Reliable Selection of the Number of Fascicles in Diffusion Images by Estimation of the Generalization Error*

Benoit Scherrer**, Maxime Taquet**, and Simon K. Warfield

Computational Radiology Laboratory, Department of Radiology
Boston Children's Hospital, 300 Longwood Avenue, Boston, MA, 02115, USA

Abstract. A number of diffusion models have been proposed to overcome the limitations of diffusion tensor imaging (DTI) which cannot represent multiple fascicles with heterogeneous orientations at each voxel. Among them, generative models such as multi-tensor models, CHARMED or NODDI represent each fascicle with a parametric model and are of great interest to characterize and compare white matter properties. However, the identification of the appropriate model, and particularly the estimation of the number of fascicles, has proven challenging. In this context, different model selection approaches have been proposed to identify the number of fascicles at each voxel. Most approaches attempt to maximize the quality of fit while penalizing complex models to avoid overfitting. However, the choice of a penalization strategy and the trade-off between penalization and quality of fit are rather arbitrary and produce highly variable results. In this paper, we propose for the first time to determine the number of fascicles at each voxel by assessing the generalization error. This criterion naturally prevents overfitting by comparing how the models predict new data not included in the model estimation. Since the generalization error cannot be directly computed, we propose to estimate it by the 632 bootstrap technique which has low bias and low variance. Results on synthetic phantoms and in vivo data show that our approach performs better than existing techniques, and is robust to the choice of decision threshold. Together with generative models of the diffusion signal, this technique will enable accurate identification of the model complexity at each voxel and accurate assessment of the white matter characteristics.

1 Introduction

Diffusion tensor imaging (DTI) is well known to be unable to represent the diffusion signal arising from multiple fascicles crossing in one voxel. Various approaches have been proposed to overcome this limitation. Among them, generative models such as multi-tensor models [12,9], CHARMED [2] or NODDI

* This work was supported in part by NIH grants R01 RR021885, R01 EB008015, R03 EB008680, R01 LM010033, UL1 RR025758-03 and 1U01NS082320. MT is supported by F.R.S-FNRS.

** These authors contributed equally.

J.C. Gee et al. (Eds.): IPMI 2013, LNCS 7917, pp. 742–753, 2013.

[13] seek to represent the signal contribution from different populations of water molecules such as the signal contribution from unrestricted diffusion and from each individual fascicle. These models are based on underlying biological assumptions and are of great interest to characterize and compare white-matter properties. For example, assessment of the free water diffusion arising from the extracellular space may be useful for the characterization of edema or inflammation [8]. Modeling of each individual fascicle may be useful to characterize properties such as the fascicle density, the axonal diameter distribution or the myelin integrity. This is not feasible with the tensor representation of the signal in DTI which conflates the signal contribution from multiple sources. However, accurate estimation of a generative model requires identification of the number of fascicles present in each voxel, which corresponds to identifying the appropriate model complexity. This remains a challenging and open problem.

Assessing the number of fascicles at each voxel is a model selection problem. To date, most approaches select between diffusion models of different complexity by minimizing the fitting error. Because both the model estimation and assessment is achieved on the same dataset, this favors complex models and may overfit the data. For this reason, the criterion usually integrates a component penalizing complex models.

The most common criterion for the selection of generative diffusion models is the F-test. Alexander *et al.* [1] compare the spherical harmonic expansion of the ADC truncated at different orders by means of a series of ANOVA F-Tests. In this strategy, complex models are penalized by their necessity to significantly decrease the fitting error when compared to simpler models. Kreher *et al.* [5] use F-tests to select the appropriate number of fascicles by observing the variance of the ADC. Scherrer and Warfield [9] use a similar F-test strategy applied to the signal residuals rather than the ADC. Besides the F-test, other approaches based on the quality of fit have been proposed. Behrens *et al.* [3] used a Bayesian Automatic Relevance Determination (ARD) approach which starts with the most complex model and gradually prune the unnecessary variables. However, this was shown inefficient in tractography and required to manually force the number of fascicles [6]. The Bayesian Information Criterion (BIC), a weighted sum of the fitting error and a penalizing term, has been suggested as well but was shown to yield suboptimal results, even on synthetic data [10].

A more reliable paradigm to avoid overfitting when selecting between models is to compare how each model performs for *new data* not included in the model estimation. This relates to the *generalization error* (GE). Typically, a model not complex enough to represent a dataset will have a large GE, and so will a too complex model which overfits the data. To the best of our knowledge, minimization of the GE has never been used to determine the number of fascicles at each voxel. Leave-one-out cross validation follows this paradigm. However, it does not lead to a consistent estimate of the model [11] and its results are highly variable. Other cross-validation methods, such as K-fold cross validation, reduce this variance at the cost of a higher bias. By contrast, the 632 bootstrap method proposed by Efron [4] reduces this variability while remaining almost unbiased.

