Supplemental Document for "Fiber Direction Estimation, Smoothing and Tracking in Diffusion MRI"

Raymond K. W. Wong* Thomas C. M. Lee† Debashis Paul†
Jie Peng†
*Department of Statistics, Iowa State University
*Department of Statistics, University of California, Davis

September 7, 2015

Abstract

This document provides supplementary material to the article "Fiber Direction Estimation, Smoothing and Tracking in Diffusion MRI" written by the same authors.

S1 Practical maximum likelihood estimation for model (2)

In an attempt to find the global maximizer of the likelihood (3), we develop an efficient algorithm through an approximation of model (2). This algorithm essentially performs a grid search, but it makes use of the geometry of the problem so it is fast. It includes three major steps: (i) lay down a grid for $(\alpha_j, \mathbf{m}_j^\intercal)$'s, (ii) evaluate the maximized likelihood function w.r.t. τ_j's on the grid, and (iii) return the grid point that maximizes the likelihood function. One can then use this returned grid point as a starting value in a gradient method for obtaining ML estimation of model (2). Such a strategy results in better numerical stability and accuracy in finding ML estimates.

S1.1 An approximation of model (2)

Let $\mathbf{c}_j = (\alpha_j, \mathbf{m}_j^\intercal)^\intercal$, $\mathbf{c} = (\mathbf{c}_1^\intercal, \ldots, \mathbf{c}_J^\intercal)^\intercal$ and $\mathcal{C}_j$ be the set of grid points for $\mathbf{c}_j$. For simplicity, we take the same set of grid points, $\mathcal{C}$, for all j. To lay down a grid for $\mathbf{m}_j$'s, we apply the sphere tessellation using Icosahedron, which is depicted in Figure S1. The tessellation algorithm starts

with Icosahedron (a regular polyhedron with 20 triangular faces); and then repeatedly divide each triangular face into four smaller triangles and rescale the newly formed vertices. Here, we only pick unique vertices up to a sign for the formation of the grid. In our implementation, we utilize randomly rotated versions of the tessellation with two subdivisions, which results in a grid with 321 directions corresponding to those unique vertices (up to a sign change) in Figure S1 (Right). If $\mathbf{c} \in \prod_{j=1}^{J} \mathcal{C}_j = \mathcal{C}^J$, model (2) can be rewritten as

$$\bar{S}(\mathbf{u}) = \sum_{k=1}^{K} \tilde{\beta}_k x(\mathbf{u}, \tilde{\mathbf{m}}_k, \tilde{\alpha}_k), \tag{S1}$$

where $K = |\mathcal{C}|$, $x(\mathbf{u}, \tilde{\mathbf{m}}_k, \tilde{\alpha}_k) = S_0 \exp\{-b\tilde{\alpha}_k(\mathbf{u}^\intercal \tilde{\mathbf{m}}_k)^2\}$, $(\tilde{\alpha}_k, \tilde{\mathbf{m}}_k) \in \mathcal{C}$ and $\tilde{\beta}_k \in [0, 1)$. One may notice that, in this reformulation, the non-zero $\tilde{\beta}_k$'s are τ_j's in model (2). If $\mathbf{c} \notin \mathcal{C}^J$, i.e. the set of parameters is not a grid point, then equation (S1) serves as an approximation to $\bar{S}(\mathbf{u})$ in model (2) as long as the grid is dense enough in the parameter space.

Furthermore, under the commonly used scales of b-values and tensors, $x(\mathbf{u}, \tilde{\mathbf{m}}_k, \tilde{\alpha}_k)$ and $x(\mathbf{u}, \tilde{\mathbf{m}}_{k'}, \tilde{\alpha}_{k'})$ are highly correlated if $\tilde{\mathbf{m}}_k = \tilde{\mathbf{m}}_{k'}$. Thus, $x(\mathbf{u}, \tilde{\mathbf{m}}_k, \tilde{\alpha}_k)$ is proportional to $x(\mathbf{u}, \tilde{\mathbf{m}}_k, \tilde{\alpha}'_k)$ approximately. Note that the proportional constant can be combined with $\tilde{\beta}_k$ to form a new coefficient in linear model (S1). Inspired by this observation, we reduce the grid size by setting $\tilde{\alpha}_k = \tilde{\alpha}$ for all k to a common value $\tilde{\alpha}$ and using new coefficients β_k's to take care of the proportional constants due to the discrepancy between α_j's and $\tilde{\alpha}$. From our experience, we set $\tilde{\alpha} = 2/b$. With all these approximations, we consider fitting the following model:

$$\bar{S}(u) = \sum_{k=1}^{K} \beta_k x_k(\mathbf{u}), \tag{S2}$$

where $x_k(\mathbf{u}) = x(\mathbf{u}, \tilde{\mathbf{m}}_k, \tilde{\alpha})$ and $\beta_k \geq 0$. For our purpose, we want to identify non-zero β_k's because those $\tilde{\mathbf{m}}_k$'s associated with non-zero $\hat{\beta}_k$'s can be regarded as selected diffusion directions. Note that model (S2) converts the expensive grid search to an estimation problem of a linear model (with respect to β_k's) with non-negative constraints. A fast algorithm for fitting this model with Rician noise assumption is given in Section S1.3 below. As it turns out, the non-negativity constraints often result in a sparse estimate of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)^\intercal$; i.e., only a subset of directions is selected. In particular, if the estimate of the unconstrained problem (i.e., β_k's are allowed to be negative) is not located in the first quadrant of the parameter space, the corresponding constrained solution will be sparse.

Even though the solution is often sparse, the number of selected directions is usually larger than

J, the true number of tensor components. This is partly due to colinearity of $x_k(\mathbf{u})$'s resulting from the use of a dense grid on the directions $\tilde{\mathbf{m}}_k$'s.

In below we propose to first divide the selected directions into I groups and then generate stable estimates of $\mathbf{m}_j$'s via gradient methods (Section S1.2). Finally, Bayesian information criterion (BIC) (Schwarz, 1978) is used to choose an appropriate I as the estimate for J (Section 3.3 of the main article).

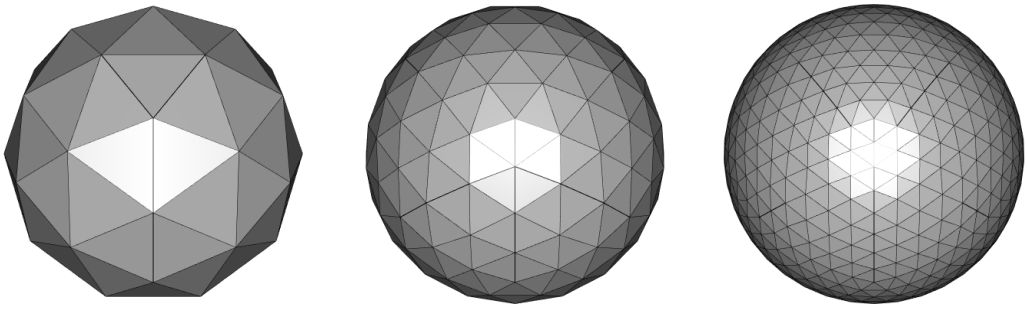

Figure S1: Sphere tessellations through triangulation using Icosahedron with level of subdivisions: 1 (Left), 2 (Middle) and 3 (Right).

S1.2 Clustering of the selected directions

Write the above ML estimate of β_k as $\hat{\beta}_k$ for $k = 1, \ldots, K$. Suppose there are $L > 0$ non-zero $\hat{\beta}_k$'s, without loss of generality, $k = 1, \ldots, L$. Thus, $\tilde{\mathbf{m}}_1, \ldots, \tilde{\mathbf{m}}_L$ are the selected directions. Now, we develop a strategy to cluster the selected directions into I groups, for a set of $I \in \{1, \ldots, L\}$. To perform clustering, we require a metric measure on the space of directions $\mathcal{M}$. A natural metric is

$$d^*(\mathbf{u}, \mathbf{v}) = \arccos(|\mathbf{u}^\intercal \mathbf{v}|), \tag{S3}$$

where $\mathbf{u}, \mathbf{v} \in \mathcal{M}$. Note that, $d^*(\mathbf{u}, \mathbf{v})$ is the acute angle between $\mathbf{u}$ and $\mathbf{v}$. With this distance metric, one can define dissimilarity matrix for a set of directions and make use of a generic clustering algorithm. Our choice is the Partition Around Medoids (PAM) (Kaufman and Rousseeuw, 1990) due to its simplicity. The detailed procedure is described in Algorithm S1 below, where the input vectors are the selected directions. Due to the sparsity of $\hat{\beta}_j$'s and efficient algorithms of PAM, this clustering strategy is practically fast. Let $\check{\mathbf{m}}_1, \ldots, \check{\mathbf{m}}_I$ be the resulting group (Karcher) means. They are used as the starting value for gradient-based methods, such as L-BFGS-B algorithm (Byrd *et al.*, 1995), for obtaining $\hat{\boldsymbol{\gamma}}(I)$, the ML estimate of γ under model (2) with I tensor components.

More specifically, the starting value is set as $((1/I, \tilde{\alpha}, \breve{\mathbf{m}}_1^\intercal), \ldots, (1/I, \tilde{\alpha}, \breve{\mathbf{m}}_I^\intercal))^\intercal$.

