A convolutional Wasserstein distance for tractography evaluation : complementary study to state-of-the-art measures

Thomas Durantel, Julie Coloigner, Olivier Commowick

A CONVOLUTIONAL WASSERSTEIN DISTANCE FOR TRACTOGRAPHY EVALUATION: COMPLEMENTARITY STUDY TO STATE-OF-THE-ART MEASURES

Thomas Durantel, Julie Coloigner and Olivier Commowick

Univ Rennes, INRIA, CNRS, INSERM, IRISA UMR 6074, Empenn ERL U-1228, F-35000, Rennes, France

ABSTRACT

Evaluation and comparison of tractograms are crucial and open problems that need to be solved in order to evaluate the false positives rate and variability of tractography algorithms. In this context, a lot of measures have been developed and are typically used to judge the quality of a tractogram. They however do not rely on the same quantities extracted from the tractograms to compare and may thus not evaluate the same aspects of tractogram quality. To evaluate this aspect and the measures redundancy or complementarity, we perform a quantitative analysis of the most common ones, both on simulated data and in real circumstances. We also propose a new evaluation measure based on optimal transport theory. We show that, when used in conjunction, these measures can provide a more in depth comparison of tractograms and thus a more complete evaluation.

1. INTRODUCTION

Diffusion MRI and fiber tractography are promising methods for the evaluation of the brain nervous fiber pathways. Since the emergence of these techniques at the end of the previous century [1], many new methods have appeared. The first ones were based on tensors, one of the simplest diffusion models [2]. Since then, numerous studies have shown the limitations of tensor-based tractography in complex fiber structures [3]. To overpass these limitations, new algorithms, using either other diffusion models (orientation distribution functions [4] or multi-tensors [5]) or new probabilistic tractography methods [6, 7], have been developed. These methods have allowed a better reconstruction of complex fiber bundles like the corticospinal tract (CST) or the corpus callosum (CC) [8], but still do not reach an acceptable false positive rate to be used in a clinical context [9].

The large number of existing techniques results in a lot of variability in the reconstruction of fiber bundles, which prevents, on one hand, clinicians from correctly interpreting the results, and on the other hand makes it difficult to evaluate and compare tractography algorithms [10]. Hence, many challenges [11, 12] have been organized to 1- study the sources of this variability, 2- provide a general framework allowing to more easily compare tractography algorithms. In those challenges, a few probabilistic methods [13, 14, 15] have shared the first place.

When organizing such challenges or evaluating an algorithm, the question of evaluation of tractograms becomes essential. While visual evaluations are crucial to quickly judge the quality of tractograms, they do not allow for an in-depth comparison and a quantitative performance analysis. Therefore, comparison being at the core of these challenges, measures have been developed to quantify and study differences between methods. First, measures such as the Dice score, its generalized version [16] or density correlation [10] analyze tractograms as an image derived from the fiber bundles being compared. Second, the anatomical plausibility can be studied with topographic regularity [17] or, as in TractoMeter, the number of valid connections (VC, the percentage of streamlines that correctly connect both ends of a fiber bundle) and the number of valid streamlines (VS, the percentage of streamlines that are part of VC, and that respect the general shape of the bundle).

As mentioned above, tractograms evaluation measures are a crucial and actual problem that lacks of a general answer. Multiple measures are available in the literature. However, as they work on different objects or quantities derived from tractograms, it is probable that one measure does not explain fully tractograms variations and errors, nor that they expose the same kind of errors. This has however never been explored in the literature. We thus propose first to study and compare the behavior of the most common of these measures when the reference and tested tractograms diverge from each other. This is done using simulated data by applying linear and non linear transformations, and also comparing results of different tractography algorithms in real cases. We also define and evaluate a new tractogram evaluation measure based on optimal transport and demonstrate its interest in combination to other measures.

2. METHODOLOGY

We first present four measures, based either on streamlines or on images derived from streamlines, typically used in recent works [11, 10] for evaluating tractogram quality: generalized Dice score, density correlation, valid connections and valid streamlines. We then introduce our new measure based on

optimal transport.

2.1. Generalized Dice score

The Dice and generalized Dice scores measure overlap between a reference and a test image. There are several generalizations of the binary Dice score. We have chosen the one proposed in [16]. Unlike the original Dice score, that assumes binary image inputs, the generalized Dice score handles weighted images. In the case of tractograms, these images are fiber density images (i.e. the weight of a voxel is the number of streamlines that pass through it). That property allows this measure to give more importance to the most dense regions.