In this paper we use for the first time the 632 bootstrap (B632) method to determine the number of fascicles present in each voxel from diffusion-weighted images. Section 2 introduces the methods used, from the definition of the generalization error (Section 2.1) to the use of its estimate in the selection of the number of fascicles (Section 2.5). Section 3 presents results on both synthetic phantoms and in vivo data. Finally, Section 4 discusses the results and concludes.

2 Material and Methods

In this section, we present the generalization error minimization framework and how it can be utilized to perform model selection and determine the number of fascicles in each voxel. We start by explaining the fundamental difference between generalization error and fitting error. We then introduce different methods to estimate the generalization error, and explain why the 632 bootstrap should be used in this context. We subsequently provide an expression to estimate the variance of generalization error. Finally, we detail how these methods are applied to the problem of selecting the number of fascicles at each voxel.

2.1 Generalization Error and Fitting Error

Let $\boldsymbol{z} = \{z_1, ..., z_n\}$ with $z_i = (x_i, y_i)$ be the set of n observed data points, in which x_i are inputs to the model (*e.g.*, the b-values and gradient directions in diffusion images, see Section 2.5) and y_i are outputs (*e.g.*, the signal attenuation in diffusion images). These data are used to build a generative model $r_{\boldsymbol{z}}(x)$ that tries to predict the output y from an input x. Ideally, the optimal model would minimize the *generalization error*, that is the error made on a new hypothetical data point $z_0 = (x_0, y_0)$. The generalization error conditional on the observed data is :

$$E_g|\boldsymbol{z} = E_{z_0 \sim F}\left[|y_0 - r_{\boldsymbol{z}}(x_0)|^2 \middle| \boldsymbol{z}\right], \tag{1}$$

where $E[.]$ is the statistical expectation and $z_0 \sim F$ indicates that the expectation is taken over the new data point that follows the distribution F. To account for the variability of the observed data points, the unconditional generalization error can be defined as the expectation of (1) over all $\boldsymbol{z}$:

$$E_{g,n} = E_{z_i \overset{\text{iid}}{\sim} F}\Big\{E_g|\boldsymbol{z}\Big\} = E_{z_i \overset{\text{iid}}{\sim} F}\Big\{E_{z_0 \sim F}\left[|y_0 - r_{\boldsymbol{z}}(x_0)|^2 \middle| \boldsymbol{z}\right]\Big\}, \tag{2}$$

where the index n indicates that n samples were used to optimize the model $r_{\boldsymbol{z}}$. The generalization errors (1) and (2) cannot be directly computed because the distribution F is unknown. One simple solution would be to estimate $F(z)$ by the empirical distribution $\hat{F}(z) = \frac{1}{n} \forall\, z \in \boldsymbol{z}$. For the conditional generalization error (1), this yields the following estimate:

$$\begin{aligned} \hat{E}_g^{\text{fit}} &= E_{z_0 \sim \hat{F}}\left[|y_0 - r_{\boldsymbol{z}}(x_0)|^2 \middle| \boldsymbol{z}\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} |y_i - r_{\boldsymbol{z}}(x_i)|^2, \end{aligned} \tag{3}$$

that is the common *fitting error*. This estimate is a biased estimate of E_g since the data $\boldsymbol{z}$ are used both to optimize the parameters of the model $r_{\boldsymbol{z}}$ and to estimate its error. In particular, in many modeling problem (including multi-fascicle modeling), it is always possible to find a model that yields $\hat{E}_g^{\text{fit}} = 0$ provided that it is complex enough. In the following sections, we will therefore explore other estimates of E_g and investigate their bias and variance.

2.2 Cross-Validation Estimates

To circumvent the overfitting problem of $\hat{E}_g^{\text{fit}}$, one could estimate the model by omitting one data point in the training sample $\boldsymbol{z}$ and evaluate the model prediction for this data point. This is the idea behind the leave-one-out cross-validation (LOOCV) method. Let $\boldsymbol{z}_{(-i)}$ be the training samples without z_i. The resulting estimate of the generalization error reads:

$$\hat{E}_g^{\text{CV}} = \frac{1}{n} \sum_{i=1}^{n} |y_i - r_{\boldsymbol{z}_{(-i)}}(x_i)|^2. \tag{4}$$

This is an unbiased estimator of the generalization error $E_{g,n-1}$. For large n, the bias of $\hat{E}_g^{\text{CV}}$ as an estimator of $E_{g,n}$ is positive but low. Its variance, however, is large, leading to high root mean squared errors, despite the low bias [4].