S1.3 Estimation of the linear model (S2)

This section describes a fast algorithm that we developed for estimating $\hat{\beta}_j$ in model (S2). With this model one can write the log-likelihood of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)^\intercal$ as

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{m} \left[\log\left(\frac{y_i}{\sigma^2}\right) - \frac{y_i^2 + (\sum_{k=1}^K \beta_k x_{ik})^2}{2\sigma^2} + \log I_0 \left\{ \frac{y_i(\sum_{k=1}^K \beta_k x_{ik})}{\sigma^2} \right\} \right],$$

where $y_i = S(u_i)$ and $x_{ik} = x_k(u_i)$ for $i = 1, \ldots, m, k = 1, \ldots, K$. And now we consider minimizing

$$-\ell(\boldsymbol{\beta}) \qquad \text{subject to} \qquad \beta_k \geq 0 \quad \forall k \tag{S4}$$

with respect to $\boldsymbol{\beta}$. Now, differentiating ℓ with respect to β_j, we have

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{m} \left\{ -\frac{(\sum_{k=1}^K \beta_k x_{ik}) x_{ij}}{\sigma^2} + \frac{y_i x_{ij}}{\sigma^2} t_i(\boldsymbol{\beta}) \right\},$$

where

$$t_i(\boldsymbol{\beta}) = I_1 \left\{ \frac{y_i(\sum_{k=1}^K \beta_k x_{ik})}{\sigma^2} \right\} \Big/ I_0 \left\{ \frac{y_i(\sum_{k=1}^K \beta_k x_{ik})}{\sigma^2} \right\}$$

with

$$I_v(x) = \frac{1}{\pi} \int_0^\pi \exp(x \cos\phi) \cos(v\phi) d\phi$$

as the v-th (for nonnegative integer v) order modified Bessel function of the first kind (Abramowitz and Stegun, 1964). One can show that the solution $\hat{\boldsymbol{\beta}}$ of minimizing (S4) satisfies

$$\hat{\boldsymbol{\beta}}_j = \left[\frac{\sum_{i=1}^m \left\{ t_i(\hat{\boldsymbol{\beta}}) y_i - \sum_{k \neq j} \hat{\beta}_k x_{ik} \right\} x_{ij}}{\sum_{i=1}^m x_{ij}^2} \right]_+ \qquad \forall j. \tag{S5}$$

If we know $t_i(\hat{\boldsymbol{\beta}})$'s, (S5) gives an update formula for one β_k at a time, similarly as in common coordinate descent algorithms. Since coordinate descent algorithm is of an iterative basis, we propose to further approximate $t_i(\hat{\boldsymbol{\beta}})$ by substituting the latest update of $\boldsymbol{\beta}$ into t_i. This leads to the following coordinate descent like strategy for finding $\hat{\boldsymbol{\beta}}$:

- Outer loop: Approximate $r(\hat{\boldsymbol{\beta}})$ using the latest update of $\boldsymbol{\beta}$.
- Inner loop: Coordinate updates through (S5) until convergence.

For inner loop, very often, many coefficients remain zero after thresholding, which leads to unchanged of their values. Since the update of a particular coefficient depends on the partial sum of other coefficients, the inner loop is usually computationally efficient and converges in a fast manner.

This algorithm requires an initial value of $\boldsymbol{\beta}$. Motivated by the typical non-linear estimator of a single fiber model, we can choose the initial value as a constrained least square estimator which minimizes

$$\sum_{i=1}^{m}\left(y_i - \sum_{k=1}^{K}\beta_k x_{ik}\right)^2 \qquad \text{subject to} \qquad \beta_k \geq 0 \qquad \forall k.$$

Note that this is a quadratic programming problem, which can be solved efficiently by existing algorithms.

S2 Simulation study of voxel-wise estimation

This section provides simulation results for the voxel-wise estimation procedure proposed in Section 3. Observed signal intensities were simulated from model (1) with Rician noise under three settings:

1. Single tensor case: $J = 1$, $\mathbf{m}_1 = (1, 0, 0)^\intercal$.
2. Two tensor case with perpendicular crossing and unbalanced components: $J = 2$, $\mathbf{m}_1 = (1, 0, 0)^\intercal$, $\mathbf{m}_2 = (0, 1, 0)^\intercal$, $p_1 = 0.7$, $p_2 = 0.3$.
3. Two tensor case with 50 degree crossing and balanced components: $J = 2$, $\mathbf{m}_1 = (\cos(\pi/9), \sin(\pi/9), 0)^\intercal$, $\mathbf{m}_2 = (\sin(\pi/9), \cos(\pi/9), 0)^\intercal$, $p_1 = 0.5$, $p_2 = 0.5$.

All FAs and largest eigenvalues of underlying tensors are set to 0.9 and 4×10^{-3} respectively. Moreover, b, S_0 and σ are set to 1000, 1000 and 50 respectively. This has a signal-to-noise ratio ($\mathsf{SNR} := S_0/\sigma$) 20, which is typical for dMRI studies. $\mathcal{U}$ is obtained from the sphere tessellation with 3 subdivision using octahedron and $|\mathcal{U}| = 33$. For each setting, we simulate 200 voxel-wise data sets and compare the following methods:

- golden: Optimization of (3) via Broyden-Fletcher-Goldfarb-Shanno (BFGS) method with starting values set as the true parameter values. (J is known.)
- global-aic: Global optimization of (3) via GENOUD (Sekhon and Mebane, 1998) with Akaike Information criterion (AIC) for selection of J.

- global-bic: Similar to global-aic but with BIC.
- prop-aic: Our proposed method with AIC.
- prop-bic: Our proposed method with BIC.

Note that the AIC is derived as

$$\mathsf{AIC}(I) = -2l(\hat{\boldsymbol{\gamma}}(I)) + 8I.$$

The simulation results are summarized in Table S1. With the information of true parameters, golden can be treated as a golden standard. Excluding golden, prop-bic has the highest proportion of correct estimation of J and attains around 99% correct recovery, which leads to our choice of BIC over AIC. In addition, note that prop-bic over-selects J when it does not estimate J correctly. This is one of the reasons why a removal step (Step 12 of Algorithm S4) is designed in our smoothing procedure. As said, our goal is the diffusion direction $\mathbf{m}$. Conditional on the correct estimation of J, the squared error of $\mathbf{m}$ is defined as

$$\min_{\{k_1,\ldots,k_J\in\{1,\ldots,J\}:k_i\neq k_j\}} \sum_{j=1}^{J} d^{*2}(\mathbf{m}_j, \hat{\mathbf{u}}_{k_j}),$$

where $\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_J$ are the estimated diffusion directions. From Table S1 all methods have root MSEs of $\mathbf{m}$ ranging from 1.5 to 1.6, 4.5 to 4.6 and 5.1 to 5.7 degree in the three settings respectively, and so these methods do not have big difference in terms of tracking. Given the accurate estimation of J and the computational benefit (over general global optimization methods), prop-bic performs the best among the compared methods.

S3 Choice of bandwidth

This section presents our bandwidth selection methods for the smoothing method in Section 4. These methods are based on the idea of cross-validation (CV). Let $\breve{\mathbf{m}}_i^{-i}$ be the smoothed version of $\hat{\mathbf{m}}_i$ when all directions sharing the same voxel with $\hat{\mathbf{m}}_i$ are not used in the smoothing. Since the choice of h may affect the number of clusters (steps 3 and 4 of Algorithm S4), $\hat{\mathbf{m}}_i$ may have been removed (step 12 of Algorithm S4). Thus, $\breve{\mathbf{m}}_i^{-i}$ is not always defined. Let o_i be the indicator of the existence of $\breve{\mathbf{m}}_i^{-i}$. The CV score is the mean of $\{d^{*2}(\hat{\mathbf{m}}_i, \breve{\mathbf{m}}_i^{-i}) : o_i = 1\}$.

Even after direction smoothing, the number of diffusion directions within a voxel may still be over-estimated. These spurious directions can have a great effect on the CV score, similar to the effect of outliers.

Table S1: Simulation results for voxel-wise estimation. **Correct-select**: proportion of $\hat{J} = J$. **Over-select**: proportion of $\hat{J} >= J$. $\mathbf{m}$, α, τ : MSE of $\mathbf{m}$, α and τ (computed on $\hat{J} = J$), with corresponding standard error stated in brackets. Note that the MSE of $\mathbf{m}$ is in squared degree.

Setting	Method	Correct-select	Over-select	$\mathbf{m}$	α	τ
1	golden	100%	100%	2.48 (3.06e-03)	5.70e-02 (4.66e-03)	2.69e-04 (2.32e-05)
	global-aic	75%	100%	2.39 (3.32e-03)	5.50e-02 (5.40e-03)	2.65e-04 (2.40e-05)
	global-bic	98%	100%	2.48 (3.11e-03)	5.65e-02 (4.70e-03)	2.67e-04 (2.33e-05)
	prop-aic	89%	100%	2.41 (3.07e-03)	5.53e-02 (5.01e-03)	2.73e-04 (2.51e-05)
	prop-bic	99.5%	100%	2.48 (3.07e-03)	5.65e-02 (4.65e-03)	2.69e-04 (2.33e-05)
2	golden	100%	100%	20.5 (1.99e-02)	1.23 (2.83e-01)	4.01e-04 (2.81e-05)
	global-aic	81.5%	100%	21.0 (2.19e-02)	1.20 (3.40e-01)	3.93e-04 (2.97e-05)
	global-bic	97%	100%	21.3 (2.07e-02)	1.92 (7.46e-01)	4.05e-04 (2.88e-05)
	prop-aic	91.5%	100%	21.1 (2.13e-02)	1.37 (3.43e-01)	4.10e-04 (2.93e-05)
	prop-bic	99.5%	100%	20.7 (2.01e-02)	1.33 (3.16e-01)	4.00e-04 (2.81e-05)
3	golden	100%	100%	28.7 (3.85e-02)	5.21 (3.24)	2.72e-03 (2.61e-04)
	global-aic	74.5%	100%	26.5 (3.80e-02)	2.02 (4.21e-01)	2.51e-03 (2.92e-04)
	global-bic	95.5%	100%	32.3 (7.20e-02)	3.38 (1.04)	2.95e-03 (3.47e-04)
	prop-aic	93.5%	100%	27.6 (3.83e-02)	5.37 (3.46)	2.60e-03 (2.65e-04)
	prop-bic	99%	100%	28.6 (3.86e-02)	5.23 (3.27)	2.70e-03 (2.62e-04)

To alleviate this issue, the trimmed mean of $\{d^{*2}(\hat{\mathbf{m}}_i, \breve{\mathbf{m}}_i^{-i}) : o_i = 1\}$ and the median of $\{d^{*}(\hat{\mathbf{m}}_i, \breve{\mathbf{m}}_i^{-i}) : o_i = 1\}$ are used to form robust CV scores. They are called trimmed CV score and Median CV score respectively. We choose h as the minimizer of either one of these scores. See Section S6 for their numerical comparison.

