Bures-Wasserstein Barycentric Coordinates with Application to Diffusion Tensor Image Smoothing

Hanning Tang*, Xiaojing Shen*, Hua Zhao‡, Zhiguo Wang* and Pramod K. Varshney¶
*College of Mathematics, Sichuan University, Chengdu, China
‡Beijing Institute of Tracking and Telecommunication Technology, Beijing, China
¶Department of Electrical Engineering and Computer Science, Syracuse University, NY, 13244, USA

***Abstract*—This article considers the Wasserstein barycentric coordinates problem for Gaussian distributions which is the inverse problem of the Wasserstein barycenter problem. These coordinates take into account the underlying geometry of the measure space of Gaussian distributions and are thus meaningful for applications such as diffusion analysis and distributed information fusion. When the probability supports are discrete and identical, the theory of Wasserstein barycentric coordinates is well developed. However, for general probability distributions, the computation of Wasserstein barycentric coordinates is intractable since the technical hurdles involve solving a non-convex and non-concave optimization problem. For Gaussian distributions, we derive the closed-form expression of the derivatives for the objective function and propose a projected gradient descent method to solve the problem. Finally, we illustrate its application in diffusion tensor image (DTI) denoising including simulated DTI with different noise levels and DTI of the human brain.**

***Index Terms*—Optimal transport, Wasserstein barycenter coordinates, diffusion tensor image smoothing**

I. Introduction

Wasserstein barycenter [1] has been introduced as a more desirable notion of average using the Wasserstein distance [2] since it incorporates the geometry of the underlying space. For example, when considering the average of the images, use of the Euclidean distance is unsuitable for many applications [3]. Wasserstein barycenters have been applied in many areas and with many potential applications, such as in graphics, neuroscience, statistics, economics, information fusion, and machine learning [4]–[13]. When considering the set of all non-degenerate Gaussian measures endowed with the 2-Wasserstein metric, W_2, this set forms a geodesic metric space [14], [15]. When limited to the centered cases, Gaussian distributions are identified with their covariance matrices, and the Wasserstein distance induces a metric on the space of positive definite matrices. This metric is also known as the Bures or Bures-Wasserstein metric [16].

In practice, Wasserstein barycenters are estimated using iterative algorithms [3], [17]–[23]. When considering the probability measure with common discrete support, the computation of Wasserstein barycenters is a convex optimization problem with additional linear constraints. However, for continuous probability measures, one can define the Wasserstein gradient of the functional W_2 and consider a non-Euclidean gradient descent approach. In the Gaussian cases, the gradient descent takes the form of a concrete and tractable update equation on the mean and the covariance matrix of the candidate barycenter. Such an algorithm was first proposed in [18], where it is described as a fixed-point algorithm. This algorithm was further generalized in [20], [24] and the linear convergence rate of this algorithm was shown in [21].

The Wasserstein barycentric coordinates problem is the inverse problem associated with Wasserstein barycenters [4]. Given N probability measures, a vector of N weights can be associated with another input probability measure, such that the barycenter of given measures with those weights approximates as closely as possible the input probability measure. The corresponding barycenter can be interpreted as a reconstructed version of the original input, with respect to the prior contained in those reference measures. For general probability distributions, the Wasserstein barycentric coordinates problem is non-convex and non-concave, and the technical hurdles of finding a local solution are computing the Wasserstein barycenter and optimizing the coordinates. An approach to solving this inverse problem for discrete measures was proposed in [4]. In this paper, we consider this problem for Gaussian distributions. The major novelty in this paper lies in the fact that the barycentric coordinates, in the sense of the optimal transport metric, give our barycentric coordinate system an intuitive and geometrically faithful flavor. We also provide efficient algorithms to compute the Wasserstein barycentric coordinates for Gaussian probability measures. Our major contributions in this paper are as follows:

- We define the Wasserstein barycentric coordinates for Gaussian distributions and derive an analytical expression of the derivatives of the objective function.
- We provide a algorithm to find a local solution of the problem using projected gradient descent by solving the Gaussian Wasserstein barycenter and approximate the derivatives at each iteration.
- We apply the Wasserstein barycentric coordinates to the problem of diffusion tensor image (DTI) smoothing. The simulation result shows that the proposed method has better performance than existing method.

The structure of this paper is as follows. In Section II, we revisit the basic concept of optimal transportation and Wasserstein barycenter. Section III presents the definition of the Wasserstein barycentric coordinates, its analytical expression of the gradient, and a general first-order optimization algorithm

to solve this problem. Application to diffusion tensor field smoothing is given in Section IV. Simulation results are presented in Section V. In Section VI, concluding remarks are provided.

II. Preliminaries

A. Notations

In this paper, we present scalars, vectors and matrices in lower case letters, bold lower case letters and upper case letters, respectively. We use $\mathbb{R}$, $\mathbb{R}^n$, $\mathbb{R}^{m\times n}$ to denote the set of real numbers, n-dimensional real vectors and $m \times n$ real matrices, respectively. And we use $\mathbb{S}^n$, $\mathbb{S}^n_+$, $\mathbb{S}^n_{++}$ to denote the set of symmetric, symmetric positive semi-definite and symmetric positive definite $n \times n$ matrices, respectively. The set of Borel probabilities on $\mathbb{R}^n$ with a finite second moment is denoted by $\mathcal{P}_2(\mathbb{R}^n)$.

For a vector $\mathbf{x}$, x_i denotes its i-th entry, $||\mathbf{x}||$ denotes its Euclidean norm, $|\mathbf{x}|$ denotes its entry-wise absolute value. For a matrix $\mathbf{X}$, x_{ij} denotes its (i,j)-th entry. $\text{vec}(\mathbf{X})$, $\mathbf{X}^T$ and $\text{Tr}(\mathbf{X})$ denotes the vectorization, the transpose and the trace of $\mathbf{X}$, respectively. We also use $\mathbf{x} \geq 0$ to denote $x_i \geq 0$ for all i. The identity matrix of size $n \times n$ is denoted by $\mathbf{I}_n$. The i-th standard basis vector is denoted by $\mathbf{e}_i$. For any $\mathbf{X} \in \mathbb{R}^{m\times n}$ and $\mathbf{Y} \in \mathbb{R}^{m'\times n'}$, the Kronecker product $\mathbf{X} \otimes \mathbf{Y}$ is defined as

$$\mathbf{X} \otimes \mathbf{Y} = \begin{pmatrix} x_{11}\mathbf{Y} & \dots & x_{1n}\mathbf{Y} \\ \dots & \dots & \dots \\ x_{m1}\mathbf{Y} & \dots & x_{mn}\mathbf{Y} \end{pmatrix}.$$

For any $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n\times n}$ the Kronecker sum is defined as $\mathbf{X} \oplus \mathbf{Y} = \mathbf{I}_n \otimes \mathbf{Y} + \mathbf{X} \otimes \mathbf{I}_n$ [25]. The n dimensional standard simplex is denoted by $\Delta_n = \{\mathbf{x} \in \mathbb{R}^n | \mathbf{x} \geq 0, \sum_i^n x_i = 1\}$.