2.2. Density correlation

The density correlation is also based on the use of fiber density images. It computes the cross-correlation coefficient between the reference density image and the test density image. It is given by:

$$\rho = \frac{\sum_i (A_i - \overline{A})(B_i - \overline{B})}{\sqrt{\sum_i (A_i - \overline{A})^2 \sum_i (B_i - \overline{B})^2}} \tag{1}$$

with $\rho \in [-1, 1]$, A_i and B_i the value of the corresponding image at voxel i, and $\overline{A}$ and $\overline{B}$ the average voxel value of reference and test density images.

One advantage of this measure over the generalized Dice score is that it can yield negative results, thus giving information about a negative correlation.

2.3. Valid Connections

In many cases, obtaining the ground truth of tractography is difficult. Some measures thus assume that we do not know the complete ground truth tractogram but rather only the start and end regions (start and end ROIs) at which the fibers should end. In this case, the evaluation measure can be defined by counting fibers of the evaluated tractogram that actually start and stop in these regions [11]. This percentage is named valid connections (VC) and was used in the TractoMeter challenge.

2.4. Valid Streamlines

Going one step further than VC, we can consider that, in addition to the start and end ROIs, we also know the envelope of the ground truth tractogram. If the full ground truth tractogram is available, this envelope can be extracted as a binary image highlighting voxels where at least one fiber passes through. The percentage of valid streamlines (VS) is then the percentage of all streamlines that are part of Valid Connections, and that do not leave the ground truth envelope. Hence the percentage of valid streamlines will always be lower than the percentage of valid connections.

2.5. Wasserstein distance

We propose in this paper a novel evaluation measure based on the computation of the Wasserstein distance, derived from optimal transport theory. First, we will introduce these concepts and then how to use them to evaluate tractograms.

Let μ_0 and μ_1 be, respectively, a source and a target distributions and $\pi(x, y)$ a transport plan that describes the amount of mass transported from μ_0 at location x to μ_1 at location y. The 2-Wasserstein distance, simply called Wasserstein distance in the following, defines a distance that can be used to measure the discrepancy between two distributions. It is given by Eq. 2:

$$W^2_{2,\gamma}(\mu_0, \mu_1) = \inf_{\pi \in \Pi} [\iint_{M \times M} d(x, y)^2 \pi(x, y)\, dx\, dy - \gamma H(\pi)] \tag{2}$$

with $d(x, y)$ the distance function between two elements $x \in M$ and $y \in M$, and M a compact, connected Riemannian manifold.

The last term, $\gamma H(\pi)$, is an entropic regularization term which is used to smooth the function and thus to provide a unique solution to the problem. Many methods have been devised in the literature to obtain this solution [18, 19]. The most common is the Sinkhorn algorithm [20]. However, this algorithm has two majors limitations. First, being an iterative algorithm, it can take a long time to converge. Second, when comparing two distributions of fibers (our tractograms), the distributions are made of individual Diracs at each point of each fiber. The number of distances to compute between the two distributions therefore becomes quickly overly large both in computation times and memory consumption (e.g. in our case, a normal fiber bundle can have 50000 points, so that 50000^2 distances need to be computed and stored to be computationally efficient).

Consequently, we have chosen to rely on a newly introduced and particularly efficient method: the convolutionnal Wasserstein distance algorithm proposed by Solomon et al. [21]. It relies on the Varadhan's formula to replace the kernel of the Sinkhorn algorithm by a Gaussian kernel from the heat equation when dealing with distributions represented as 3D images. Considering the tractograms as 3D density images as above for the generalized Dice score, we can perform the Sinkhorn iterations through 3D Gaussian kernel convolutions, which are both very computation and memory efficient since we can perform them directly on the images and thus do not have to store the distance matrix explicitly.

3. EXPERIMENTS

To compare the behaviour of the evaluated measures and to test the Wasserstein distance (WD), two sets of experiments, one on simulated transformations and another one on real data, were performed.

In the first set of experiments, the 5 evaluation measures (Dice, correlation, VC, VS and WD) were computed on 4 controlled transformations cases, often found at least partially in real cases. The data used here is the Cortico-Spinal Tract (CST) of 5 subjects of the Human connectome Project (HCP). This ground truth data was obtained by Wasserthal et al. [22] by performing fiber tracking using Mrtrix and multiple times filtered, first by ROIs, then by an expert. The 4 experiments were designed as follows:

Translation The reference track is translated along the X axis from 0 to 12.5 mm, in step of 0.0125 mm.