In our numerical illustrations, the bandwidth h is chosen differently for single fiber regions and crossing fiber regions. Further, if one has enough computational resource, adaptive choice of bandwidth can also be achieved by dividing voxels into blocks according to their spatial locations and performing cross validation.

S4 Algorithms

This section presents various algorithms developed in the main paper.

Algorithm S1 ClustDir: PAM based clustering for direction vectors

Input: Set of direction vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$, number of cluster N_c

Output: Group mean $\{\mathbf{v}_1^*, \ldots, \mathbf{v}_{N_c}^*\}$, group label $\{e_1, \ldots, e_n\}$

procedure ClustDir($\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$, N_c)

  for $i, j = 1$ to n **do** $D_{ij} \leftarrow d^*(\mathbf{v}_i, \mathbf{v}_j)$

  Define $\mathbf{D}$ as the dissimilarity matrix with elements D_{ij}'s

  Apply PAM with dissimilarity matrix $\mathbf{D}$ to cluster $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ into N_c groups

  for $i = 1$ to n **do** $e_i \leftarrow$ group label of $\mathbf{v}_i$

  for $j = 1$ to N_c **do** Compute group (Karcher) means:

$$v_j^* \leftarrow \arg\min_{\mathbf{v} \in \mathcal{M}} \sum_{i=1}^{n} I\{e_i = j\} d^{*2}(\mathbf{v}_i, \mathbf{v})$$

  return $(\{\mathbf{v}_1^*, \ldots, \mathbf{v}_{N_c}^*\}, \{e_1, \ldots, e_n\})$

Algorithm S2 Algorithm for voxel-wise estimation

Input: Observed signal intensities $\{S(\mathbf{u}), \mathbf{u} \in \mathcal{U}\}$, set of gradient vectors $\mathcal{U}$, non-diffusion weighted intensity S_0, standard deviation of the noise σ, b-value b, FA threshold r, upper bound of the number of directions $\tilde{I}$

Output: The selected number of diffusion directions, $\hat{J}$ and, if $\hat{J} > 0$, the corresponding ML estimate $\boldsymbol{\gamma}(\hat{J})$

Description: To perform voxel-wise estimation

Compute FA

if FA $< r$ **then**

  Declare there is no major diffusion direction: $\hat{J} \leftarrow 0$

else

  Estimate $\boldsymbol{\beta}$ (Appendix S1.3) and determine the selected directions.

  for $I = 1, \ldots, \min\{\tilde{I}, L\}$ **do**

    Cluster the selected directions into I groups (Algorithm S1)

    Perform optimization with a gradient method (Section 3.3) and obtain ML estimate $\hat{\boldsymbol{\gamma}}(I)$

    Compute $\mathsf{BIC}(I)$

  Compute $\mathsf{BIC}(0)$

  Estimate the number of diffusion directions: $\hat{J} \leftarrow \text{argmin}_{I \in \{0, \ldots, \min\{\tilde{I}, L\}\}} \mathsf{BIC}(I)$

Algorithm S3 ClustDirN: PAM based clustering algorithm for direction vectors with automatic choice of number of clusters

Input: Set of direction vector $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$, maximum number of cluster K, angular threshold ξ

Output: Group mean $\{\mathbf{v}_1^*, \ldots, \mathbf{v}_C^*\}$, number of clusters C

procedure ClustDirN($\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$, K, ξ)
  if $n = 1$ **then**
    the case of only one input direction: declare only one cluster
    $C \leftarrow 1$
  else if $n = 2$ **then**
    the case of two input directions: declare only one cluster if the angular separation of these directions are small
    if $d^*(\mathbf{v}_1, \mathbf{v}_2) \leq \xi$ **then** $C \leftarrow 1$ **else** $C \leftarrow 2$
  else if $n = 3$ **then**
    the case of three input directions
    $\psi \leftarrow$ the distance (S3) between the two cluster means of ClustDir($\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$, 2) (Algorithm S1)
    if $\psi \leq \xi$ **then**
      $C \leftarrow 1$
    else
      if minimum pairwise distance of $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\} \leq \xi$ **then** $C \leftarrow 2$ **else** $C \leftarrow 3$
  else
    the case of more than three input directions: use Shilhouette criterion
    $\psi \leftarrow$ distance between the two cluster means of ClustDir($\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$, 2)
    if $\psi \leq \xi$ **then**
      Claim there is only one cluster if the angular separation is small: $C \leftarrow 1$
    else
      for $k = 2$ to K **do**
        $a_k \leftarrow$ average silhouette computed using ClustDir($\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$, k)
      Estimate the number of clusters as the maximizer of average silhouette: $C \leftarrow \arg\min_j \{a_j\}$
  $(\{\mathbf{v}_1^*, \ldots, \mathbf{v}_C^*\}, \{e_1, \ldots, e_n\}) \leftarrow$ ClustDir($\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$, C)
  return $(\{\mathbf{v}_1^*, \ldots, \mathbf{v}_C^*\}, C)$

Algorithm S4 Algorithm for direction smoothing

Input: Target voxel $\mathbf{s}^*$, voxel-wise estimate $\{(\mathbf{s}_k, \hat{\mathbf{m}}_k), k = 1, \ldots, T\}$, estimated number of fibers $\{\hat{J}(\mathbf{s}) : \mathbf{s} \in \mathcal{S}\}$, kernel function K, bandwidth matrix $\mathbf{H}$, threshold c, maximum number of cluster (Algorithm S3) K, angular threshold (Algorithm S3) ξ

Output: Updated number of directions and updated directions at $\mathbf{s}^*$

Description: To perform smoothing for diffusion directions at $\mathbf{s}^*$

1: **for** $k = 1$ to T **do** Compute spatial weight: $w_k \leftarrow K_{\mathbf{H}}(\mathbf{s}_k - \mathbf{s}^*)$

2: **for** $k = 1$ to T **do** Standardize spatial weights: $w_k \leftarrow w_k / \sum_{j=1}^{T} w_j$

3: Sort w_k's in decreasing order such that $w_{l_1} \geq \cdots \geq w_{l_T}$

4: Identify neighborhood for clustering (Section 4.2):
Compute $L \leftarrow \min_{M \in \{1,\ldots,T\}} \mathbb{1}\{\sum_{m=M+1}^{T} w_{l_m} \leq c\}$ (The summation $\sum_{m=T+1}^{T} w_{l_m}$ is defined as 0.)

5: Clustering via Algorithm S3: $(\{\mathbf{u}_1, \ldots, \mathbf{u}_C\}, C) \leftarrow$CLUSTDIRN$(\{\hat{\mathbf{m}}_{l_1}, \ldots, \hat{\mathbf{m}}_{l_L}\}, K, \xi)$

6: **if** $C \geq \hat{J}(\mathbf{s}^*)$ **then**

7: Match the smoothed directions, $\{u_1, \ldots, u_C\}$, to the voxel-wise estimates at $\mathbf{s}^*$, $\{\hat{\mathbf{m}}_1(s^*), \ldots, \hat{\mathbf{m}}_{\hat{J}(s^*)}(s^*)\}$:

$$\left(\hat{k}_1, \ldots, \hat{k}_{\hat{J}(\mathbf{s}^*)}\right) \leftarrow \underset{\{k_1 \ldots, k_{\hat{J}(\mathbf{s}^*)} \in \{1,\ldots,C\} : k_i \neq k_j\}}{\arg\min} \sum_{j=1}^{\hat{J}(\mathbf{s}^*)} d^*(\hat{\mathbf{m}}_j(\mathbf{s}^*), \mathbf{u}_{k_j})$$

8: **for** $j = 1$ to $\hat{J}(\mathbf{s}^*)$ **do** $\hat{\mathbf{m}}_j(\mathbf{s}^*) \leftarrow \mathbf{u}_{\hat{k}_j}$

9: **else**

10: Match the voxelwise estimates at $\mathbf{s}^*$ to the smoothed directions: :

$$\left(\hat{k}_1, \ldots, \hat{k}_C\right) \leftarrow \underset{\{k_1 \ldots, k_C \in \{1,\ldots,\hat{J}(\mathbf{s}^*)\} : k_i \neq k_j\}}{\arg\min} \sum_{j=1}^{C} d^*(\hat{\mathbf{m}}_{k_j}(\mathbf{s}^*), \mathbf{u}_j)$$

11: **for** $j = 1$ to C **do** $\hat{\mathbf{m}}_{\hat{k}_j}(\mathbf{s}^*) \leftarrow \mathbf{u}_j$

