, as do the estimation methods in Rajan et al. (2010). The reason for this is unclear, but we note that the underlying premise of their method is the existence of large homogeneous segments in the image. This aspect may have been violated. The methods discussed in this paper does not build on such assumptions and seem to perform very well.

In this section, we have demonstrated application of our σ-estimation methodology to a large number of phantom datasets as well as on four three-dimensional clinical datasets. Our estimates, especially using BIC, were the closest to the "ground truth" values when the latter was available, and provide some confidence in the performance of our refined methodology. An added plus of our estimation method over the others is its unique ability to readily provide an estimate of the standard error of the estimated $\hat{\sigma}$.

4 Conclusions

In this paper, we provide a refinement of the automated methodology developed in Maitra and Faden (2009) for estimating the noise parameter in magnitude MR images that is applicable irrespective of whether there is a substantial number of background voxels in the image. Our refinement consists of recognizing that the data arising in magnitude MR images is from complex k-space data, for which the phase information has been discarded. Using these observations as additional missing information, we develop a EM algorithm which has the advantage that the complete data belongs to a (Gaussian) REF. Thus, the M-step in the EM algorithm is of closed form, and we are led to an algorithm that does not make use of iterative computationally demanding and potentially unstable maximization steps. The EM algorithm requires initializing parameter values for which we have provided an intelligent and informed stochastic approach. It also provides, very readily, standard errors of our estimates. We have also detailed two broad approaches to estimating the number of components in the mixture model. Performance on experiments on simulated and physical phantom data as well on four clinical datasets was very encouraging, with BIC-based estimation as a consistent top-performer. When compared with recent methodology introduced by Rajan et al. (2010) and Coupè et al. (2010), our methodology using BIC and ICL is quite competitive in simulation experiments and vastly outperforms the others on two-dimensional physical phantom and three-dimensional clinical MR datasets.

A few points need to be made in this context. First, we note that because our algorithm no longer has iterative numerical methods for implementing the M-step, this means that computations are substantially faster than before. Additionally, we note that though not implemented here, the EM algorithm can be substantially sped up using acceleration methods as in Louis (1982) or McLachlan and Krishnan (2008). While also not pursued in this paper, we note that the estimates of the signal and associated clustering probabilities provide the ingredients for a model-based segmentation algorithm. Another issue pertains to smoothing and dependent data. We have tried to address this concern

by sampling from a sub-grid with offset l (chosen to be 8 in our simulation experiments). It may be possible to explicitly include the dependence structure in our estimation. This is especially true in the context of image segmentation, where the goal is to classify every voxel, unlike the estimation of one parameter (σ), so that a coarser sub-grid may not be possible. Separately, it may be desirable to investigate and develop further the mode-finding methodology in the local-noise-estimation-based methodology of Rajan et al. (2010). Further, as a reviewer has very kindly pointed out, our suggested approaches may not be applicable in the context of images (such as in parallel imaging) with a spatially-varying noise parameter. It would be of interest to investigate performance and to develop methodology in this context. Thus, while a promising automated method for noise estimation in magnitude MR images has been developed, a few issues meriting further attention remain.

A Appendix: Conditional distribution of γ_i given R_i

We first state and prove the following

Theorem A.1. *Let $\boldsymbol{W}_i$ be a realization from the one-trial J-class multinomial distribution with class probability vector $\boldsymbol{\pi}$. Conditional on $W_{i,j} = 1$ for some $j = 1, 2, \ldots, J$, let $U_i \sim N(\nu_j, \sigma^2)$ be independent of $V_i \sim N(0, \sigma^2)$. Write (U_i, V_i) in its polar form* i.e. $(U_i, V_i) = (R_i \cos\gamma_i, R_i \sin\gamma_i)$. *Then the conditional distribution of γ_i given R_i and the event $W_{i,j} = 1$ is $\mathcal{M}(0, R_i\nu_j/\sigma^2)$,* i.e. *it is von-Mises-distributed with mean angular direction 0 and concentration parameter $R_i\nu_j/\sigma^2$. Consequently, $\mathbb{E}(\cos\gamma_i \mid R_i, W_{i,j} = 1) = I_1(\frac{R_i\nu_j}{\sigma^2})/I_0(\frac{R_i\nu_j}{\sigma^2})$.*