Rotation The track is rotated around the Z axis (roughly the CST main axis) and the centre of gravity of the bundle by an angle from 0 to 360 degrees, with increments of 1 degree

Deformation A stationary velocity field (SVF) [23] is generated by randomly generating Gaussian weight functions inside the reference fiber bundle mask. To each weight function is associated an individual translation in a random direction, with a random magnitude. The SVF is the weighted combination of these translations using the Gaussian weight functions. In our case, we generated for each deformation from one to 75 Gaussian weight functions, with a random translation between 0 and 5 mm. This SVF is then integrated and applied to the tractograms. To characterize the global amplitude of each generated deformation, we compute the average vector norm at each voxel, inside the bundle envelope.

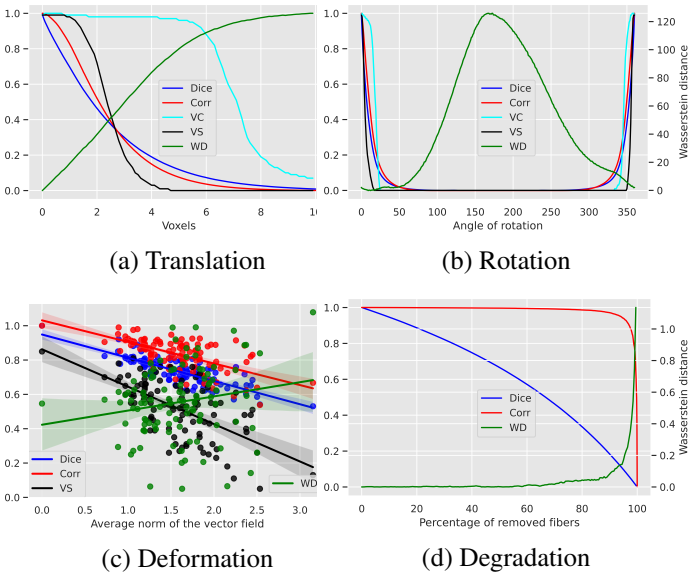
Degradation The track is degraded by randomly removing fibers from 0 to 100 percent of the reference fiber bundle

For each of these simulated datasets, cross-correlation coefficients were further computed between each pair of measure curves computed. This allowed to quantify the complementarity between these measures (the smaller the correlation the more complementary information).

The second set of experiments is based on data from the HCP and more particularly on 3 bundles, the CST, the Optic radiation (OR) and the first Superior longitudinal fascicle (SLF) tracked with the parallel transport frame algorithm [15], the Mrtrix3 iFOD2 [13] and the DIPY Probabilistic PFT algorithm [14]. For the 3 algorithms, only default parameters were used and 10000 streamlines were generated by running a whole brain tractography, then by filtering the results with the start and end ROIs, segmented with TractSeg. The evaluation was then performed between the ground truth tracks given by [22] and the resulting tractograms. For that experiment, we may note that the evaluation framework forces VC to be at 100%, thus only the other measures were computed. The purpose of these experiments is not, here, to compare these methods but to study the behavior of our measures in comparison with the state of the art methods.

4. RESULTS

We present in this section representative results obtained for the two sets of experiments.



(a) Translation (b) Rotation

(c) Deformation (d) Degradation

Fig. 1. Results for simulated data.

For translation, Fig. 1.a shows that all measures follow the same behavior, except for VC that seems to provide different information. All combinations have indeed a correlation coefficient above 0.95, except for $\rho(Corr, VC) = 0.60$, $\rho(Dice, VC) = 0.63$, $\rho(WD, VC) = 0.70$ and $\rho(VC, VS) = 0.43$. We have a maximum of 0.99 for $\rho(Corr, Dice)$.

For rotation, Fig. 1.b shows the same result as the translation, even more pronounced. We found $\rho(WD, Corr) = 0.41$, $\rho(WD, Dice) = 0.41$, $\rho(WD, VC) = 0.37$ and, $\rho(WD, VS) = 0.23$. All other are above 0.90 and the maximum is $\rho(Corr, Dice) = 0.99$.

For the deformation (Fig. 1.c), linear regression (plus a 95% confidence interval, in transparent on the plot) were computed and plotted to study the correlation between each measure against deformation. VC were not plotted because the major part of the deformation did not occur on the ending parts of the streamlines. The R^2 of each regression are as follows: 0.64 for Dice score, 0.40 for the density correlation, 0.38 for VS and 0.03 for WD. The regression plots show that, the Density correlation, the Dice score and VS percentage (red, blue and black) appear to follow the expected trend: they decrease when the average norm increases. However, the relationship between these measures and the average norm may not be linear and the x-axis is probably to simplistic to really conclude here without further experiments.