The variance of the LOOCV can be decreased by keeping more than one element out of the dataset at each iteration of model training and testing. This defines K-fold cross-validation methods. They result in unbiased estimates of $E_{g,n-\frac{n}{K}}$ which, however, present an increased bias for the estimation of E_g.

2.3 632 Bootstrap

The bootstrap smoothing method can be used to lower the variance of the cross-validation estimate [4]. This technique estimates E_g in (2) by providing two different estimates for the distribution F. The distribution F of z_0 is estimated by the empirical distribution $\hat{F}$ and the distribution F of $\boldsymbol{z}$ is estimated from bootstrap samples $\boldsymbol{z}^*_{(-i)}$ of the empirical distribution $\hat{F}_{(-i)}$ which excludes the sample z_i used for testing. Formally, an expression of the estimator comes by inverting the order of the expectations in (2) and by subsequently replacing the distributions F by their estimates:

$$\hat{E}_g^{\text{BS}} = \frac{1}{n} \sum_{i=1}^{n} E_{\hat{F}_{(-i)}} \left[|y_i - r_{\boldsymbol{z}^*_{(-i)}}(x_i)|^2 \right].$$

An equivalent expression of $\hat{E}_g^{\text{BS}}$ that is closer to its implementation is:

$$\hat{E}_g^{\text{BS}} = \sum_{i=1}^{n} \left[\sum_{b=1}^{B} \delta(N_i^b) \left| y_i - r_{\boldsymbol{z}^*_{(-i)}}(x_i) \right|^2 \Bigg/ \sum_{b=1}^{B} \delta(N_i^b) \right], \tag{5}$$

where N_i^b is the number of times sample i is used in the training set of the b^{th} bootstrap replicate and $\delta(x)$ is the Dirac function. The factor $\delta(N_i^b)$ guarantees that sample i can be used as a testing sample in bootstrap replicate b.

Much like cross-validation, the bootstrap estimate is biased because it relies on fewer point than the number n of available samples. LOOCV uses $(n-1)$ points and its bias is therefore limited. By contrast, $\hat{E}_g^{\text{BS}}$ uses, on average, $[1-(1-\frac{1}{n})^n]n$ points which is approximately equal to $0.632n$ for large n. This makes the bias of $\hat{E}_g^{\text{BS}}$ more critical. Efron [4] proposed to counterbalance the positive bias of $\hat{E}_g^{\text{BS}}$ by the negative bias of $\hat{E}_g^{\text{fit}}$, introducing the 632 bootstrap estimator:

$$\hat{E}_g^{632} = 0.368\,\hat{E}_g^{\text{fit}} + 0.632\,\hat{E}_g^{\text{BS}}\,. \tag{6}$$

The coefficients are defined so that the testing samples used to estimate $\hat{E}_g^{632}$ are at the same average distance from the training sample as would be a random point drawn directly from F [4]. This estimator has outperformed others in many applications, mostly when the signal-to-noise ratio is moderate to low [7].

2.4 Standard Error of the Difference Estimator

In many model selection problems including the estimation of the number of fascicles from DWI, model classes are nested: simpler models are particular cases of complex models. A more complex model can be arbitrarily close to a simpler one and its improvement of the generalization error may be due to chance alone. To reliably select between the two models, we need to assess whether this improvement is statistically significant.

Assessing the statistical significance of the difference in generalization error estimates between a model A and a model B, $\hat{\Delta}_{AB}^{632} = \hat{E}_{g,A}^{632} - \hat{E}_{g,B}^{632}$, requires the standard error of this difference to be estimated. After rearranging the terms of $\hat{\Delta}_{AB}^{632}$, taking advantage of the linearity of the expectation, we have:

$$\begin{aligned}
\hat{\Delta}_{AB}^{632} &= \frac{0.368}{n}\sum_{i=1}^{n} E_{z_i \overset{\text{iid}}{\sim} \hat{F}}\left[\left|y_i - r_{\boldsymbol{z}}^A(x_i)\right|^2 - \left|y_i - r_{\boldsymbol{z}}^B(x_i)\right|^2\right] \\
&\quad + \frac{0.632}{n}\sum_{i=1}^{n} E_{z_i \overset{\text{iid}}{\sim} \hat{F}_{(-i)}}\left[\left|y_i - r_{\boldsymbol{z}^*_{(-i)}}^A(x_i)\right|^2 - \left|y_i - r_{\boldsymbol{z}^*_{(-i)}}^B(x_i)\right|^2\right] \\
&\triangleq \frac{0.368}{n}\sum_{i=1}^{n}\hat{\Delta}_{AB,i}^{\text{fit}} + \frac{0.632}{n}\sum_{i=1}^{n}\hat{\Delta}_{AB,i}^{\text{BS}} \triangleq 0.368\,\hat{\Delta}_{AB}^{\text{fit}} + 0.632\,\hat{\Delta}_{AB}^{\text{BS}}
\end{aligned} \tag{7}$$