B. Wasserstein distance

Given two probability measures with finite second moments μ and $v \in \mathcal{P}_2(\mathbb{R}^d)$, the 2-Wasserstein distance between μ and v is defined as the value of the following Monge-Kantorovich problem [2]:

$$W_2(\mu, v) := \sqrt{\inf_{\gamma\in\Pi(\mu,v)} \int_{\mathbb{R}^d\times\mathbb{R}^d} ||x-y||_2^2 \mathbf{d}\gamma(x,y)},$$

where $\Pi(\mu, v)$ denotes the set of probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ having μ and v as marginals [2]. Wasserstein distance exhibits some attractive features and has been used extensively in statistics, machine learning, and so on [2], [26].

In particular, when $v_i \sim \mathbf{N}_d(\mathbf{x}_i, \mathbf{P}_i)$, the square Wasserstein distance between v_1 and v_2 has a closed form expression [27]

$$\begin{aligned} W_2^2(v_1, v_2) =& \text{Tr}(\mathbf{P}_1 + \mathbf{P}_2 - 2(\mathbf{P}_1^{1/2}\mathbf{P}_2\mathbf{P}_1^{1/2})^{1/2}) \\ &+ \|\mathbf{x}_1 - \mathbf{x}_2\|_2^2. \end{aligned}$$

C. Wasserstein barycenters

Agueh and Carlier defined Wasserstein barycenters [1] as Fréchet means in the space of probability measures endowed with the Wasserstein metric.

Definition 1 (Agueh and Carlier [1])**.** *Given a set of probability distributions $\{v_i\}_{i=1}^N$ with $v_i \in \mathcal{P}_2(\mathbb{R}^d)$. The Wasserstein barycenter or Fréchet mean with weights $\mathbf{w} \in \mathbb{R}^N$ is defined as an optimal solution of the following problem*

$$\inf_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} F(\mu) := \sum_{i=1}^k w_i W_2^2(v_i, \mu), \tag{1}$$

where $\mathbf{w} = (w_1, \dots, w_N)$ satisfying $\sum_{i=1}^N w_i = 1$ and $w_i \geq 0$. Under mild condition this problem has a unique solution which is referred as the barycenter of the measures v_i with weights w_i, denoted by $\overline{v} = bar(v_i, w_i)_{i=1,\dots,N}$.

Lemma 1 (Agueh and Carlier [1])**.** *In the non-degenerate Gaussian case $v_i \sim \mathbf{N}_d(\mathbf{x}_i, \mathbf{P}_i)$, there is a unique solution $\overline{v}$ to (1). Moreover, $\overline{v} \sim \mathbf{N}_d(\overline{\mathbf{x}}, \mathbf{P})$, where $\overline{\mathbf{x}} = \sum_{i=1}^N w_i\mathbf{x}_i$ and $\mathbf{P}$ is the unique positive definite root of the matrix equation*

$$\sum_{i=1}^N w_i(\mathbf{P}^{\frac{1}{2}}\mathbf{P}_i\mathbf{P}^{\frac{1}{2}})^{\frac{1}{2}} = \mathbf{P}. \tag{2}$$

Wasserstein barycenter is very different from the usual linear averaging $\sum_{i=1}^N w_i v_i$, which corresponds to the barycenter using Euclidean distance. There is no closed-form expression for $\overline{v}$ in general. In practice, when v_i are distributions with a fixed common discrete support, the computation of Wasserstein barycenters is a convex optimization problem with additional structure [3], [28]–[30]. For more general cases where the discrete support points are not fixed, the Wasserstein barycenter problem is non-convex, which is estimated using iterative, first order algorithms [3], [18]–[20]. In particular, one can define the Wasserstein gradient of a functional in measure space and consider an intrinsic gradient descent-based approach to estimate Wasserstein barycenter. Such an algorithm was proposed in [18], where it is described as a fixed-point algorithm for Gaussian distributions. They proved that the fixed-point algorithm converges to the true barycenter as the number of iterations goes to infinity. The results were further generalized in [20] for general regular distributions. Recently, the linear convergence of the gradient descent algorithm in Wasserstein space was proved in [21] for Gaussian distributions. The corresponding iteration is equivalent to the fixed point iteration in [18]

$$\begin{aligned} \mathbf{S}_{k+1} =& \mathcal{F}(\mathbf{S}_k, \mathbf{w}_k, \{\mathbf{P}_i\}) \\ :=& \mathbf{S}_k^{-1/2}(\sum_{i=1}^N w_i(\mathbf{S}_k^{1/2}\mathbf{P}_i\mathbf{S}_k^{1/2})^{1/2})^2\mathbf{S}_k^{-1/2}. \end{aligned} \tag{3}$$

$\mathbf{S}_k$ converge to the solution of equation (2) as $k \to \infty$.

D. Wasserstein barycentric coordinates

The main problem addressed in this paper is Wasserstein barycentric coordinates. Given N probability measures $\{v_i\}$ and weights $\mathbf{w} \in \Delta_N$, we assume that the Wasserstein barycenter of these probability measures is well defined. Let $P(\mathbf{w})$ denotes the barycentric map associates to the weights $\mathbf{w}$:

$$P(\mathbf{w}) = \text{bar}(v_i, w_i)_{i=1,\dots,N}. \tag{4}$$

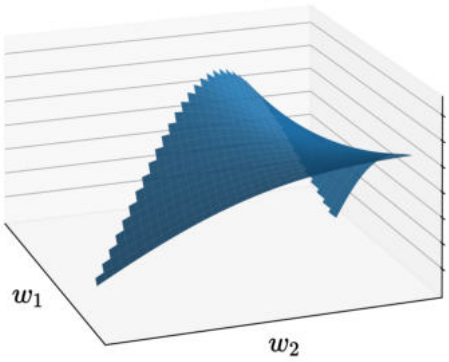


Fig. 1: A example of $\mathcal{E}(\mathbf{w})$ where $\mathbf{w} = (w_1, w_2, 1-w_1-w_2)$. It is clear that this function is neither convex nor concave in this case. See Appendix A for details.