12: $\hat{J}(\mathbf{s}^*) \leftarrow C$ and remove non-updated $\hat{\mathbf{m}}_j(\mathbf{s}^*)$'s

Algorithm S5 Algorithm for fiber tracking

Input: Target voxel $\mathbf{s}^*$, initial direction $\mathbf{v}^*$, (smoothed) voxel-wise estimate $\{(\mathbf{s}_k, \hat{\mathbf{v}}_k), k = 1, \ldots, T\}$, maximum number of projection N_{proj}, angular threshold ξ

Output: Recorded locations and directions

Description: To perform fiber tracking

1: Initialization: $\mathbf{x} \leftarrow \mathbf{s}^*$; $\mathbf{v} \leftarrow \mathbf{v}^*$; $Z \leftarrow$ True
Here, $\mathbf{x}$ represents the current location, $\mathbf{v}$ represents the current direction, Z is an indicator of whether the tracking should continue
2: Record $\mathbf{x}, \mathbf{v}$
3: **while** Z **do**
4: Move from $\mathbf{x}$ in the direction of $\mathbf{v}$ until hitting the boundary of the voxel
5: $\mathbf{x} \leftarrow$ boundary point of the voxel
6: $K \leftarrow$ number of fiber directions at the next voxel
7: **if** $K = 0$ **then**
8: $\tilde{Z} \leftarrow$ False, where $\tilde{Z}$ is an indicator of whether a viable direction exists
9: **else**
10: $\{\mathbf{v}_1, \ldots, \mathbf{v}_K\} \leftarrow$ fiber directions at the next voxel
11: Identify the direction with smallest angular separation: $j \leftarrow \arg\min_k d^*(\mathbf{v}, \mathbf{v}_k)$
12: **if** $d^*(\mathbf{v}, \mathbf{v}_j) \leq \xi$ **then**
13: $\mathbf{v} \leftarrow \text{sign}(\mathbf{v} \cdot \mathbf{v}_j)\mathbf{v}_j$; $\tilde{Z} \leftarrow$ True
14: **else**
15: $\tilde{Z} \leftarrow$ False
16: **if** not $\tilde{Z}$ **then**
Project the tracking and check if there is any viable direction after N_{proj} voxels:
17: $\tilde{\mathbf{x}} \leftarrow \mathbf{x}$; $\tilde{\mathbf{v}} \leftarrow \mathbf{v}$
18: **for** $n = 1$ to N_{proj} **do**
19: Projection: run lines 4 to 15 with all $\mathbf{x}$ and $\mathbf{v}$ replaced by $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{v}}$
20: **if** $\tilde{Z}$ **then**
21: Record $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{v}}$; **break**
22: **if** not $\tilde{Z}$ **then** $Z \leftarrow$ False (Stop the tracking if there is no viable direction after N_{proj} voxels)
23: **else**
24: Record $\mathbf{x}, \mathbf{v}$

S5 Proofs and Technical details

S5.1 Proof of Theorem 1

To prove Theorem 1, it suffices to prove the following proposition.

Proposition 1. *Let* $f(\mathbf{u}) = \sum_{j=1}^{p} c_j e^{-a_j(\mathbf{m}_j^\intercal \mathbf{u})^2}, \quad \mathbf{u} \in \mathcal{V}$, *be a spherical function, where* $c_j \in \mathbb{R}$, $a_j > 0$, $\mathbf{m}_j \in \mathcal{V}$ *for* $j = 1, \cdots, p$ *and* $\mathbf{m}_j \neq \pm\mathbf{m}_i$ *for* $1 \leq i \neq j \leq p$. *Then if* $f(\mathbf{u}) = 0$ *for all* $\mathbf{u} \in \mathcal{V}$, *we have* $c_j = 0$ *for* $j = 1, \ldots, p$.

This proposition says that, $f_j(\mathbf{u}) := e^{-a_j(\mathbf{m}_j^\intercal \mathbf{u})^2}$, $j = 1, \ldots, p$, are linearly independent spherical functions.

Proof. We first consider the spherical Laplacian operator $\Delta_{\mathcal{V}}$. It can be shown that, if $f = f(x_1, x_2, x_3)$, $x_1^2 + x_2^2 + x_3^2 = 1$, is a spherical function only depending on x_3, then

$$\Delta_{\mathcal{V}} f = \frac{\partial\left((1 - x_3^2)\frac{\partial f}{\partial x_3}\right)}{\partial x_3}.$$

Particularly, applying $\Delta_{\mathcal{V}}$ on the function e^{-ax^2}, $|x| \leq 1$, we get

$$\Delta_{\mathcal{V}}\left(e^{-ax^2}\right) = P_1(x; a)e^{-ax^2},$$

where $P_1(x; a)$ is a 4th order polynomial in x with the leading term being $-(2a)^2x^4$. Successively applying $\Delta_{\mathcal{V}}$ l times, by induction, we get

$$\Delta_{\mathcal{V}}^l\left(e^{-ax^2}\right) = P_l(x; a)e^{-ax^2}, \quad l \geq 0,$$

where $P_l(x; a)$ is a $4l$th polynomial in x with the leading term being $(-1)^l(2a)^{2l}x^{4l}$.

Now applying $\Delta_{\mathcal{V}}^l$ to $f_j(\mathbf{u}) := e^{-a_j(\mathbf{m}_j^\intercal \mathbf{u})^2}$, $\mathbf{u} \in \mathcal{V}$. Since the spherical Laplacian operator $\Delta_{\mathcal{V}}$ is invariant to orthogonal transformations, so

$$\begin{aligned} \Delta_{\mathcal{V}}^l f_j &= \Delta_{\mathcal{V}}^l\left(e^{-a_j x^2}\right), \quad x = \mathbf{m}_j^\intercal \mathbf{u} \\ &= P_l(x; a_j)e^{-a_j x^2} = P_l(\mathbf{m}_j^\intercal \mathbf{u}; a_j)f_j, \quad l \geq 0. \end{aligned}$$

If $f = \sum_{j=1}^{p} c_j f_j \equiv 0$ on $\mathcal{V}$, then for $l \geq 0$

$$\begin{aligned} \Delta_{\mathcal{V}}^l f &= \sum_{j=1}^{p} c_j \Delta_{\mathcal{V}}^l f_j \\ &= \sum_{j=1}^{p} c_j P_l(\mathbf{m}_j^\intercal \mathbf{u}; a_j)e^{-a_j(\mathbf{m}_j^\intercal \mathbf{u})^2} \equiv 0, \quad \mathbf{u} \in \mathcal{V}. \end{aligned}$$

Consider the p by p matrix function:

$$\tilde{\mathbf{P}}(\mathbf{u}) = \left(P_l(\mathbf{m}_j^\intercal \mathbf{u}; a_j)e^{-a_j(\mathbf{m}_j^\intercal \mathbf{u})^2}\right)_{l=0,\ldots,p-1}^{j=1,\ldots,p}, \quad \mathbf{u} \in \mathcal{V}.$$

In order to show $c_j = 0$ for $j = 1, \ldots, p$, we only need to show that there exists $\mathbf{u}^* \in \mathcal{V}$ such that $det(\tilde{\mathbf{P}}(\mathbf{u}^*)) \neq 0$. Note that, $e^{-a_j(\mathbf{m}_j^\intercal \mathbf{u})^2}$ does not depend on l and is everywhere nonzero, so $det(\tilde{\mathbf{P}}(\mathbf{u}^*)) \neq 0$ if and only if $det(\mathbf{P}(\mathbf{u}^*)) \neq 0$ where

$$\mathbf{P}(\mathbf{u}) = \left(P_l(\mathbf{m}_j^\intercal \mathbf{u}; a_j)\right)_{l=0,\ldots,p-1}^{j=1,\ldots,p}, \quad \mathbf{u} \in \mathcal{V}. \tag{S6}$$

This is shown in the following lemma and it completes the proof of the proposition. □

Lemma 1. *There exists $\mathbf{u}^* \in \mathcal{V}$ such that $det(\mathbf{P}(\mathbf{u}^*)) \neq 0$ for $\mathbf{P}$ defined in equation (S6) with $a_j > 0$, $\mathbf{m}_j \in \mathcal{V}$ for $j = 1, \ldots, p$ and $\mathbf{m}_j \neq \pm \mathbf{m}_i$ for $1 \leq i \neq j \leq p$.*

Proof. For any $\mathbf{u}, \mathbf{v} \in \mathcal{V}$ such that $\mathbf{u}^\intercal \mathbf{v} = 0$, consider the curve on $\mathcal{V}$:

$$\mathbf{u}(t) = \mathbf{u}\cos t + \mathbf{v}\sin t = \frac{1}{2}(\mathbf{u} - i\mathbf{v})w + \frac{1}{2}(\mathbf{u} + i\mathbf{v})\bar{w}, \quad w = e^{it} = \cos t + i \sin t.$$

As $\bar{w} = 1/w$, $P_l(\mathbf{m}_j^\intercal \mathbf{u}(t); a_j)$ is a Laurent polynomial in w with leading terms being $\lambda_j(\mathbf{u}, \mathbf{v})^l w^{4l}$ and $\lambda'_j(\mathbf{u}, \mathbf{v})^l w^{-4l}$, where

$$\begin{aligned} \lambda_j(\mathbf{u}, \mathbf{v}) &= -4a_j^2 \left(\frac{1}{2}\mathbf{m}_j^\intercal(\mathbf{u} - i\mathbf{v})\right)^4, \quad j = 1, \ldots, p, \\ \lambda'_j(\mathbf{u}, \mathbf{v}) &= -4a_j^2 \left(\frac{1}{2}\mathbf{m}_j^\intercal(\mathbf{u} + i\mathbf{v})\right)^4, \quad j = 1, \ldots, p. \end{aligned} \tag{S7}$$

Therefore, $w^{2p(p-1)} det(P_l(\mathbf{m}_j^\intercal \mathbf{u}(t); a_j))$ is a $4p(p-1)$ order polynomial in w with the leading term being $w^{2p(p-1)} det(\lambda_j(\mathbf{u}, \mathbf{v})^l w^{4l})$ which is nonzero if and only if $det(\lambda_j(\mathbf{u}, \mathbf{v})^l)$ is nonzero (since $w \neq 0$).