One could estimate the standard error of $\hat{\Delta}_{AB}^{\text{BS}}$ as $[\sum_i(\hat{\Delta}_{AB,i}^{\text{BS}} - \hat{\Delta}_{AB}^{\text{BS}})^2/n^2]^{1/2}$. This would assume that the $\hat{\Delta}_{AB,i}^{\text{BS}}$ are independent, which is not the case. A better estimate can be obtained by the *delta-method-after-bootstrap* approach [4]. This method is nonparametric and allows the computation of the standard error for any statistics that (1) is smooth in the observed data $\boldsymbol{z}$, (2) is invariant under permutations of the points z_i and, (3) only depends on the empirical

distribution $\hat{F}$. With this method, one can show that the standard error of $\hat{\Delta}^{\mathrm{BS}}_{AB}$ can be estimated by:

$$\hat{SE}^{\mathrm{BS}} = \left[\sum_{i=1}^{n} \hat{D}_i^2\right]^{1/2} \quad \text{with} \quad \hat{D}_i = \left(2 + \frac{1}{n-1}\right) \frac{\hat{\Delta}^{\mathrm{BS}}_{AB,i} - \hat{\Delta}^{\mathrm{BS}}_{AB}}{n} + \frac{\sum_{b=1}^{B} (N_i^b - \bar{N}_i)\bar{q}^b}{\sum_{b=1}^{B} \delta(N_i^b)}$$

$$\text{and} \quad \bar{q}^b = \sum_{i=1}^{n} \delta(N_i^b) \left[\left| y_i - r^A_{\boldsymbol{z}^*_{(-i)}}(x_i) \right|^2 - \left| y_i - r^B_{\boldsymbol{z}^*_{(-i)}}(x_i) \right|^2\right],$$

where $\bar{N}_i$ is the average N_i^b over all B bootstrap replicates. The same approach cannot be used for the standard error of $\hat{\Delta}^{\mathrm{fit}}_{AB}$ because it is a non-smooth function of the samples $\boldsymbol{z}$. Efron [4] proposes to estimate the standard error of $\hat{\Delta}^{632}_{AB}$ as:

$$\hat{SE}^{632} \approx \frac{\hat{\Delta}^{632}_{AB}}{\hat{\Delta}^{\mathrm{BS}}_{AB}} \hat{SE}^{\mathrm{BS}}. \tag{8}$$

To infer whether a model A is better than a model B, the estimate of the difference between their generalization errors (7) can be compared to the estimate of the standard error of this difference (8). This is the cornerstone of the selection of the number of fascicles problem.

2.5 Selection of the Number of Fascicles

The main idea to identify the number of fascicles at each voxel is to progressively increase the complexity of the model as long as a substantial decrease in the generalization error can be achieved and to stop when the decrease is no more significant or when the generalization error starts to increase. More specifically, the steps for the selection of the number of fascicles are:

1. For each pair of consecutive models (model with $m-1$ fascicles and m fascicles), compute the difference of generalization error estimates using (7). Let $\Delta_m = \hat{\Delta}^{632}_{m-1,m}$ be this difference. To use expression (5) for this estimate, the bootstrap replicates must be identical for all models.
2. Compute the standard error s_m of the estimate Δ_m using (8).
3. Select the model with m_{opt} fascicles such that:

$$m_{\mathrm{opt}} = \inf\left\{m \middle| \Delta_m - \theta^{\mathrm{B632}} s_m \geq 0, \Delta_{m+1} - \theta^{\mathrm{B632}} s_{m+1} < 0\right\}, \tag{9}$$

   where θ^{B632} is the number of standard error above which the difference should be to be deemed significant. For this expression to hold, we set $s_0 = \Delta_0 = \Delta_M = 0$ and $s_M = 1$, where M is the maximum number of fascicles.

We could compare every model to every other one and extend the selection rule (9) to express the need for a model to be significantly better than all the simpler ones and not significantly worse than the more complex ones. In our experiment, we did not observe a difference between the two rules. Rule (9) has the advantage that it can be applied as the model complexity is progressively increased, avoiding the need to optimize further complex models at voxels where a simple model has already been selected.