For any probability measure v_0, the Wasserstein barycentric coordinate is a vector $\mathbf{w} \in \Delta_N$ such that $P(\mathbf{w})$ is equal or close to v_0 with respect to a loss function $\mathcal{L}$.

Definition 2. *Let $v_0, v_1, \ldots, v_N \in \mathcal{P}_2(\mathbb{R}^d)$. The barycentric coordinates of v_0 with respect to $\{v_i\}_i^N$ are given by any optimal solution of the problem*

$$\underset{\mathbf{w}\in\Delta_N}{\operatorname{argmin}} \ \mathcal{E}(\mathbf{w}) := \mathcal{L}(v_0, P(\mathbf{w})). \tag{5}$$

Wasserstein barycentric coordinates were first considered in [4] with assumptions that v_i are discrete distributions with common fixed support point.

In this paper, we consider the Wasserstein barycentric coordinates for Gaussian distributions, i.e. $v_i \sim \mathbf{N}_d(\mathbf{x}_i, \mathbf{P}_i), i = 0, \ldots, N$, where $\mathbf{P}_i$ is a positive definite matrix. And we set the loss function as $\mathcal{L}(\mu, v) = W_2^2(\mu, v)$. In this setting, the objective function (5) could be a non-convex function as shown in Fig. 1. Thus, our goal is to find a stationary point of the objective function.

III. Bures-Wasserstein barycentric coordinates for Gaussian distributions

In this section, we consider finding a stationary point of problem (5) through gradient descent. It is shown that the Wasserstein barycenter of Gaussian distributions is also Gaussian [1]. We denote the mean and the covariance matrix of $P(\mathbf{w})$ as $\mathbf{x}(\mathbf{w})$ and $\mathbf{P}(\mathbf{w})$, respectively, i.e. $P(\mathbf{w}) \sim \mathbf{N}_d(\mathbf{x}(\mathbf{w}), \mathbf{P}(\mathbf{w})))$.

Then problem (5) can be formulated as

$$\begin{aligned}\min_{\mathbf{w}\in\Delta_N} \ \mathcal{E}(\mathbf{w}) =& W_2^2(v_0, P(\mathbf{w})) \\ =& \mathrm{Tr}(\mathbf{P}_0 + \mathbf{P}(\mathbf{w}) - 2(\mathbf{P}_0^{1/2}\mathbf{P}(\mathbf{w})\mathbf{P}_0^{1/2})^{1/2}) \\ &+ \|\mathbf{x}_0 - \mathbf{x}(\mathbf{w})\|_2^2. \end{aligned} \tag{6}$$

When $\forall\ 0 \le i, j \le N, \mathbf{P}_j\mathbf{P}_i = \mathbf{P}_i\mathbf{P}_j$, the solution of the matrix equation (2) is

$$\mathbf{P}(\mathbf{w}) = (\sum_{i=1}^N w_i \mathbf{P}_i^{\frac{1}{2}})^2.$$

Problem (6) can be simplified as a quadratic optimization problem with linear constraints

$$\min_{\mathbf{w}\in\Delta_N} \ \mathbf{w}^T\Lambda\mathbf{w} + B\mathbf{w}$$

where $\Lambda = \{(\mathbf{x}_0 - \mathbf{x}_i)^T(\mathbf{x}_0 - \mathbf{x}_j) + \mathrm{Tr}(\mathbf{P}_i^{\frac{1}{2}}\mathbf{P}_j^{\frac{1}{2}})\}_{ij} \in \mathbb{R}^{N\times N}$, $B = (-2\mathrm{Tr}(\mathbf{P}_i^{1/2}\mathbf{P}_0^{1/2}))_i \in \mathbf{R}^N$.

A. Analytical expression of the gradient

The gradient of the objective function $\mathcal{E}(\mathbf{w})$ with respect to $\mathbf{w}$ can be computed using the chain rule. We show below that $\nabla\mathcal{E}$ can be analytically derived.

Lemma 2. *For any given positive definite matrices $\{\mathbf{P}_i\}_{i=1}^N$,*

$$\frac{\partial Tr(\mathbf{P}(\mathbf{w}))}{\partial w_j} = 2Tr(\mathbf{H}_j),$$

where $\mathbf{H}_j = (\mathbf{P}_\mathbf{w}^{1/2}\mathbf{P}_j\mathbf{P}_\mathbf{w}^{1/2})^{1/2}$.

Proposition 1. *For any given positive definite matrices $\{\mathbf{P}_i\}_{i=0}^N$ and $\{\mathbf{x}_i\}_{i=0}^N$, the derivative of the objective function $\mathcal{E}(\mathbf{w})$ is*

$$\begin{aligned}\frac{\partial\mathcal{E}(\mathbf{w})}{\partial w_j} =& 2(\mathbf{x}(\mathbf{w}) - \mathbf{x}_0)^T\mathbf{x}_j + 2Tr(\mathbf{H}_j) \\ &- vec(\mathbf{H}_0)^T\mathbf{V}vec(\mathbf{H}_j),\end{aligned} \tag{7}$$

where $\mathbf{V} = (\sum_{i=1}^N w_i(\mathbf{H}_i^{-1} \oplus \mathbf{H}_i^{-1})^{-1})^{-1}$, $\mathbf{H}_j = (\mathbf{P}_\mathbf{w}^{1/2}\mathbf{P}_j\mathbf{P}_\mathbf{w}^{1/2})^{1/2}$, $\mathbf{P}_\mathbf{w} = \mathbf{P}(\mathbf{w})$ for simplicity of notation.

Proof of Lemma 2 and Proposition 1 is omitted due to space constraints. A detailed proof will be provided in a subsequent journal publication. In general, the expression (7) is hard to implement. The hypothesis is that one is able to solve the matrix equation (2) exactly, which is not the case in practice, since the matrix is only approximated by iterating a sufficient number of times the map in (3) to converge to a sufficient accuracy.

B. Projected gradient descent method

Based on Proposition 1, We present a projected gradient descent method to solve the problem (6) in Algorithm 1 by iteratively solving the matrix equation (2).

Algorithm 1: Projected gradient descent (PGD)

Input: $v_i \sim \mathbf{N}(\mathbf{x}_i, \mathbf{P}_i)$, $i = 1, 2, \ldots, N$, $\{\alpha_k\}$
Let $\mathbf{w}_0 = \frac{1}{N}\mathbf{1}$, $\mathbf{P} = \mathbf{I}$,
for $k = 1, \ldots, K$ **do**
  for $t = 1, \ldots, T_k$ **do**
    $\mathbf{P} = \mathcal{F}(\mathbf{P}, \mathbf{w}_{k-1}, \{\mathbf{P}_i\})$ (using (3))
  end
  Calculate $\nabla\mathcal{E}(\mathbf{w}_{k-1})$ using (7)
  $\mathbf{w}_k = \mathrm{Proj}_{\Delta_N}(\mathbf{w}_{k-1} - \alpha_k\nabla\mathcal{E}(\mathbf{w}_{k-1}))$
end
return $\overline{v} \sim \mathbf{N}(\overline{\mathbf{x}}, \mathbf{P})$, *where* $\overline{\mathbf{x}} = \sum_{i=1}^N w_{K_i}\mathbf{x}_i$, $\mathbf{w}_K = (w_{K_1}, \ldots, w_{K_N})^T$.