Note

$$det(\lambda_j(\mathbf{u}, \mathbf{v})^l) = \prod_{k<j} (\lambda_j(\mathbf{u}, \mathbf{v}) - \lambda_k(\mathbf{u}, \mathbf{v})).$$

By Lemma 2 (the following lemma), there exists $\mathbf{u}^0, \mathbf{v}^0 \in \mathcal{V}, (\mathbf{u}^0)^\intercal \mathbf{v}^0 = 0$ such that: $\lambda_j(\mathbf{u}^0, \mathbf{v}^0)$ are all distinct for $j = 1, \ldots, p$. Therefore, $w^{2p(1-p)} det(P_l(\mathbf{m}_j^\intercal \mathbf{u}^0(t); a_j))$ has a nonzero coefficient for the $4p(p-1)$ (highest order) term and thus is not constantly zero. Therefore, there exists t^* such that $det(P_l(\mathbf{m}_j^\intercal \mathbf{u}^0(t^*); a_j)) \neq 0$.

Now let $u^* = \mathbf{u}^0 \cos t^* + \mathbf{v}^0 \sin t^*$, then we prove the lemma. □

Lemma 2. *If $a_j > 0$ and $\mathbf{m}_j \neq \pm \mathbf{m}_i$ for $1 \leq i \neq j \leq p$, then there exists $\mathbf{u}^0, \mathbf{v}^0 \in \mathcal{V}, (\mathbf{u}^0)^\intercal \mathbf{v}^0 = 0$ such that: $\lambda_j(\mathbf{u}^0, \mathbf{v}^0)$ defined by equation (S7) are all distinct for $j = 1, \ldots, p$.*

Proof. If the result does not hold, then there exists $1 \leq j \neq k \leq p$ such that

$$a_j \left(\mathbf{m}_j^\intercal(\mathbf{u} - i\mathbf{v})\right)^2 = a_k \left(\mathbf{m}_k^\intercal(\mathbf{u} - i\mathbf{v})\right)^2$$

holds for at least three pairs of $\mathbf{u}, \mathbf{v} \in \mathcal{V}, \mathbf{u}^\intercal\mathbf{v} = 0$ where the three $\mathbf{u}$ are linearly independent. Therefore

$$\sqrt{a}_j \mathbf{m}_j^\intercal \mathbf{u} = \pm\sqrt{a}_k \mathbf{m}_k^\intercal \mathbf{u}$$

for at least three $\mathbf{u}$ on $\mathcal{V}$ which are linearly independent. Therefore

$$\sqrt{a}_j \mathbf{m}_j = \pm\sqrt{a}_k \mathbf{m}_k.$$

Since $\mathbf{m}_j, \mathbf{m}_k \in \mathcal{V}$ and $a_j, a_k > 0$, this means

$$a_j = a_k, \quad \mathbf{m}_j = \pm\mathbf{m}_k,$$

which is a contradiction. □

S5.2 Proof and technical details of Theorem 2

We need the following assumptions for Theorem 2. Write $\mathcal{B}_\delta(\boldsymbol{\theta}_0) = \{\boldsymbol{\theta} \in \mathbb{R}^2 : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| < \delta\}$, for $\delta > 0$. Throughout our discussion, we use the L_2-norm for matrix.

Assumption 1. *There exists* $\epsilon > 0$ *such that* $\text{supp}(\mathbf{v}_1|S_1 = s) \subseteq \{\mathbf{v} \in \mathcal{M} : d^*(\mathbf{v}, \mathbf{v}_0) \leq \pi/2 - \epsilon\}$, *in a neighborhood of* s_0.

Assumption 2. $h \to 0$ *and* $nh \to \infty$.

Assumption 3. $K(\cdot)$ *is bounded, compactly supported kernel function satisfying (i)* $\int K(x)dx = 1$ *and (ii)* $\int xK(x)dx = 0$.

Assumption 4. *The density of* S, $f_S(\cdot)$, *is twice continuously differentiable in a neighborhood of* s_0 *and* $f_S(s_0) > 0$.

Assumption 5. $m_j(\cdot)$ *is twice continuously differentiable in a neighborhood of* s_0, *for* $j = 1, 2$.

Assumption 6. $\Sigma_{jk}(\cdot)$ *is continuous in a neighborhood of* s_0, *for* $j, k = 1, 2$.

Assumption 7. $\Psi_{jk}(\cdot)$ *is continuous in a neighborhood of* s_0, *for* $j, k = 1, 2$.

Assumption 8. $E\{[\boldsymbol{\psi}_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_0)]_{j,k}^2 | S_1 = s\} \leq C_{jk}$ *for all* s, *for* $j, k = 1, 2$.

Assumption 9. *Let* $\gamma(\delta, s) = \mathbb{E}[\sup_{\tilde{\boldsymbol{\theta}} \in \mathcal{B}_\delta(\boldsymbol{\theta}_0)} \|\boldsymbol{\psi}_2(\boldsymbol{\theta}_1, \tilde{\boldsymbol{\theta}}) - \boldsymbol{\psi}_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_0)\| | S_1 = s]$. *There exists a neighborhood of* s_0, $\mathcal{W}(s_0)$, *such that*

$$\tilde{\gamma}(\delta) = \sup_{s \in \mathcal{W}(s_0)} \gamma(\delta, s) = o(1) \quad as \quad \delta \to 0.$$

Assumption 10. $\boldsymbol{\Psi}(s_0)$ *is positive definite.*

Assumption 1 is a technical assumption for avoiding the unnecessary complication arising from the representation of geodesic distance as a function of the working coordinate system. As a result of Assumption 1, one can use a representation of $\pm\mathbf{v}$, which aligns with $\mathbf{v}_0$, and reduces the geodesic distance of $\mathcal{M}$ to the geodesic distance of $\mathcal{V}$. This assumption is usually satisfied by our procedure, as a results of thresholding and clustering. Assumptions 2–8 are standard conditions for consistency and distributional limits for smoothing estimators. Assumption 9 states that the local modulus of continuity of the Hessian of the discrepancy measure ($d^2(\boldsymbol{\omega}, \boldsymbol{\theta})$) between the principal diffusion directions converges to zero, uniformly over the locations. Thus, this condition imposes a minimal degree of smoothness of the Hessian of the discrepancy measure, which is needed to ensure negligibility of higher order terms in a second order Taylor expansion of the objective function around the true diffusion direction at any given location, and is used in proving Lemma 7. Assumption 10 is also a standard requirement to ensure identifiability of the true diffusion direction at s_0.

Lemma 3. *Assume that Assumption 1 hold.* $\psi(\boldsymbol{\omega}, \boldsymbol{\theta})$ *is twice continuously differentiable in a neighborhood of* $\boldsymbol{\theta}_0 = \mathbf{0}$, $m(s_0) = \mathbf{0}$ *and* $M_n^{(1)}(\mathbf{0}) = -2\sum_{i=1}^n hK_h(S_i - s_0)\boldsymbol{\theta}_i$.

Proof of Lemma 3. Under Assumption 1, for $\boldsymbol{\theta}$ close to $\boldsymbol{\theta}_0$,

$$d(\boldsymbol{\theta}_i, \boldsymbol{\theta}) = \arccos(|\rho_{\mathbf{v}_0}(\mathbf{v}_i)^\intercal \phi^{-1}(\boldsymbol{\theta})|) = \arccos(\rho_{\mathbf{v}_0}(\mathbf{v}_i)^\intercal \phi^{-1}(\boldsymbol{\theta})).$$

Note that $\rho_{\mathbf{v}_0}(\mathbf{v}_i) \in \mathcal{V}$ is represented by $\boldsymbol{\theta}_i$. Thus, for $\boldsymbol{\theta}$ close to $\boldsymbol{\theta}_0$, $d(\boldsymbol{\theta}_i, \boldsymbol{\theta})$ coincides with the geodesic distance of $\mathcal{V}$ between points represented by logarithm coordinates $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}$. Now, Lemma 3 follows from Bhattacharya and Bhattacharya (2012, Theorem 5.3) applied to the Manifold $\mathcal{V}$. Note that the cited theorem develops the coordinate system through the logarithm map at the intrinsic mean, which is not the same in our case. However, the requirement for developing the system at the intrinsic mean is for deeper results stated in their theorem, which is irrelevant to our use of their theorem. □

Lemma 4. *Assume that Assumptions 1-5 hold. Let* $\mathbf{Y}_i = hK_h(S_i - s_0)m(S_i)$*, for* $i = 1, \dots n$*. Then*

$$\sum_{i=1}^{n} \mathbf{Y}_i = nh^3 \int x^2 K(x)dx \left\{ m^{(1)}(s_0)f^{(1)}(s_0) + \frac{1}{2}m^{(2)}(s_0)f_S(s_0) \right\} + O_p(\sqrt{nh^3}),$$

where $m^{(1)}$ *and* $m^{(2)}$ *are interpreted as vectors of first and second derivatives of elements of* m *respectively.*