For degradation (Fig. 1.d), VC and VS, which by nature cannot measure a degradation (as the overall shape and the endings of the bundle are not affected), are not computed. Corr and WD are almost identical (except for the curves inversion). Dice follows a different trend, possibly leading to more information. In addition, $\rho(WD, Corr) = 0.51$ and

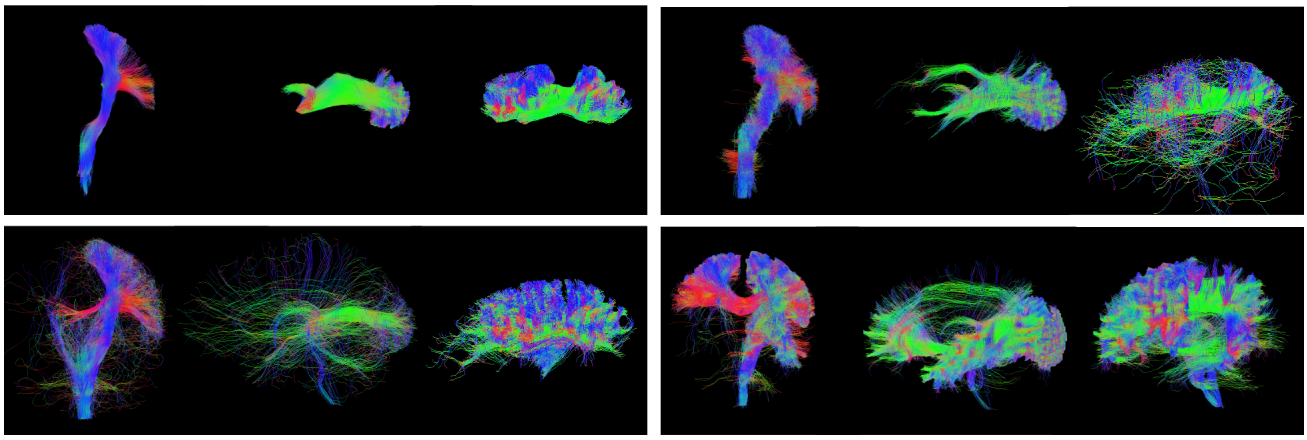


Fig. 2. Resulting tractograms for each algorithm. From left to right and top to bottom : the ground truth, Mrtrix IFOD2, Trekker and Dipy PFT. And from left to right in each case : the CST, the OR and the SLF

$\rho(WD, Dice) = 0.48$. All other results are under 0.10. The minimum is $\rho(Corr, Dice) = 0.002$.

Results for the real data experiment are shown in Table 1. Fig. 2 presents the resulting tractograms on which measurements were made. We can see that, in first analysis, the measures describe well the overall quality of the tractograms.

	Dice	Dens. corr.	WD	VS (%)
CST	0.23/0.18/0.53	0.47/0.36/0.54	37/74/72	42/43/78
OR	0.26/0.22/0.08	0.43/0.34/0.37	48/24/99	45/33/33
SLF	0.29/0.15/0.04	0.55/0.20/0.38	27/16/74	50/38/24

Table 1. Results for real data experiment. Red text denotes results for IFOD2, blue for Trekker, black for Dipy PFT.

5. DISCUSSION AND CONCLUSION

The cross correlation results for rotation and translation show that, for these transformations, the Dice score and the density correlation do not give complementary information. For translation, VC appears in each combination with the lowest coefficient, meaning that for this type of transformation, streamlines based measures can bring complementary information to image based measures. In the rotation case, the Wasserstein distance is the least correlated with the others, thus, that gives more information on tractogram quality in complement to other measures. The R^2 of the linear regressions for deformation plots show that these measures have difficulties to evaluate properly deformations in tractograms. However, this result should be taken with caution since, in this experiment, we tried to reflect local deformations with an average global information which may not be enough to capture the deformations variations. Designing more local evaluation of the deformation magnitude would be beneficial. Finally, for the degradation, all coefficients are relatively low, in particular the density correlation and the Dice score that are almost uncorrelated showing their complementarity.

For real data, results show that, globally, VC, Dice and the correlation follow the same trend. The Wasserstein distance seems to give slightly different information. A more in-depth real data study will bring more information on that aspect.

As shown above, the Wasserstein distance can, in certain cases, be used to deliver information that the other measures cannot. However, we would like to point out some of the actual limits of this distance and therefore some future works that could improve this measure. First, the measure computes a distance between all points of the two tractograms instead of matching fibers, which is not topologically right. A way to transport fibers from a tractogram to another should be looked into. Second, special attention must be given to the quantity of mass transported. Indeed, balanced (the same amount of mass on both sides) and unbalanced (different amounts of mass) optimal transport are two different problems, which require different solutions. In this work, balanced optimal transport is forced by normalising the tractograms densities. This can lead to unexpected situations, e.g. with the degradation experiment, where, for a high percentage of degradation, we force an entire tractogram to be transported onto a few fibers.