Since $\mathbf{P}(\mathbf{w})$ is a Lipschitz continuous function, the objective function $\mathcal{E}(\mathbf{w})$ is also Lipschitz. By choosing an appropriate stepsize α, Algorithm 1 converges to a stationary point [31].

Note that the necessary optimal condition of the problem (6) is $\nabla\mathcal{E}(\mathbf{w}^*)^T\mathbf{w}^* - \min_i \nabla_i\mathcal{E}(\mathbf{w}^*) = 0$ [32]. For given tolerance $\epsilon > 0$, we set the stopping condition of Algorithm 1 as

$$\nabla\mathcal{E}(\mathbf{w}_n)^T\mathbf{w}_n - \min_i \nabla_i\mathcal{E}(\mathbf{w}_n) \leq \epsilon$$

IV. Application: diffusion tensor image smoothing

Diffusion tensor imaging (DTI) is a technique that is popular for imaging the white matter of the brain which is based on nuclear magnetic resonance [33]. The assumption is made that the random diffusion of water molecules at a given position in a biological tissue is Gaussian distributed with zero mean [34]. A diffusion tensor image is a semi-definite positive matrix-valued field in which the matrix associated with the current volume element (or voxel) is the covariance matrix of the local diffusion distribution. Assuming Gaussian diffusion of water molecules at a given voxel, the noiseless diffusion weighted signal intensity in direction $\mathbf{q} = (q_1, q_2, q_3)^T \in \mathbb{R}^3$ is given by

$$S_{\mathbf{q}} = S_{\mathbf{0}} \exp(-b\mathbf{q}^T\mathbf{D}\mathbf{q}),$$

where $S_{\mathbf{0}}$ is the baseline intensity determined by the background constant field, b is a fixed experimental constant and $\mathbf{D}$ is a 3×3 positive definite matrix which is referred to as the diffusion tensor at that voxel. The diffusion weighted signals of all voxels form a diffusion weighted image corresponds to the direction $\mathbf{q}$.

A three-dimensional diffusion model is estimated by creating diffusion weighted images of multiple directions. In order to calculate the 6 independent elements in the 3×3 symmetric matrix $\mathbf{D}$, at least 7 images are needed.

DTI may provide a low signal-to-noise ratio in some applications. In those cases, smoothing of DTI is essential. Different methods of smoothing have been explored in the literature. There are mainly two types of smoothing methods. The first type of methods smooth the diffusion weighted images and then estimate the DTI and the second type of methods directly smooth the estimated fusion tensor field.

We propose a tensor field smoothing method based on Gaussian Wasserstein barycentric. This method directly smooths the tensor field using the Bures-Wasserstein barycenter coordinates. For each voxel, we use the Gaussian distributions of its neighbors as references and the Gaussian distributions of this voxel as an input of the Wasserstein barycenter coordinates. Then we use the reconstructed Gaussian distributions as the smoothed tensor. By performing the optimization of the coordinates, this method uses similar distributions of the neighbors to take average. The reason for doing so is that the diffusion tensor of the water molecules is mostly determined by the fiber of the brain, and the diffusion tensors of water in the same fiber should be similar to each other.

Given a three-dimensional DTI $\{\mathbf{P}_{ijk}\}_{N^3}$, $\mathbf{P}_{ijk} \in \mathbb{R}^{3\times3}$ is positive definite matrix, denote $v_{ijk} = \mathbf{N}_3(\mathbf{0}, \mathbf{P}_{ijk})$. Consider the following problem:

$$\begin{aligned} \min_{\mathbf{w}\in\Delta_{|\mathcal{N}_{ijk}|}} \quad & \mathcal{E}(\mathbf{w}) = W_2^2(v_{ijk}, P(\mathbf{w})) \\ \text{s.t.} \quad & P(\mathbf{w}) = \operatorname*{argmin}_{\mu} \sum_{lmn\in\mathcal{N}_{ijk}} w_{lmn} W_2^2(v_{lmn}, \mu), \\ & i,j,k = 1,\ldots,N \end{aligned} \tag{8}$$

where $\mathcal{N}_{lmn} = \{(l,m,n) | \|(l,m,n) - (i,j,k)\|_1 = 1\}$ is the index set of neighboring voxels of (i,j,l). For a given voxel (i,j,k), we denote the corresponding solution of (8) as $\mathbf{w}^*_{ijk}$. The reconstructed covariance matrix $\mathbf{P}(\mathbf{w}^*_{ijk})$ is the smoothing result of $\mathbf{P}_{ijk}$. Then the smoothed DTI is $\{\mathbf{P}(\mathbf{w}^*_{ijk})\}_{N^3}$.

In general, the optimal solution to this problem is sparse, which means that this method only averages similar voxels among neighboring voxels. By considering the local structure, this method avoids the loss of information on the size and shape of structures and also avoids the blur on the edge. Simulation results and the smoothing performance on DTI of the human brain are presented in Section V.

V. Experimental results

In this section, we present numerical results to demonstrate the effectiveness and efficiency of our proposed method. To begin with, we illustrate the convergence rate of the projected gradient descent. Then, we evaluate the performance for diffusion tensor image smoothing.

A. Randomly generated Gaussian distributions

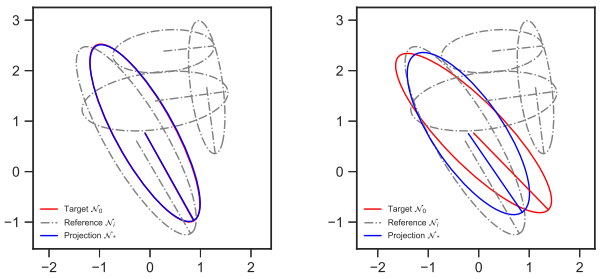
We initiated our experiment by considering two-dimensional Gaussian distributions as depicted in Fig. 2. For more general cases, we examined problem (6) with randomly generated parameters. Specifically, the vector the vectors $\mathbf{x}_i$ were drawn from a uniform distribution $\mathbf{U}([0,1]^d)$, while the matrices $\mathbf{P}_i \in \mathbb{S}_+^{d\times d}, i = 0,1,\ldots,N$ were generated randomly as $\mathbf{P}_i = \mathbf{A}\mathbf{A}^T + 0.01\mathbf{I}$ where each entry of matrix $\mathbf{A}$ followed a normal distribution $\mathbf{A}_{ij} \sim \mathbf{N}(0,1)$.