Proof of Lemma 4. Since $\mathbf{Y}_i$'s are independently and identically distributed, we have

$$\sum_{i=1}^{n} \mathbf{Y}_i = n\mathbb{E}(\mathbf{Y}_1) + O_p\left\{ \sqrt{n\mathbb{E}(\mathbf{Y}_1^2)} \right\}.$$

We compute $\mathbb{E}(\mathbf{Y}_1)$ and $\mathbb{E}(\mathbf{Y}_1^2)$ below. Write $\mathbf{Y}_1 = (Y_{1,1}, Y_{1,2})^\intercal$. For $j = 1, 2$, by dominated convergence theorem with boundedness and continuity assumptions of f_S and m_j, and $m(s_0) = 0$, from Lemma 3, we have

$$\begin{aligned}
\mathbb{E}(Y_{1,j}) &= \mathbb{E}\left\{ hK_h(S_1 - s_0)m_j(S_1) \right\} \\
&= h \int K_h(s - s_0)m(s)f_S(s)ds \\
&= h \int K(x)m_j(s_0 + hx)f_S(s_0 + hx)dx \\
&= h \int K(x) \left\{ m_j^{(1)}(s_0)hx + \frac{1}{2}m_j^{(2)}(s_0)h^2x^2 \right\} \\
&\qquad \times \left\{ f_S(s_0) + f_S^{(1)}(s_0)hx + \frac{1}{2}f_S^{(2)}(s_0)h^2x^2 \right\} dx + o(h^3) \\
&= h^3 \int x^2 K(x)dx \left\{ m_j^{(1)}(s_0)f_S^{(1)}(s_0) + \frac{1}{2}m_j^{(2)}(s_0)f_S(s_0) \right\} + o(h^3).
\end{aligned}$$

Similarly, for $j = 1, 2$,

$$\mathbb{E}(\mathbf{Y}_1^2) = \mathbb{E}\left\{ h^2 K_h^2(S_1 - s_0)m_j^2(S_i) \right\} = h^3 \int x^2 K^2(x)dx\{m_j^{(2)}(s_0)\}^2 f_S(s_0) + o(h^3).$$

Thus,

$$\sum_{i=1}^{n} \mathbf{Y}_i = nh^3 \int x^2 K(x)dx \left\{ m^{(1)}(s_0)f_S^{(1)}(s_0) + \frac{1}{2}m^{(2)}(s_0)f_S(s_0) \right\} + O_p(\sqrt{nh^3}).$$

□

Lemma 5. *Assume that Assumptions 1-4 and 6 hold. Let* $\tilde{\mathbf{Y}}_i = hK_h(S_i - s_0)(\boldsymbol{\theta}_i - m(S_i))$*, for* $i = 1, \dots n$*. Then*

$$\frac{1}{\sqrt{nh}} \sum_{i=1}^{n} \tilde{\mathbf{Y}}_i \implies \mathcal{N}_2\left(\mathbf{0}, \int K^2(x)dx f_S(s_0)\boldsymbol{\Sigma}(s_0) \right).$$

Proof of Lemma 5. We will use the Linderberg-Feller central limit theorem for showing the asymptotic normality of $\sum_{i=1}^n \tilde{\mathbf{Y}}_i/\sqrt{nh}$. First, it is trivial that, for fixed n, $\tilde{\mathbf{Y}}_i$'s are independently and identically distributed, with $\mathbb{E}(\tilde{\mathbf{Y}}_1) = 0$. Next, we study the variance of $\sum_{i=1}^n \tilde{\mathbf{Y}}_i/\sqrt{nh}$, which is $\mathbb{E}(\mathbf{Y}_1\mathbf{Y}_1^\intercal)/h$. Now, write $\mathbf{Y}_1 = (Y_{1,1}, Y_{1,2})^\intercal$. For $j,k = 1,2$,

$$\begin{aligned}
\frac{1}{h}\mathbb{E}(Y_{1,j}Y_{1,k}) &= h\int K_h^2(s-s_0)\mathbb{E}\left[\{\boldsymbol{\theta}_{1,j} - m_j(s_1)\}\{\boldsymbol{\theta}_{1,k} - m_k(s_1)\}\,|S_1 = s\right] f_S(s)ds \\
&= h\int K_h^2(s-s_0)\Sigma_{jk}(s)f_S(s)ds \\
&= \int K^2(x)f(s_0+hx)\Sigma_{jk}(s_0+hx)dx \\
&= \int K^2(x)dx f(s_0)\Sigma_{jk}(s_0) + o(1),
\end{aligned}$$

by dominated convergence theorem with boundedness and continuity assumptions of f_S and Σ_{jk}.

And, next, we have to verify the Linderberg-Feller condition. In our case, it can be reformulated as, for any $\varepsilon > 0$,

$$\lim_{n\to\infty}\sum_{i=1}^n \mathbb{E}\left(\left\|\frac{\tilde{\mathbf{Y}}_i}{\sqrt{nh}}\right\|^2 I\left\{\left\|\frac{\tilde{\mathbf{Y}}_i}{\sqrt{nh}}\right\| > \varepsilon\right\}\right) = 0.$$

We verify this condition by showing $\lim_{n\to\infty}\Pr(\|\tilde{\mathbf{Y}}_1/\sqrt{nh}\| > \varepsilon) = 0$, for any $\varepsilon > 0$. This is equivalent to $\|\tilde{\mathbf{Y}}_1/\sqrt{nh}\| = o_p(1)$, which we verify by looking at the second moment of $\|\tilde{\mathbf{Y}}_1/\sqrt{nh}\|$.

$$\begin{aligned}
\mathbb{E}\left(\left\|\frac{\tilde{\mathbf{Y}}_1}{\sqrt{nh}}\right\|^2\right) &= \frac{1}{nh}\mathbb{E}\left\{h^2K_h^2(S_1-s_0)\|\boldsymbol{\theta}_1 - \mathbf{m}(S_1)\|^2\right\} \\
&= \frac{h}{n}\int K_h^2(s-s_0)\mathbb{E}(\|\boldsymbol{\theta}_1 - \mathbf{m}(S_1)\|^2|S_1 = s)f_S(s)ds \\
&= \frac{h}{n}\int K_h^2(s-s_0)\mathbb{E}(\text{trace}\left[\{\boldsymbol{\theta}_1 - m(S_1)\}\{\boldsymbol{\theta}_1 - \mathbf{m}(S_1)\}^\intercal\right]|S_1 = s)f_S(s)ds \\
&= \frac{h}{n}\int K_h^2(s-s_0)\text{trace}\left\{\boldsymbol{\Sigma}(s)\right\}f_S(s)ds \\
&= \frac{1}{n}\int K(x)\text{trace}\left\{\boldsymbol{\Sigma}(s_0+hx)\right\}f_S(s_0+hx)dx \\
&= \frac{1}{n}\left[\{\Sigma_{11}(s_0) + \Sigma_{22}(s_0)\}f_S(s_0) + o(1)\right].
\end{aligned}$$

Thus, $\|\tilde{\mathbf{Y}}_1/\sqrt{nh}\| = o_p(1)$ and by continuous mapping theorem, $\|\tilde{\mathbf{Y}}_1/\sqrt{nh}\|^2 = o_p(1)$.

$$\sum_{i=1}^n \mathbb{E}\left(\left\|\frac{\tilde{\mathbf{Y}}_i}{\sqrt{nh}}\right\|^2 I\left\{\left\|\frac{\tilde{\mathbf{Y}}_i}{\sqrt{nh}}\right\| > \varepsilon\right\}\right) = \mathbb{E}\left(n\left\|\frac{\tilde{\mathbf{Y}}_1}{\sqrt{nh}}\right\|^2 I\left\{\left\|\frac{\tilde{\mathbf{Y}}_1}{\sqrt{nh}}\right\| > \varepsilon\right\}\right)$$

Call the term inside the expectation of the right hand side as Z_n. From above, $\mathbb{E}(n\|\tilde{\mathbf{Y}}_1/\sqrt{nh}\|^2) < \infty$, for sufficiently large n. Note that $Z_n \leq n\|\tilde{\mathbf{Y}}_1/\sqrt{nh}\|^2$. Thus, by dominated convergence theorem with application of Skorohod Representation Theorem to extend the result to weakly convergent sequence of random variables, we have $\lim_{n\to\infty} \mathbb{E}(n\|\tilde{\mathbf{Y}}_1/\sqrt{nh}\|^2) = 0$ and thus Linderberg-Feller condition is verified. Hence, by Linderberg-Feller central limit theorem, we have

$$\frac{1}{\sqrt{nh}} \sum_{i=1}^{n} \tilde{\mathbf{Y}}_i \implies \mathcal{N}_2\left(\mathbf{0}, \int K^2(x)dx f_S(s_0)\boldsymbol{\Sigma}(s_0)\right).$$

□

Lemma 6. *Assume that Assumption 1-4, 7 and 8 hold.*

$$M_n^{(2)}(\boldsymbol{\theta}_0) = nh\boldsymbol{\Psi}(s_0)f_S(s_0)\{1 + o_p(1)\}$$

Proof of Lemma 6. Note that $M_n^{(2)}(\boldsymbol{\theta}_0) = \sum_{i=1}^{n} hK_h(S_i - s_0)\boldsymbol{\psi}_2(\boldsymbol{\theta}_i, \boldsymbol{\theta}_0)$. To understand the asymptotic behavior of $M_n^{(2)}(\boldsymbol{\theta}_0)$, we study the asymptotic expansion of $hK_h(S_1 - s_0)\boldsymbol{\psi}_2(\boldsymbol{\theta}_i, \boldsymbol{\theta}_0)$ through computing its first two moments.