Proof. From the characterization of the Rice distribution in Rice (1944, 1945), we know that $R_i \sim \varrho(x; \sigma, \nu_j)$ given that $W_{i,j} = 1$. From standard results on probability distributions of transformation of variables, we get that the conditional joint density of (R_i, γ_i) given that $W_{i,j} = 1$ is $f_{R_i,\gamma_i|W_{i,j}=1}(x, \gamma) = \frac{x}{2\pi\sigma^2} \exp[-(x^2 - 2x\nu_j \cos\gamma_i + \nu_j^2)/2\sigma^2]$.Thus $f(\gamma_i|R_i = x, W_{i,j} = 1) = \exp[\frac{x\nu_j}{\sigma^2}\cos\gamma_i]/2\pi\mathrm{I}_0(\frac{x\nu_j}{\sigma^2})$ for $0 < \gamma_i < 2\pi$, which is the density of $\mathcal{M}(0, \frac{x\nu_j}{\sigma^2})$. From results in Mardia and Jupp (2000) on the expectation of the cosine of a von-Mises-distributed random variable, we get that $\mathbb{E}(\cos\gamma_i \mid R_i, W_{i,j} = 1) = \mathrm{I}_1(\frac{R_i\nu_j}{\sigma^2})/\mathrm{I}_0(\frac{R_i\nu_j}{\sigma^2})$. This proves Theorem A.1. □

Acknowledgment

I thank R. P. Gullapalli and S. R. Roys for the phantom and clinical datasets and to an Associate Editor and two reviewers whose helpful and insightful comments on an earlier version of this manuscript greatly improved its content. I also acknowledge partial support by the National Science Foundation Awards NSF CAREER DMS-0437555.

References

Ahmed, O. A. (2005). New denoising scheme for magnetic resonance spectroscopy signals. *IEEE Transactions on Medical Imaging*, 24(6):809–816.

Aja-Fernández, S., Tristán-Vega, A., and Alberola-Lòpez, C. (2009). Noise estimation in single- and multiple-coil magnetic resonance data based on statistical models. *Magnetic Resonance Imaging*, 27(10):1397–1409.

Altman, M., Gill, J., and McDonald, M. (2003). *Numerical Issues in Statistical Computing for the Social Scientist*. Wiley-Interscience, New York.

Bammer, R., Skare, S., Newbould, R., Liu, C., Thijs, V., Ropele, S., Clayton, D. B., Krueger, G., Moseley, M. E., and Glover, G. H. (2005). Foundations of advanced magnetic resonance imaging. *NeuroRx*, 2:167–196.

Biernacki, C., Celeux, G., and Gold, E. M. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:719–725.

Biernacki, C., Celeux, G., and Govaert, G. (2003). Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics and Data Analysis*, 413:561–575.

Brown, T. R., Kincaid, B. M., and Ugurbil, K. (1982). NMR chemical shift imaging in three dimensions. *Proceedings of the National Academy of Sciences, USA*, 79:3523–3526.

Brummer, M. E., Mersereau, R. M., Eisner, R. L., and Lewine, R. R. J. (1993). Automatic detection of brain contours in MRI data sets. *IEEE Transactions on Medical Imaging*, 12(2).

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal of Scientific Computing*, 16:1190–1208.

Cocosco, C., Kollokian, V., Kwan, R., and Evans, A. (1997). Brainweb: Online interface to a 3d MRI simulated brain database. *NeuroImage*, 5(4).

Coupè, P., Manjn, J. V., Gedamu, E., Arnold, D., Robles, M., and Collins, D. L. (2010). Robust Rician noise estimation for MR images. *Medical Image Analysis*, 14:483–493.

Daubechies, I. (1992). *Ten Lectures on Wavelets*. CBMS-NSF Regional Conference Series in Applied Mathematics. SIAM, Philadelphia.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

Donoho, D. L. (1995). De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, 41(3):613–627.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97:611–631.

Glad, I. K. and Sebastiani, G. (1995). A Bayesian approach to Synthetic Magnetic Resonance Imaging. *Biometrika*, 82(2):237–250.

Hartigan, J. A. and Wong, M. A. (1979). A k-means clustering algorithm. *Applied Statistics*, 28:100–108.

Hennessy, M. J. (2000). A three-dimensional physical model of MRI noise based on current noise sources in a conductor. *Journal of Magnetic Resonance*, 147:153169.

Hinshaw, W. S. and Lent, A. H. (1983). An introduction to NMR imaging: From the Bloch equation to the imaging equation. *Proceedings of the IEEE*, 71(3).

Keribin, C. (2000). Consistent estimation of the order of finite mixture models. *Sankhyā, Series A*, 62:49–66.