The step size of PGD was set to 0.01. In the inner loop of PGD, we continued iterating until the difference of the gradient between two iterations is less than $\epsilon = 10^{-8}$. Fig. 3 illustrates the convergence behavior of PGD.

B. Diffusion tensor image smoothing

Here we construct a simulated tensor field on a $128\times128\times3$ three-dimension grid consisting of the background regions with identical isotropic tensors and the banded regions with three parallel vertical bands and three horizontal bands (for each of the three slices), where within each band tensors are identical and aligned in either the x- or the y- direction. The bands are of various widths and degrees of anisotropy according to Simulation I in [35].

At each voxel, we simulate the raw diffusion weighted image (DWI) data using the true tensor at that voxel and the Gaussian noise model. When signal to noise ratio is



(a) $\min \mathcal{E}(\mathbf{w}) = 0$ (b) $\min \mathcal{E}(\mathbf{w}) > 0$

Fig. 2: Two-dimensional Gaussian example.

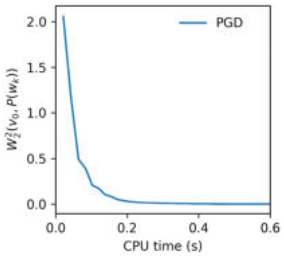


Fig. 3: CPU time *vs* error ($N = 6, d = 3$), $W_2^2(v_0, P(w_k))$ is the objective function. The result displays the average results over 50 times of Monte Carlo experiments.

greater than 5, the error distribution can be approximated by a Gaussian distribution [36],

$$S_{\mathbf{q}} = S_{\mathbf{0}} \exp(-b\mathbf{q}^T\mathbf{D}\mathbf{q} + v), \ v \sim \mathbf{N}(0, \sigma^2).$$

Specifically, we set $S_0 = 1000$, $b = 1$ and use 9 gradient directions of the following vectors

$$(1,0,1),(1,1,0),(0,1,1),(3,2,1),(0.9,0.45,0.2),$$
$$(1,0,0),(0,1,0),(0,0,1),(2,1,1.3).$$

Finally, at each voxel, weighted least squares is applied to the DWI data to derive the observed tensors as inputs for the smoothing procedure.

The fractional anisotropy (FA) [37] is the most widely used anisotropy measure for visualization. FA is a normalized variance of the eigenvalues of $\mathbf{D}$ defined by

$$FA = \frac{1}{\sqrt{2}} \frac{\sqrt{(\lambda_1 - \overline{\lambda})^2 + (\lambda_2 - \overline{\lambda})^2 + (\lambda_3 - \overline{\lambda})^2}}{\sqrt{\lambda_1^2 + \lambda_2^2 + \lambda_3^2}},$$

where λ_i is the eigenvalues of $\mathbf{D}$ and $\overline{\lambda} = (\lambda_1 + \lambda_2 + \lambda_3)/3$. Intuitively, FA measures how far the tensor $\mathbf{D}$ is from the identity matrix $\mathbf{I}$ and the range of FA is $[0, 1]$. Another type of image color FA can represent the orientation of the major eigenvector using a mapping to colors [38]. In this image, blue represents superior-inferior directions, green represents the anterior-posterior directions and red represents the left-right directions. The brightness of the color is controlled by FA.

In this experiment, we consider smooth slice 2 for comparison. The color FA of the ground truth tensor field are presented in Fig. 4. Note that slice 1 and slice 2 are identical while slice 3 is different. The noisy simulation ($\sigma^2 = 1/5$) and smoothed result of the slice 2 are presented in Fig. 5. It shows that the image smoothed by the proposed method has the highest peak signal-to-noise ratio (PSNR) and preserves sharp edges when two bands overlaps.

For further comparison, we consider the noise parameter $\sigma^2 = 1/5,\ 1/10,\ 1/20,\ 1/30$ which are referred to different noise levels. Errors in terms of l_2 and W_2 distance between the smoothed and the ground truth tensors are summarized in TABLE I. At all noise levels, our proposed method has smaller errors and also improves the peak signal-to-noise ratio (PSNR) 8 dB for the color FA images in all situations. It shows in Fig. 5 (b) that the edges of these bands are still sharp and the area within is smoothed by the proposed method. NLM [39], P2S [40], MPPCA [41], LPCA [42] and GO [43] are smoothing methods for denoising DWI data. But directly smoothing the tensor field considers the physical relationship between DWI of different gradient directions, which results in proposed method having a better performance. Compared with the kernel method [35], proposed method takes the local structure of the tensor field into consideration by optimizing the weights of each voxel.

We also applied our method on real DTI data from [44]. Fig. 6 presents the original DTI and smoothing results. It shows that the proposed method provides smooth images and preserves white matter structures well without evident blurring, compared to other methods. The clearer structure may help with providing insights into the micro-structure of the human brain.

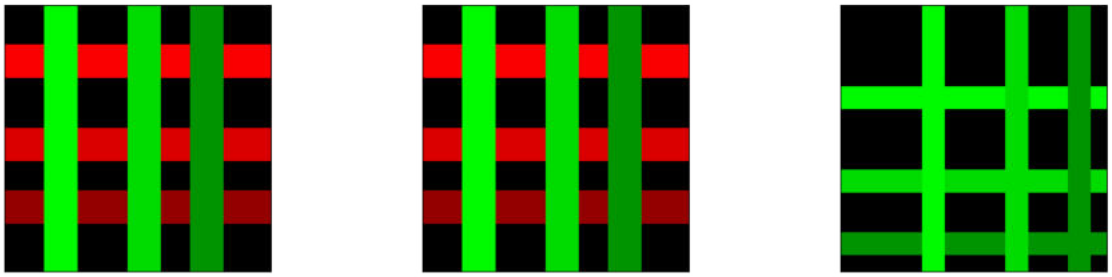
VI. Conclusion

This paper introduced Bures-Wasserstein barycentric coordinates for Gaussian distributions. In spite of the fact that this problem is non-convex and non-concave, we proposed a projected gradient descent method for local solutions with a low computational overhead with respect to the Wasserstein barycenter problem. We further illustrated its application in DTI data smoothing, which provides a new method that can simultaneously smooth DTI and keep the edges sharp.