For $j, k = 1, 2$,

$$\mathbb{E}\left\{hK_h(S_1 - s_0)[\boldsymbol{\psi}_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_0))]_{j,k}\right\} = \int hK_h(s - s_0)\Psi_{jk}(s)f_S(s)ds = h\left\{\Psi_{jk}(s_0)f_S(s_0) + o(1)\right\},$$

by dominated convergence theorem with boundedness and continuity assumptions of f_S and Ψ_{jk}. As the second moment, since $E\{[\boldsymbol{\psi}_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_0)]_{j,k}^2 | S_1 = s\}$ is bounded,

$$\mathbb{E}\left\{h^2K_h^2(S_1 - s_0)[\boldsymbol{\psi}_2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_0))]_{j,k}^2\right\} \leq C_{jk}\int h^2K_h^2(s - s_0)f_S(s)ds = h\left\{C_{jk}f_S(s_0)\int K^2(x)dx + o(1)\right\}$$

by dominated convergence theorem with boundedness and continuity of f_S. Thus,

$$M_n^{(2)}(\boldsymbol{\theta}_0) = nh\boldsymbol{\Psi}(s_0)f_S(s_0)\{1 + o_p(1)\}.$$

□

Lemma 7. *Assume that Assumptions 1-4 and 7-10 hold. Let $\boldsymbol{\theta} \in \mathbb{R}^2$. For all sufficiently small $\delta > 0$,*

$$\lim_{n\to\infty} \Pr\left[\inf_{\tilde{\boldsymbol{\theta}} \in \mathcal{B}_\delta(\boldsymbol{\theta}_0)} \frac{1}{nh}\left\{(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\intercal M_n^{(2)}(\tilde{\boldsymbol{\theta}})(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\right\} \geq \frac{1}{2}f_S(s_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\intercal\boldsymbol{\Psi}(s_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\right] = 1.$$

Proof of Lemma 7. In this proof, we will prepare the uniform result that is required to show consistency of our estimator. Write $\mathbf{T}_n(\tilde{\boldsymbol{\theta}}) = (1/n)\sum_{i=1}^n K_h(S_i - s_0)\left\{\boldsymbol{\psi}_2(\boldsymbol{\theta}_i, \tilde{\boldsymbol{\theta}}) - \boldsymbol{\psi}_2(\boldsymbol{\theta}_i, \boldsymbol{\theta}_0)\right\}$. Note that

$$\sup_{\tilde{\boldsymbol{\theta}}\in\mathcal{B}_\delta(\boldsymbol{\theta}_0)} \|\mathbf{T}_n(\tilde{\boldsymbol{\theta}})\| \le \frac{1}{n}\sum_{i=1}^n \left\{ K_h(S_i - s_0) \sup_{\tilde{\boldsymbol{\theta}}\in\mathcal{B}_\delta(\boldsymbol{\theta}_0)} \left\|\boldsymbol{\psi}_2(\boldsymbol{\theta}_i, \tilde{\boldsymbol{\theta}}) - \boldsymbol{\psi}_2(\boldsymbol{\theta}_i, \boldsymbol{\theta}_0)\right\| \right\}.$$

By dominated convergence theorem and Assumption 9, we have

$$\begin{aligned}
\mathbb{E}\left(\sup_{\tilde{\boldsymbol{\theta}}\in\mathcal{B}_\delta(\boldsymbol{\theta}_0)} \|\mathbf{T}_n(\tilde{\boldsymbol{\theta}})\| \right) &\le \int K_h(s - s_0)\gamma(\delta, s) f_S(s) ds \\
&= \int K(x)\gamma(\delta, s_0 + hx) f_S(s_0 + hx) dx \\
&\le \tilde{\gamma}(\delta) f(s_0) + o(1).
\end{aligned}$$

By Assumption 9, we have $\lim_{\delta\to 0} \operatorname{limsup}_{n\to\infty}[\sup_{\tilde{\boldsymbol{\theta}}\in\mathcal{B}(\boldsymbol{\theta}_0)} \|\mathbf{T}_n(\tilde{\boldsymbol{\theta}})\|] = 0$ in probability. For a given $\boldsymbol{\theta} \in \mathbb{R}^2$, note that

$$\sup_{\tilde{\boldsymbol{\theta}}\in\mathcal{B}_\delta(\boldsymbol{\theta}_0)} \left|(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\intercal \mathbf{T}_n(\tilde{\boldsymbol{\theta}})(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\right| \le \left(\sup_{\tilde{\boldsymbol{\theta}}\in\mathcal{B}_\delta(\boldsymbol{\theta}_0)} \|\mathbf{T}_n(\tilde{\boldsymbol{\theta}})\| \right)^2 \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2.$$

Thus, by Lemma 6 and Assumption 10, for all sufficiently small $\delta > 0$,

$$\lim_{n\to\infty} \Pr\left[\inf_{\tilde{\boldsymbol{\theta}}\in\mathcal{B}_\delta(\boldsymbol{\theta}_0)} \frac{1}{nh}\left\{(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\intercal M_n^{(2)}(\tilde{\boldsymbol{\theta}})(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\right\} \ge \frac{1}{2} f_S(s_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\intercal \boldsymbol{\Psi}(s_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \right] = 1.$$

□

Proof of Theorem 2(a). To show the consistency result, we look into the Taylor's expansion of $M_n(\boldsymbol{\theta})$ around $\boldsymbol{\theta}_0$. Consider $\boldsymbol{\theta} \in \mathcal{B}_\delta(\boldsymbol{\theta}_0)$ and by Taylor's expansion, we have

$$M_n(\boldsymbol{\theta}) - M_n(\boldsymbol{\theta}_0) = M_n^{(1)}(\boldsymbol{\theta}_0)^\intercal(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\intercal M_n^{(2)}(\boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}_0),$$

where $\boldsymbol{\theta}^*$ lies on the line segment joining $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}$. First, by Lemma 4 and 5,

$$\frac{1}{nh} M_n^{(1)}(\boldsymbol{\theta}_0) = -\frac{2}{nh}\sum_{i=1}^n \left(\mathbf{Y}_i + \tilde{\mathbf{Y}}_i\right) = o_p(1).$$

Then, from Lemma 7, we have, for all sufficiently small $\delta > 0$,

$$\lim_{n\to\infty} \Pr\left[\inf_{\boldsymbol{\theta}^*\in\mathcal{B}_\delta(\boldsymbol{\theta}_0)} \frac{1}{nh}\left\{(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\intercal M_n^{(2)}(\boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\right\} \ge \frac{1}{2} f_S(s_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\intercal \boldsymbol{\Psi}(s_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \right] = 1.$$

Thus, by Assumption 10, there exists a local minimum in $\mathcal{B}_\delta(\boldsymbol{\theta}_0)$ asymptotically. That means, for any $\delta > 0$, there exists a sequence of roots, $\hat{\boldsymbol{\theta}}_n$, to $M_n^{(1)}(\boldsymbol{\theta}) = 0$ such that,

$$\lim_{n\to\infty} \Pr(\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| < \delta) = 1.$$

This completes the proof of Theorem 2(a). □

Proof of Theorem 2(b). To show the distributional result, we expand $M_n^{(1)}(\boldsymbol{\theta})$ by Taylor's expansion. Expanding at $\hat{\boldsymbol{\theta}}_n$, stated in Theorem 2(b),

$$0 = M_n^{(1)}(\hat{\boldsymbol{\theta}}_n) = M_n^{(1)}(\boldsymbol{\theta}_0) + M_n^{(2)}(\boldsymbol{\theta}_n^*)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0),$$

where $\boldsymbol{\theta}_n^*$ lies on the line segment joining $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}_n$. Note that $\|\boldsymbol{\theta}_n^* - \boldsymbol{\theta}_0\| \leq \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| = o_p(1)$. And

$$\frac{1}{nh} M_n^{(2)}(\boldsymbol{\theta}_n^*) = f_S(s_0)\boldsymbol{\Psi}(s_0)\{1 + o_p(1)\}$$

since, with $\|\boldsymbol{\theta}_n^* - \boldsymbol{\theta}_0\| = o_p(1)$, one can show that $\mathbb{E}\{\mathbf{T}_n(\boldsymbol{\theta}_n^*)\} = o_p(1)$ along the proof of Lemma 7. As for $M_n^{(1)}(\boldsymbol{\theta}_0)$, by Lemma 4,

$$\begin{aligned}\frac{1}{\sqrt{nh}} M_n^{(1)}(\boldsymbol{\theta}_0) = (-2)\sqrt{nh^5} \int x^2 K(x) dx \left\{ m^{(1)}(s_0) f^{(1)}(s_0) + \frac{1}{2} m^{(2)}(s_0) f_S(s_0) \right\} \\ + (-2)\frac{1}{\sqrt{nh}} \sum_{i=1}^{n} \tilde{\mathbf{Y}}_i + o_p(1)\end{aligned}$$

Thus, by Slutsky's theorem,

$$\sqrt{nh} \left\{ (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) - h^2 \boldsymbol{\eta} \right\} \implies \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Omega}).$$

□

S6 Simulation study

This section presents simulation results of the proposed DiST procedure.

We simulate 200 diffusion tensor data sets from the tensor field displayed in Figure S2 (Left). The tensors all have the principal eigenvalues being 4×10^{-3} and FA (5) being 0.9. The b-value is set to be 1000 across all voxels. This mimics the b-value and diffusivity (reflected by the numerical scale of the tensor) in real dMRI studies.