Koay, C. G. and Basser, P. J. (2006). Analytically exact correction scheme for signal extraction from noisy magnitude MR signals. *Journal of Magnetic Resonance*, 179(2):317322.

Ljunggren, S. (1983). A simple graphical representation of fourier-based imaging methods. *Journal of Magnetic Resonance*, 54:338–343.

Louis, T. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 44(2):226–233.

MacQueen, J. (1967). Some methods of classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297.

Maitra, R. (2009). Initializing partition-optimization algorithms. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6:144–157.

Maitra, R. and Faden, D. (2009). Noise estimation in magnitude MR datasets. *IEEE Transactions on Medical Imaging*, 28(10):1615–1622.

Maitra, R. and Melnykov, V. (2010). Simulating data to study performance of finite mixture modeling and clustering algorithms. *Journal of Computational and Graphical Statistics*, page in press.

Maitra, R. and Riddles, J. J. (2010). Synthetic Magnetic Resonance Imaging revisited. *IEEE Transactions on Medical Imaging*, 29(3):895 –902.

Mardia, K. V. and Jupp, P. E. (2000). *Directional Statistics*. Wiley, New York.

McLachlan, G. and Krishnan, T. (2008). *The EM Algorithm and Extensions*. Wiley, New York, second edition.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. John Wiley and Sons, Inc., New York.

McVeigh, E. R., Henkelman, R. M., and Bronskill, M. J. (1985). Noise and filtration in Magnetic Resonance Imaging. *Medical Physics*, 12(5):586–591.

Pasquale, F. d., Barone, P., Sebastiani, G., and Stander, J. (2004). Bayesian analysis of dynamic magnetic resonance breast images. *Journal Of The Royal Statistical Society Series, Series C*, 53(3):475–493.

Rajan, J., Poot, D., Juntu, J., and Sijbers, J. (2010). Noise measurement from magnitude MRI using local estimates of variance and skewness. *Physics in Medicine and Biology*, 55:N441–449.

Rice, S. O. (1944). Mathematical analysis of random noise. *Bell System Technical Journal*, 23:282.

Rice, S. O. (1945). Mathematical analysis of random noise. *Bell System Technical Journal*, 24:46–156.

Rohdea, G. K., Barnettc, A. S., Bassera, P. J., and Pierpaoli, C. (2005). Estimating intensity variance due to noise in registered images: Applications to diffusion tensor MRI. *NeuroImage*, 26:673–684.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Sijbers, J. (1998). *Signal and Noise Estimation from Magnetic Resonance Images*. PhD thesis, University of Antwerp.

Sijbers, J. and den Dekker, A. J. (2004). Maximum likelihood estimation of signal amplitude and noise variance from MR data. *Magnetic Resonance in Medicine*, 51:586–594.

Sijbers, J., den Dekker, A. J., Van Audekerke, J., Verhoye, M., and Van Dyck, D. (1998). Estimation of the noise in magnitude MR images. *Magnetic Resonance Imaging*, 16(1):87–90.

Sijbers, J., Poot, D., den Dekker, A. J., and Pintjens, W. (2007). Automatic estimation of the noise variance from the histogram of a magnetic resonance image. *Physics in Medicine and Biology*, 52:1335–1348.

Smith, R. C. and Lange, R. C. (2000). *Understanding Magnetic Resonance Imaging*. CRC Press LLC.

Tweig, D. B. (1983). The k-trajectory formulation of the NMR imaging process with applications in analysis and synthesis of imaging methods. *Medical Physics*, 10(5):610–21.

Wang, T. and Lei, T. (1994). Statistical analysis of MR imaging and its application in image modeling. In *Proceedings of the IEEE International Conference on Image Processing and Neural Networks*, volume 1, pages 866–870.

Weishaupt, D., Köchli, V. D., and Marincek, B. (2003). *How Does MRI Work?* Springer–Verlag, New York.

Wilde, J. P. D., Lunt, J., and Straughan, K. (1997). Information in magnetic resonance images: evaluation of signal, noise and contrast. *Medical and Biological Engineering and Computing*, 35:259–265.

Zhang, Y., Brady, M., and Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation maximization algorithm. *IEEE Transactions on Medical Imaging*, 20(1):45–47.

Zhu, C., Byrd, R. H., Lu, P., and Nocedal, J. (1994). L-BFGS-B – Fortran subroutines for large-scale bound constrained optimization. Technical report, Northwestern University.