For future work, the Gaussian assumption may fail in practice. Wasserstein barycenter coordinates for mixed Gaussian distributions or general continuous probability distributions are still open problems. The proposed method can also be considered for other practical applications in graphics, and machine learning, etc.

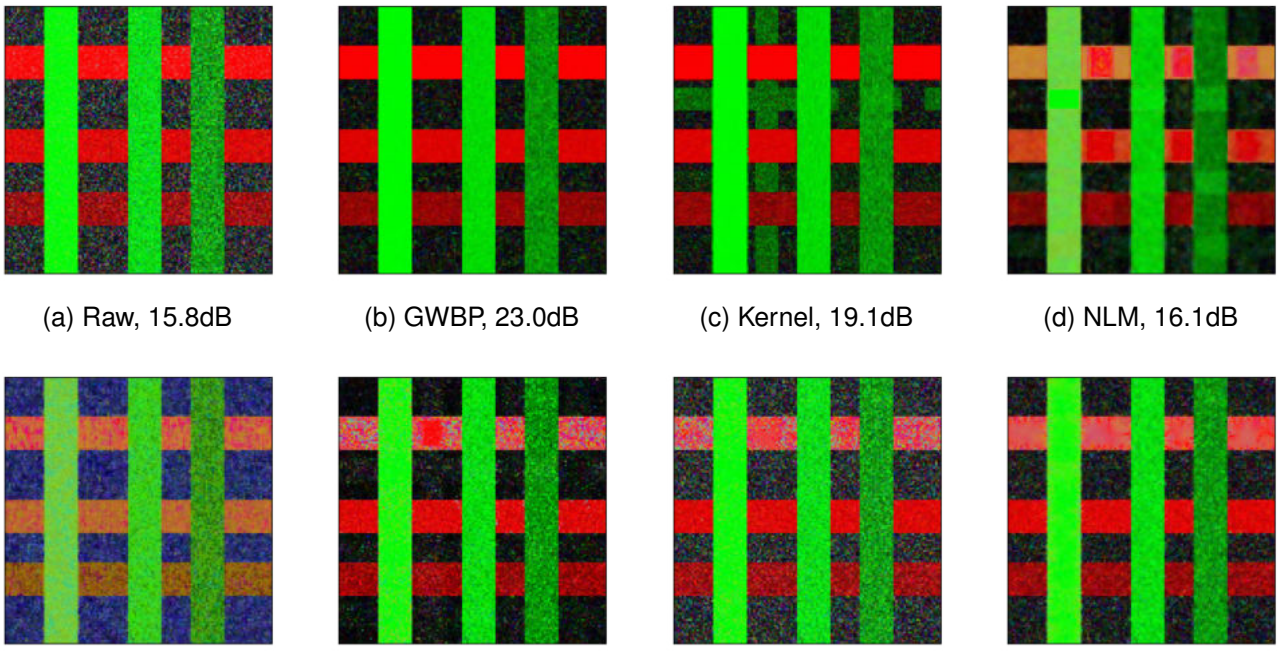
Acknowledgment

This work was supported in part by the NSFC under Grant 62203313, in part by the Natural Science Foundation of Sichuan Province under Grant 2022NSFSC1853, and in part by Sichuan Youth Science and Technology Innovation Team (Grant Nos. 2022JDTD0014, 2021JDJQ0036). Xiaojing Shen is the corresponding author. Email: xiao23332@163.com.



(a) Slice 1 (b) Slice 2 (c) Slice 3

Fig. 4: The color FA of the ground truth. Note that slice 1 and slice 2 are identical while slice 3 is different.



(a) Raw, 15.8dB (b) GWBP, 23.0dB (c) Kernel, 19.1dB (d) NLM, 16.1dB

(e) P2S, 11.5dB (f) MPPCA, 17.4dB (g) LPCA, 14.6dB (h) GO, 17.1dB

Fig. 5: Color FA images (slice 2) and its peak signal-to-noise ratio (PSNR) of simulated DTI and smoothing results. (a): Raw DTI estimated by simulated DWI, the noise level is $\sigma^2 = 1/5$; (b): The color FA images of DTI smoothed by the proposed method; (c): Kernel method using Euclidean distance with bandwidth = 0.75; (d)-(h): Results of different methods. The PSNR values computed on these images. It is noticeable that the image smoothed by the proposed method preserves sharper edges.

APPENDIX A
NONCONVEX EXAMPLE

Given $v_i \sim \mathbf{N}_2(\mathbf{0}, \mathbf{P}_i)$, $i = 0, 1, 2, 3$, where

$$\mathbf{P}_0 = \begin{pmatrix} 6 & 0 \\ 0 & 0.1 \end{pmatrix}, \mathbf{P}_1 = \begin{pmatrix} 1.5 & -2.5 \\ -2.5 & 4.5 \end{pmatrix},$$
$$\mathbf{P}_2 = \begin{pmatrix} 1.5 & 2.5 \\ 2.5 & 4.5 \end{pmatrix}, \mathbf{P}_3 = \begin{pmatrix} 1 & 0 \\ 0 & 6 \end{pmatrix}.$$

The objective function $\mathcal{E}((w_1, w_2, 1 - w_1 - w_2))$ is shown in Fig. 1. In this case, all covariance matrix are pretty ill conditioned. In common cases, $\mathcal{E}(\mathbf{w})$ is most likely convex on the unit simplex.

REFERENCES

[1] M. Agueh and G. Carlier, "Barycenters in the Wasserstein space," *SIAM Journal on Mathematical Analysis*, vol. 43, no. 2, pp. 904–924, 2011.