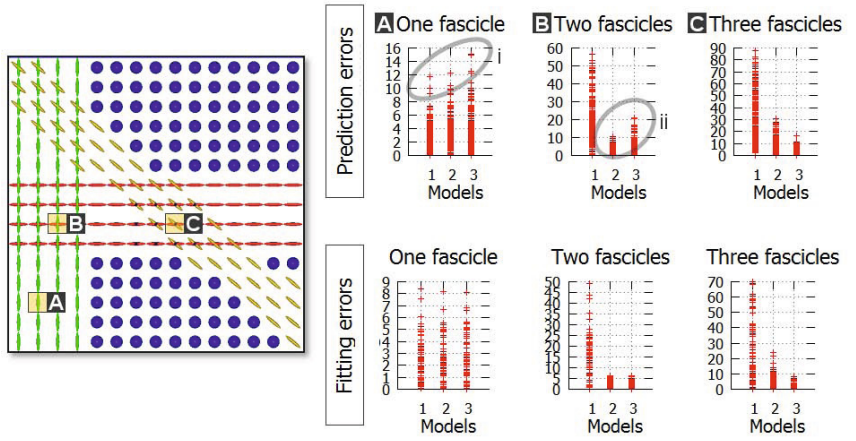


Fig. 1. (Left) Synthetic phantom used in our experiment, containing isotropic areas (blue balls), 1-fascicle areas [A], 2-fascicle areas [B] and three fascicle areas [C]. (Right) The repartition of the prediction errors, as used in the 632 bootstrap estimator, discriminates the three models, since more complex models may have higher prediction errors. On the contrary, the fitting error, as used in the F-test model selection, always decreases when the model complexity increases, due to overfitting.

2.6 Experimental Setup

Our proposed approach for the selection of the number of fascicles could be used with any generative model of the diffusion signal. In this work, we considered a multi-fascicle model in which each fascicle is represented by a tensor and the diffusion of free water is represented by an isotropic tensor. This amounts to considering the following generative model for the formation of the diffusion signal S for a b-value b and a gradient direction $\boldsymbol{g}$:

$$S = S_0 \left(f_0 e^{-bD_{\text{iso}}} + \sum_{i=1}^{m} f_i e^{-b\boldsymbol{g}^T \boldsymbol{D}_i \boldsymbol{g}} \right),$$

where S_0 is the signal with no diffusion sensitization applied, $D_{\text{iso}} = 3.0 \times 10^{-3}\text{mm}^2/s$ is the diffusion of free water in the brain at 37°C, f_i are the volumetric fractions of occupancy associated with each compartment ($\sum_{i=0}^{m} f_i = 1$), m is the number of fascicles of the model and $\boldsymbol{D}_i$ is the tensor representing the fascicle $i \in [1, m]$. Such model requires a DWI acquisition that images multiple non-zero b-values [9]. We employed the CUSP gradient encoding scheme [9] which achieves short echo time and high SNR. This acquisition was composed of five $b = 0$, 30 DW images at $b = 1000$ and 30 DW images between $b = 1000$ and $b = 3000$. The parameters of each model were estimated using a maximum a posteriori approach. We focused on model complexity ranging from $m = 0$ (isotropic diffusion) to $m = 3$ fascicles. We investigated the performance of our B632 model selection approach with both synthetic phantoms and in vivo data.

We compared it to the F-test on the signal residuals [9], for which the null hypothesis is that the fitting error of models with $m-1$ and m fascicles are equivalent by assessing the F-score:

$$F_{m-1,m} = \frac{n-1-|\mathcal{M}_m|}{|\mathcal{M}_m|-|\mathcal{M}_{m-1}|}\frac{\mathrm{SSE}_{m-1}-\mathrm{SSE}_m}{\mathrm{SSE}_{m-1}} > \theta^{\mathrm{F-test}}, \tag{10}$$

where n is the number of data, $\theta^{\mathrm{F-test}}$ is the F-score threshold above which the null hypothesis is rejected, and $|\mathcal{M}_m|$ and SSE_m are respectively the number of parameters and the sum of squared errors (fitting error) for a model with m fascicles. Various synthetic phantoms of size 15×15 were generated. The tensor profile $\boldsymbol{D}_i$ representing an individual fascicle was chosen to match typical in vivo data (trace of $2.1 \times 10^{-3}\mathrm{mm}^2/s$ and FA of 0.8). We considered regions with 0, 1, 2 and 3 fascicles (Fig. 1). The simulated DWI were corrupted by various Rician-noise levels. In vivo imaging was achieved on a healthy volunteer using a Siemens 3T Trio scanner with a 32-channel head coil and the following parameters : FOV=220mm, 68 slices, matrix=128×128, resolution=$1.72 \times 1.7 \times 2mm^3$.