At each voxel there is either one tensor or there are two tensors. For crossing fiber regions, p_1 and p_2 are set to 0.7 and 0.3 respectively, and the separation angles between the two tensors range from 66.3 to 86.6 degree. In crossing fiber regions of Figure S2 (Left), the more transparent the tensor is, the less weight it takes.

In addition, $S_0(\mathbf{s})$'s have the same value which is set to 1000. Two choices of the noise standard deviation σ are used, namely 50 and 100, which corresponds to signal-to-noise ratio (S_0/σ) of 20 and 10, respectively. The case that $\mathsf{SNR} = 20$ is typical for dMRI studies while that of $\mathsf{SNR} = 10$

corresponds to a high noise setting. The set of gradient directions $\mathcal{U}$ is obtained from the sphere tessellation with 3 subdivision using octahedron and $|\mathcal{U}| = 33$, which is within a typical range for dMRI studies nowadays. With these gradient directions, the observed signal intensities $\mathbf{S}(\mathbf{s})$'s are simulated according to the multi-tensor model (1) with the Rician noise. A total of four different procedures are compared:

- raw: voxel-wise estimation without any smoothing;
- DiST-cv: DiST with h chosen by ordinary cross-validation score;
- DiST-tcv: DiST with h chosen by 5% trimmed cross-validation score;
- DiST-mcv: DiST with h chosen by median cross-validation score.

See Section S3 of the Supplemental Material for definitions of the various cross-validation variants.

Table S2 shows numerical summaries of the simulation results. In addition to the proportion of correctly estimated number of diffusion directions, we also report the average MSE (AMSE) and the average root MSE (ARMSE), defined as follows. Conditional on the correct estimation of J, the squared error of $\mathbf{m}$ is defined as

$$\min_{\{k_1,\ldots,k_J\in\{1,\ldots,J\}:k_i\neq k_j\}} \sum_{j=1}^{J} d^{*2}(\mathbf{m}_j, \hat{\mathbf{u}}_{k_j}), \tag{S8}$$

where $\hat{\mathbf{u}}_1, \ldots, \hat{\mathbf{u}}_J$ are the estimated diffusion directions. The minimization is taken in (S8) due to the ambiguity in direction estimator assignemnts to the true directions. Here, the MSE is the mean of squared errors (S8) over voxels with $\hat{J} = J$ in one simulated data set and root MSE (RMSE) is the square root of MSE. Then AMSE and ARMSE are defined, respectively, as the averages of MSEs and RMSEs over the 200 simulated data sets.

The voxel-wise estimation (i.e. raw) works reasonably well in estimating both the number of diffusion directions J and the diffusion directions. Even for the low SNR setting, the correctness of estimation of J is around 75% and the angular error is no more than 11 degree for the crossing fiber region ($J = 2$). Moreover, smoothing substantially improves the raw results. Specifically, for the single tensor region ($J = 1$), smoothing improves upon estimation of both J and diffusion directions. For regions with two tensors ($J = 2$), smoothing only improves direction estimation. Among the three smoothing procedures, DiST-mcv works the best.

Table S3 shows the five-number summary of the maximum angular error with $\hat{J} = J = 2$ across the 200 simulated data sets. Again smoothing procedures have smaller errors than the raw

procedure and DiST-mcv is the best among all methods. For DiST-mcv, the mean and median of maximum angular errors are around 2.5 degree and 1 degree for SNR = 10 and SNR = 20, respectively. Such magnitude of errors has little impact on tracking.

We then apply the proposed tracking algorithm in Section 5 (Algorithm S5, Supplemental Material) to the estimated diffusion directions based on the raw and DiST-mcv. The tracking results of a simulation with SNR = 10 are shown in Figures S2 (Right) and S3. As can be seen in Figure S3, the lines produced by DiST are much more aligned when compared to the tracking result based on voxel-wise estimation without smoothing (raw).

Table S2: Diffusion direction estimation results. **Correct-select**: proportion of $\hat{J} = J$. **AMSE**: average of MSEs (Each MSE is computed over voxels with $\hat{J} = J$ in one simulated data set.), in squared degree, of the estimated diffusion direction, with the corresponding standard error stated in brackets. **ARMSE**: average of RMSEs (Each RMSE is computed over voxels with $\hat{J} = J$ in one simulated data set.), in degree, of the estimated diffusion direction, with the corresponding standard error stated in brackets.

SNR	J		raw	DiST-cv	DiST-tcv	DiST-mcv
10	1	Correct-select	97.12%	99.09%	99.15%	99.45%
		AMSE	9.84 (3.84e-02)	4.95 (2.94e-01)	2.70 (1.09e-01)	3.06 (1.40e-01)
		ARMSE	3.14 (6.12e-03)	2.09 (5.46e-02)	1.60 (2.60e-02)	1.69 (3.13e-02)
	2	Correct-select	75.18%	74.38%	75.37%	75.44%
		AMSE	114 (2.42)	50.9 (3.45)	40.0 (3.11)	9.81 (1.40)
		ARMSE	10.6 (1.07e-01)	6.05 (2.68e-01)	5.26 (2.49e-01)	2.49 (1.35e-01)
20	1	Correct-select	98.59%	99.46%	99.69%	99.75%
		AMSE	2.30 (8.50e-03)	1.25 (1.23e-01)	7.97e-01 (3.02e-02)	1.15 (5.47e-02)
		ARMSE	1.52 (2.80e-03)	1.02 (3.28e-02)	8.79e-01 (1.10e-02)	1.03 (2.04e-02)
	2	Correct-select	99.38%	99.94%	99.99%	99.99%
		AMSE	19.8 (2.12e-01)	6.43 (5.18e-01)	2.00 (2.84e-01)	1.48 (2.13e-01)
		ARMSE	4.43 (2.34e-02)	2.13 (9.75e-02)	1.13 (6.02e-02)	9.93e-01 (4.98e-02)

Table S3: Summary statistics of the maximum absolute error across the voxels with $\hat{J} = J = 2$.

SNR	Method	Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum
10	raw	0.530	6.63	9.86	11.8	14.6	89.3
	DiST-cv	0.132	2.32	4.99	6.97	9.59	89.3
	DiST-tcv	0.0933	2.08	4.01	6.00	8.07	89.3
	DiST-mcv	0.135	1.35	2.11	2.91	3.35	65.1
20	raw	0.350	3.20	4.67	5.20	6.65	29.5
	DiST-cv	0.0803	0.931	1.73	2.48	3.28	26.1
	DiST-tcv	0.0494	0.613	0.965	1.33	1.53	15.7
	DiST-mcv	0.0473	0.531	0.841	1.16	1.40	15.9

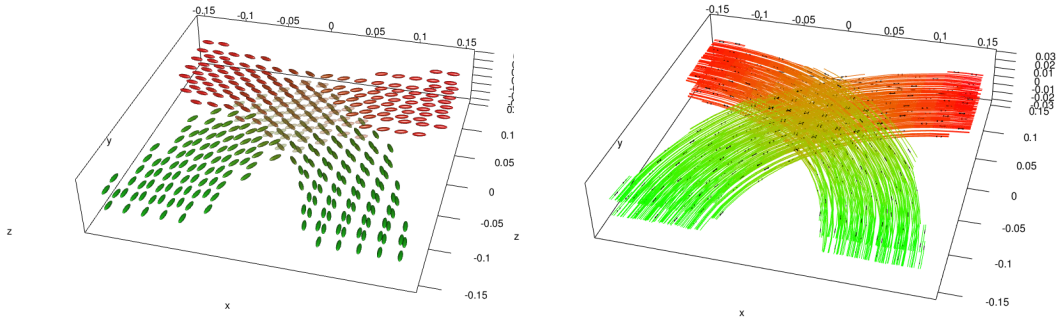

Figure S2: Left: The true tensor field used in the simulation study (Section S6). Right: Illustration of fiber tracking using DiST-mcv.

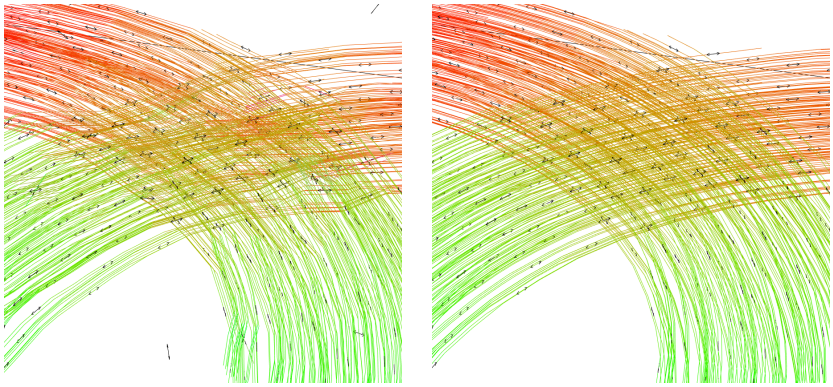

Figure S3: Illustration of fiber tracking over the crossing fiber region by raw (left) and DiST-mcv (right) respectively.

References

Abramowitz, M. and Stegun, I. A. (1964) *Handbook of Mathematical Functions: With Formulars, Graphs, and Mathematical Tables*, vol. 55. National Bureau of Standards.

Bhattacharya, A. and Bhattacharya, R. (2012) *Nonparametric inference on manifolds: with applications to shape spaces*. Cambridge University Press.

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

Kaufman, L. and Rousseeuw, P. J. (1990) *Finding groups in data: an introduction to cluster analysis*, vol. 344. New Jersey: John Wiley & Sons.

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

Sekhon, J. S. and Mebane, W. R. (1998) Genetic optimization using derivatives. *Political Analysis*, **7**, 187–210.