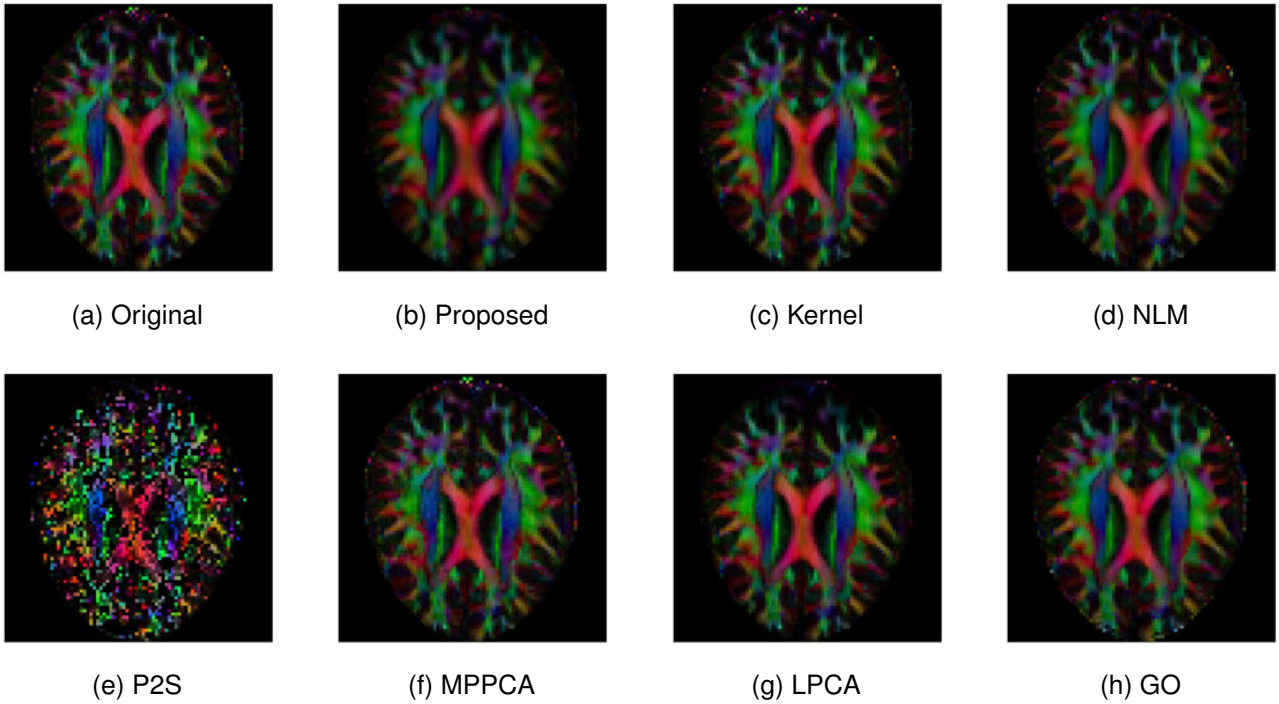
[2] C. Villani, *Topics in optimal transportation*. American Mathematical Society, 2003, no. 58.

[3] M. Cuturi and A. Doucet, "Fast computation of Wasserstein barycenters," in *International Conference on Machine Learning*, 2014, pp. 685–693.

[4] N. Bonneel, G. Peyré, and M. Cuturi, "Wasserstein barycentric coordinates: histogram regression using optimal transport." *ACM Transactions on Graphics*, vol. 35, no. 4, pp. 71–1, 2016.

TABLE I: Smoothing results at different noise levels.

	RMSE (l_2)					RMSE (W_2)					PSNR (dB)				
$1/\sigma^2$	5	10	20	30	40	5	10	20	30	40	5	10	20	30	40
Noisy DTI	2.05	1.05	0.53	0.35	0.26	0.54	0.28	0.13	0.08	0.06	15.7	21.8	27.7	31.3	33.8
Proposed	**0.30**	**0.14**	**0.07**	**0.05**	**0.03**	**0.17**	**0.08**	**0.04**	**0.03**	**0.03**	**23.0**	**29.1**	**35.2**	**38.8**	**41.3**
Kernel (h=3/4)	0.56	0.50	0.48	0.48	0.48	0.23	0.19	0.17	0.17	0.17	19.2	20.7	21.2	21.2	21.2
NLM	5.66	5.69	5.07	4.89	4.85	1.09	0.93	0.82	0.79	0.78	16.2	18.6	19.6	19.9	20.1
P2S	6.38	6.19	5.97	5.89	5.81	1.40	1.28	1.16	1.14	1.12	11.5	13.5	15.9	16.4	16.7
MPPCA	7.58	7.57	7.58	7.58	7.58	2.42	2.42	2.42	2.42	2.42	17.4	19.7	20.7	21.3	21.8
LPCA	9.32	7.35	5.49	5.32	4.55	1.07	0.88	0.72	0.67	0.65	14.6	18.3	20.9	21.9	22.1
GO	3.99	3.87	3.71	3.60	3.32	0.77	0.71	0.67	0.66	0.61	17.1	19.6	20.8	21.2	22.1



(a) Original (b) Proposed (c) Kernel (d) NLM

(e) P2S (f) MPPCA (g) LPCA (h) GO

Fig. 6: Color FA images derived from a slice of DTI ($104 \times 104 \times 72$ voxels) [44]. (a): The color FA image of original DTI data; (b): The color FA image of DTI data smoothed by the proposed method; (c): kernel method with bandwidth $= 0.5$; (d)-(h): Results of different methods. It is noticeable that the image smoothed by the proposed method is smoother.

[5] A. Gramfort, G. Peyré, and M. Cuturi, "Fast optimal transport averaging of neuroimaging data," in *International Conference on Information Processing in Medical Imaging*. Springer, 2015, pp. 261–272.

[6] A. N. Bishop, "Information fusion via the Wasserstein barycenter in the space of probability measures: Direct fusion of empirical measures and Gaussian fusion with unknown correlation," in *17th International Conference on Information Fusion (FUSION)*. IEEE, 2014, pp. 1–7.

[7] M. A. Schmitz, M. Heitz, N. Bonneel, F. Ngole, D. Coeurjolly, M. Cuturi, G. Peyré, and J.-L. Starck, "Wasserstein dictionary learning: Optimal transport-based unsupervised nonlinear dictionary learning," *SIAM Journal on Imaging Sciences*, vol. 11, no. 1, pp. 643–678, 2018.

[8] W. Yi, S. Li, B. Wang, R. Hoseinnezhad, and L. Kong, "Computationally efficient distributed multi-sensor fusion with multi-Bernoulli filter," *IEEE Transactions on Signal Processing*, vol. 68, pp. 241–256, 2019.

[9] X. Wang, G. Li, C. Quan, and P. K. Varshney, "Distributed detection of sparse stochastic signals with quantized measurements: The generalized Gaussian case," *IEEE Transactions on Signal Processing*, vol. 67, no. 18, pp. 4886–4898, 2019.

[10] J. Ye, P. Wu, J. Z. Wang, and J. Li, "Fast discrete distribution clustering using Wasserstein barycenter with sparse support," *IEEE Transactions on Signal Processing*, vol. 65, no. 9, pp. 2317–2332, 2017.

[11] M. Baum, P. K. Willett, and U. D. Hanebeck, "On Wasserstein barycenters and MMOSPA estimation," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1511–1515, 2015.

[12] E. Cazelles, A. Robert, and F. Tobar, "The wasserstein-fourier distance for stationary time series," *IEEE Transactions on Signal Processing*, vol. 69, pp. 709–721, 2020.