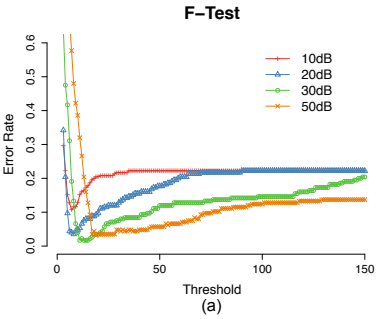
3 Results

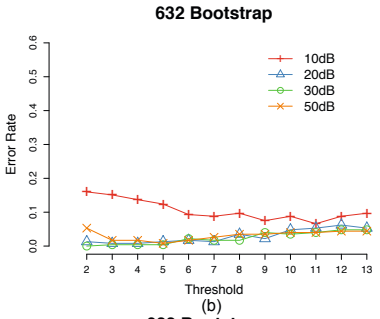
3.1 Synthetic Phantom Experiments

Synthetic phantoms offer a ground truth against which results of the model selection can be compared. We investigated the performance of the B632 and F-test approaches under four different SNR : 10dB, 20dB, 30dB and 50dB. Both the B632 and the F-test approaches require determination of a threshold (see (9) and (10)). We investigated the influence of $\theta^{\mathrm{F-test}}$ by evaluating the F-test model selection with $3 < \theta^{\mathrm{F-test}} < 150$. Similarly, we evaluated the influence of θ^{B632} by computing the B632 model selection with $0 < \theta^{\mathrm{B632}} < 13$. Note that $\theta^{\mathrm{F-test}}$ is a threshold on the F-score while θ^{B632} is the number of standard error above which a model m is considered better than a model $m-1$ (see (9)). The maximum number of bootstrap replicates for B632 was set to 150. We counted the number of errors between the ground truth and the automatic model selection results and reported the error rate.

The overall minimum error obtained by choosing the optimal thresholds is consistently higher with the F-test (Fig. 2a) than with B632 (Fig. 2b). The table in Fig. 2c summarizes those errors. In practice, the SNR is unknown and so the choice of threshold cannot depend on it. The overall minimum error rate are therefore lower bounds for what can actually be achieved in practice. The increase in error rate compared to this lower bound, due to the choice of a single threshold, is more dramatical with the F-test than with B632 (Fig. 2c, bottom rows). In particular, at 10dB, the error rate almost doubles compared to its lower bound with the F-test, while it increases only by a few percents with B632.

Finally, the evolution of the error rate with the number of bootstrap replicates assesses the stability of the estimate and allows the definition of a minimum number of bootstrap replicates required to achieve good performances. Fig. 2d shows that the error rate becomes stable after approximately 50 replicates.





		10dB	20dB	30dB	50dB
overall minimum	F-Test	24 (10%)	8 (3%)	4 (2%)	7 (3%)
	632 Bootstrap	**15 (7%)**	**2 (0.9%)**	**0 (0%)**	**2 (0.9%)**
single threshold	F-Test	44 (19%)	20 (9%)	6 (3%)	8 (4%)
	632 Bootstrap	**20 (9%)**	**3 (1%)**	**4 (2%)**	**6 (3%)**

(c)

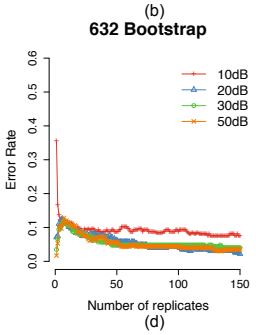


Fig. 2. Evaluation on simulated data for various noise levels. (a) For the F-test, the dependence of the error rate with the threshold $\theta^{\mathrm{F-test}}$ shows that no single threshold can be used to achieve good performance at all noise levels. (b) The error rate is less sensitive to the threshold θ^{B632} set on the generalization error estimate. (c) Number of errors out of the 225 voxels and error rate for optimal thresholds chosen independently for each SNR (top rows) and jointly for all SNR (bottom rows). B632 leads to fewer error than the F-test, in both scenarios. The difference is more striking when a single threshold is used for all SNR. (d) The 632 bootstrap estimate reaches a close to optimal value after about 50 replicates.

3.2 In-Vivo Data: Robustness to Pre-processing

The results on the synthetic phantom suggest that the F-test model selection is not robust to changes in the characteristics of the image. In particular, if a threshold is optimized for some SNR, it will yield suboptimal results at another SNR. The acquisition of DWI is usually followed by several steps of pre-processing before the diffusion model is estimated. Some of these steps aim at improving the SNR. We may wonder whether the model selection is robust to these pre-processing step.