[13] H. Liu, X. Shen, Z. Wang, and F. Meng, "Randomized multimodel multiple hypothesis tracking," *SCIENCE CHINA-INFORMATION SCIENCES*, vol. 66, no. 9, 2023.

[14] R. J. McCann, "A convexity principle for interacting gases," *Advances in mathematics*, vol. 128, no. 1, pp. 153–179, 1997.

[15] L. Malagò, L. Montrucchio, and G. Pistone, "Wasserstein riemannian geometry of Gaussian densities," *Information Geometry*, vol. 1, no. 2,

pp. 137–179, 2018.
[16] R. Bhatia, T. Jain, and Y. Lim, "On the Bures–Wasserstein distance between positive definite matrices," *Expositiones Mathematicae*, vol. 37, no. 2, pp. 165–191, 2019.
[17] A. Kroshnin, N. Tupitsa, D. Dvinskikh, P. Dvurechensky, A. Gasnikov, and C. Uribe, "On the complexity of approximating Wasserstein barycenters," in *International conference on machine learning*. PMLR, 2019, pp. 3530–3540.
[18] P. C. Álvarez Esteban, E. del Barrio, J. Cuesta-Albertos, and C. Matrán, "A fixed-point approach to barycenters in Wasserstein space," *Journal of Mathematical Analysis and Applications*, vol. 441, no. 2, pp. 744 – 762, 2016.
[19] S. Claici, E. Chien, and J. Solomon, "Stochastic Wasserstein barycenters," in *International Conference on Machine Learning*. PMLR, 2018, pp. 999–1008.
[20] Y. Zemel, V. M. Panaretos *et al.*, "Fréchet means and procrustes analysis in Wasserstein space," *Bernoulli*, vol. 25, no. 2, pp. 932–976, 2019.
[21] S. Chewi, T. Maunu, P. Rigollet, and A. J. Stromme, "Gradient descent algorithms for Bures-Wasserstein barycenters," in *Conference on Learning Theory*. PMLR, 2020, pp. 1276–1304.
[22] D. Dvinskikh, "Stochastic approximation versus sample average approximation for Wasserstein barycenters," *Optimization Methods and Software*, pp. 1–33, 2021.
[23] T. Lin, N. Ho, X. Chen, M. Cuturi, and M. I. Jordan, "Fixed-support Wasserstein barycenters: Computational hardness and fast algorithm," *arXiv preprint arXiv:2002.04783*, 2020.
[24] J. Backhoff-Veraguas, J. Fontbona, G. Rios, and F. Tobar, "Bayesian learning with Wasserstein barycenters," *arXiv preprint arXiv:1805.10833*, 2018.
[25] D. S. Bernstein, *Matrix mathematics*. Princeton university press, 2009.
[26] V. M. Panaretos and Y. Zemel, "Statistical aspects of Wasserstein distances," *Annual review of statistics and its application*, vol. 6, pp. 405–431, 2019.
[27] M. Gelbrich, "On a formula for the l_2 Wasserstein metric between measures on euclidean and hilbert spaces," *Mathematische Nachrichten*, vol. 147, no. 1, pp. 185–203, 1990.
[28] S. Guminov, P. Dvurechensky, N. Tupitsa, and A. Gasnikov, "Accelerated alternating minimization, accelerated sinkhorn's algorithm and accelerated iterative bregman projections," *arXiv preprint arXiv:1906.03622*, 2019.
[29] D. Ge, H. Wang, Z. Xiong, and Y. Ye, "Interior-point methods strike back: Solving the wasserstein barycenter problem," *arXiv preprint arXiv:1905.12895*, 2019.
[30] L. Yang, J. Li, D. Sun, and K.-C. Toh, "A fast globally linearly convergent algorithm for the computation of Wasserstein barycenters." *Journal of Machine Learning Research*, vol. 22, no. 21, pp. 1–37, 2021.
[31] G. Lan, *First-order and Stochastic Optimization Methods for Machine Learning*. Springer Nature, 2020.
[32] D. P. Bertsekas, "Nonlinear programming," *Journal of the Operational Research Society*, vol. 48, no. 3, pp. 334–334, 1997.
[33] S. Mori, *Introduction to diffusion tensor imaging*. Elsevier, 2007.
[34] P. J. Basser, J. Mattiello, and D. LeBihan, "MR diffusion tensor spectroscopy and imaging," *Biophysical journal*, vol. 66, no. 1, pp. 259–267, 1994.
[35] O. Carmichael, J. Chen, D. Paul, and J. Peng, "Diffusion tensor smoothing through weighted karcher means," *Electronic journal of statistics*, vol. 7, p. 1913, 2013.
[36] R. Salvador, A. Peña, D. K. Menon, T. A. Carpenter, J. D. Pickard, and E. T. Bullmore, "Formal characterization and extension of the linearized diffusion tensor model," *Human brain mapping*, vol. 24, no. 2, pp. 144–155, 2005.
[37] P. J. Basser and C. Pierpaoli, "Microstructural and physiological features of tissues elucidated by quantitative-diffusion-tensor MRI," *Journal of magnetic resonance*, vol. 213, no. 2, pp. 560–570, 2011.
[38] S. Pajevic and C. Pierpaoli, "Color schemes to represent the orientation of anisotropic tissues from diffusion tensor data: application to white matter fiber tract mapping in the human brain," *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, vol. 42, no. 3, pp. 526–540, 1999.
[39] P. Coupé, J. V. Manjón, M. Robles, and D. L. Collins, "Adaptive multiresolution non-local means filter for three-dimensional magnetic resonance image denoising," *IET image Processing*, vol. 6, no. 5, pp. 558–568, 2012.
[40] S. Fadnavis, J. Batson, and E. Garyfallidis, "Patch2self: denoising diffusion MRI with self-supervised learning," *Advances in Neural Information Processing Systems 33*, 2020.
[41] J. Veraart, E. Fieremans, and D. S. Novikov, "Diffusion MRI noise mapping using random matrix theory," *Magnetic resonance in medicine*, vol. 76, no. 5, pp. 1582–1593, 2016.
[42] J. V. Manjón, P. Coupé, L. Concha, A. Buades, D. L. Collins, and M. Robles, "Diffusion weighted image denoising using overcomplete local pca," *PloS one*, vol. 8, no. 9, p. e73021, 2013.
[43] E. Kellner, B. Dhital, V. G. Kiselev, and M. Reisert, "Gibbs-ringing artifact removal based on local subvoxel-shifts," *Magnetic resonance in medicine*, vol. 76, no. 5, pp. 1574–1581, 2016.
[44] V. Romain, "Diffusion MRI measured at multiple b-values," 2015, http://hdl.handle.net/1773/33311.