As an illustration, we applied the model selection methods on an in-vivo acquisition before and after automatic motion correction based on coregistration of all the DW images. The acquisition was not corrupted by any significant subject motion, and therefore this step mostly introduces a smoothing due to the interpolation when coregistering the images. Results in Fig. 3 show that the F-test model selection is strongly affected by this preprocessing step. With a threshold

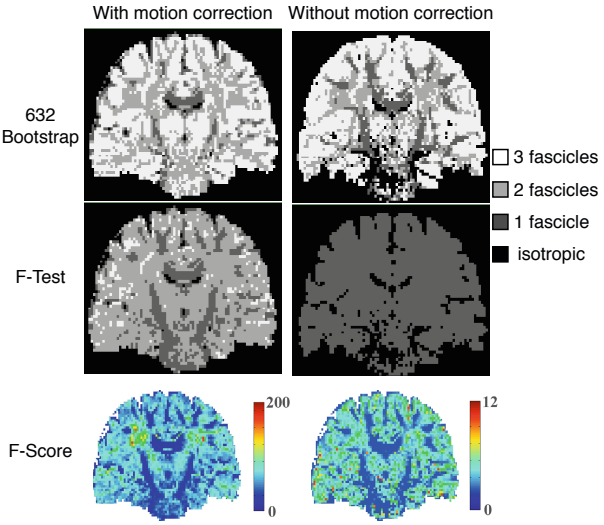


Fig. 3. Maps of the number of fascicles as detected with B632 and F-test for DWI with and without motion correction. Motion correction mostly introduced interpolation since no significant motion was present in this scan. B632 yields similar results in both cases, while the F-test selection fails to detect areas of more than one fascicle. This is due to the reduced perceived signal to noise ratio after motion correction, which affects the result of the F-test, as shown by the map of the F-score for the comparison between one and two-fascicle models.

of $\theta^{\mathrm{F-test}} = 15$, the map of the number of fascicles after motion correction does not resemble that before motion correction. In the latter, only one-tensor models and isotropic diffusion models were selected, while two- and three-tensor models are detected at many locations after motion correction. To observe two- or three-tensor models in the second map, one would need to decrease $\theta^{\mathrm{F-test}}$ since the SNR is lower, which is consistent with the synthetic results of Fig. 2(a). By contrast, for a constant $\theta^{\mathrm{B632}} = 8$, the maps of the number of fascicles detected with B632 are consistent across both images and follow the traits of the anatomy.

3.3 In Vivo Data: Cross-Testing Validation

Experiments based on in vivo data cannot rely on any ground truth to assess the error rates of the methods. To objectively compare the performance of the F-test and the B632 model selection approaches, we performed a cross-testing analysis. This procedure consists in repeatedly splitting the dataset into an *estimation set* and a *testing set*. In our experiments, we considered 70% of the data for estimation and the remaining 30% for testing. Both the model selection and estimation of the MFM parameters were carried out with the estimation set while the testing set was used to assess the performance of the two approaches. The threshold parameters were set to respectively $\theta^{\mathrm{F-test}} = 15$ and

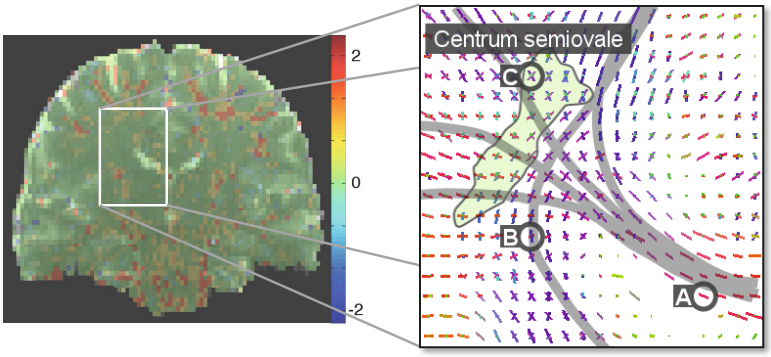


Fig. 4. (a) Comparison of the F-test and B632 model selection approaches using cross-testing. Difference between the testing error when using F-test model selection and B632 model selection. It shows that the testing error is significantly lower (positive values) when using B632. (b) Illustration that B632 and MFM estimation enables reliable detection of the number of fascicles that matches the known anatomy. (a) body of the corpus callosum, (b) crossing of the corpus callosum and the cortico-spinal tracts and (c) centrum semiovale which contains three fascicle orientations.

$\theta^{\mathrm{B632}} = 8$. The performance of the approaches was assessed by computing the mean-square prediction error on the testing set. We repeated the estimation-testing process 30 times and computed the average testing error. Fig 4 shows that the testing error with B632 is lower than with F-test. More precisely, a paired t-test on the differences between the testing errors at each voxel shows that B632 is significantly better than F-test ($p < 10^{-12}$) with an mean improvement of 0.56.

4 Discussion and Conclusion

The estimation of the generalization error allows a reliable selection of the optimal model. Results on both synthetic and in vivo data show the improved performance over model selection based on the F-test.

Validating the models by means of an external dataset, as done in Section 3.3, seems the most objective validation method. Arguably, the generalization error is therefore what all model selection approaches attempt to minimize. However, unlike the fitting error, the generalization error cannot be computed and model selection criteria can be viewed as bypasses to this conceptual limitation. In this interpretation, the fitting error is, itself, an estimate of the generalization error. Complexity penalization could then be seen as heuristic methods to correct for the high bias of this estimate. The 632 bootstrap estimate, on the other hand, directly estimates the generalization error paying attention to remove as much bias and unnecessary variance as possible.

Computing the fitting error is a lot faster than computing the 632 bootstrap which requires several estimations of the model (one for each of the B bootstrap replicates). For this reason, model selection and estimation using B632 is about

B times slower. Different approximations could be used to decrease the computational time, such as estimating sticks instead of tensors during the model selection step and defining a stopping criterion on the number of bootstrap replicates based on the current estimate of the generalization error.

In a future work, we will apply the proposed approach to select between a broader class of generative diffusion models of the diffusion signal. This will enable appropriate identification of the model complexity at each voxel and accurate assessment of the white matter characteristics.

References

1. Alexander, D.C., Barker, G.J., Arridge, S.R.: Detection and modeling of non-gaussian apparent diffusion coefficient profiles in human brain data. Magn. Reson. Med. 48(2), 331–340 (2002)
2. Assaf, Y., Basser, P.J.: Composite hindered and restricted model of diffusion (CHARMED) MR imaging of the human brain. NeuroImage 27(1), 48–58 (2005)
3. Behrens, T.E.J., Berg, H.J., Jbabdi, S., Rushworth, M.F.S., Woolrich, M.W.: Probabilistic diffusion tractography with multiple fibre orientations: What can we gain? NeuroImage 34(1), 144–155 (2007)
4. Efron, B., Tibshirani, R.: Improvements on cross-validation: The .632 + bootstrap method. Journal of the American Statistical Association 92(438), 548–560 (1997)
5. Kreher, B., Schneider, J., Mader, I., Martin, E., Hennig, J., Il'yasov, K.: Multitensor approach for analysis and tracking of complex fiber configurations. Magnetic Resonance in Medicine 54(5), 1216–1225 (2005)
6. Miller, K.L., et al.: Diffusion imaging of whole, post-mortem human brains on a clinical MRI scanner. Neuroimage 57(1), 167–181 (2011)
7. Molinaro, A., Simon, R., Pfeiffer, R.: Prediction error estimation: a comparison of resampling methods. Bioinformatics 21(15), 3301–3307 (2005)
8. Pasternak, O., Westin, C., Bouix, S., Seidman, L., Goldstein, J., Woo, T., Petryshen, T., Mesholam-Gately, R., McCarley, R., Kikinis, R., et al.: Excessive extracellular volume reveals a neurodegenerative pattern in schizophrenia onset. The Journal of Neuroscience 32(48), 17365–17372 (2012)
9. Scherrer, B., Warfield, S.K.: Parametric Representation of Multiple White Matter Fascicles from Cube and Sphere Diffusion MRI. PLoS ONE 7(11) (2012)
10. Schultz, T., Westin, C.-F., Kindlmann, G.: Multi-diffusion-tensor fitting via spherical deconvolution: A unifying framework. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010, Part I. LNCS, vol. 6361, pp. 674–681. Springer, Heidelberg (2010)
11. Shao, J.: Linear Model Selection by Cross-Validation. Journal of the American Statistical Association 88(442), 486–494 (1993)
12. Tuch, D.S., Reese, T.G., Wiegell, M.R., Makris, N., Belliveau, J.W., Wedeen, V.J.: High angular resolution diffusion imaging reveals intravoxel white matter fiber heterogeneity. MRM 48(4), 577–582 (2002)
13. Zhang, H., Schneider, T., Wheeler-Kingshott, C.A., Alexander, D.C.: Noddi: practical in vivo neurite orientation dispersion and density imaging of the human brain. Neuroimage 61(4), 1000–1016 (2012)