In conclusion, we have proposed, in this work, a comparative analysis, both on synthetic data (translation, rotation, deformation, degradation) and in real circumstance, of the behavior of a few typically used tractograms evaluation measures: the Dice score, the density correlation, the percentage of valid connections and valid streamlines. We then proposed a new measure, based on optimal transport and compared its behavior against the other measures. We show that, in most cases, it can be useful to use multiple measures in conjunction since they do not correlate and thus bring different kinds of information. We also show that, although still being in development, our new Wasserstein measure can be used to compare tractograms, bringing additional information.

6. REFERENCES

[1] P.J. Basser et al., "In vivo fiber tractography using DT-MRI data," *Magn Reson Med.*, vol. 44, pp. 625–32, 2000.

[2] S. Mori et al., "Fiber tracking: principles and strategies – a technical review," *NMR in Biomedicine*, vol. 15, no. 7-8, pp. 468–480, 2002.

[3] Y. Gao et al., "Validation of DTI tractography-based measures of primary motor area connectivity in the squirrel monkey brain," *PloS one*, vol. 8, 2013.

[4] M. Descoteaux et al., "Regularized, fast, and robust analytical q-ball imaging," *Magn Reson Med.*, vol. 58, pp. 497–510, 2007.

[5] D.S. Tuch et al., "High angular resolution diffusion imaging reveals intravoxel white matter fiber heterogeneity," *Magn Reson Med.*, vol. 48(4), pp. 577–82, 2002.

[6] A. Stamm et al., "Adaptive multi-modal particle filtering for probabilistic white matter tractography," *Information Processing in Medical Imaging.*, pp. 594–606, 2013.

[7] D. K. Jones, "Tractography gone wild: Probabilistic fibre tracking using the wild bootstrap with diffusion tensor MRI," *IEEE Transactions on Medical Imaging*, vol. 27, no. 9, pp. 1268–1274, 2008.

[8] G. Girard et al., "On the cortical connectivity in the macaque brain: A comparison of diffusion tractography and histological tracing data," *NeuroImage*, vol. 221, pp. 117201, 2020.

[9] K. Maier-Hein et al., "The challenge of mapping the human connectome based on diffusion tractography," *Nat Commun 8*, vol. 1349, 2017.

[10] K. G. Schilling et al., "Tractography dissection variability: What happens when 42 groups dissect 14 white matter bundles on the same dataset?," *NeuroImage*, vol. 243, pp. 118502, 2021.

[11] Marc-Alexandre Côté et al., "Tractometer: Towards validation of tractography pipelines," *Medical Image Analysis*, vol. 17, no. 7, pp. 844–857, 2013, Special Issue on the 2012 Conference MICCAI.

[12] Maffei et al, "The irontract challenge: Validation and optimal tractography methods for the hcp diffusion acquisition scheme," *Proc. Int. Soc. Mag. Res. Med.*, p. 849, 2020.

[13] J.-D. Tournier et al., "Improved probabilistic streamlines tractography by 2nd order integration over fibre orientation distributions," *Proceedings of the ISMRM*, vol. 1670, 2010.

[14] R. E. Smith et al., "Anatomically-constrained tractography: Improved diffusion MRI streamlines tractography through effective use of anatomical information," *NeuroImage*, vol. 62, no. 3, pp. 1924–1938, 2012.

[15] Dogu Baran Aydogan and Yonggang Shi, "Parallel transport tractography," *IEEE Transactions on Medical Imaging*, vol. 40, no. 2, pp. 635–647, 2021.

[16] W.R. Crum et al., "Generalized overlap measures for evaluation and validation in medical image analysis," *IEEE Transactions on Medical Imaging*, vol. 25, no. 11, pp. 1451–1461, 2006.

[17] Dogu Baran Aydogan and Yonggang Shi, "Tracking and validation techniques for topographically organized tractography," *NeuroImage*, vol. 181, pp. 64–84, 2018.

[18] N. Bonneel and D. Coeurjolly, "Spot: Sliced partial optimal transport," *ACM Trans. Graph.*, vol. 38, no. 4, July 2019.

[19] A. Figalli, "The optimal partial transport problem," *Arch Rational Mech Anal*, vol. 195, pp. 533–560, 2010.

[20] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in *Advances in Neural Information Processing Systems*, 2013, vol. 26.

[21] J. Solomon et al., "Convolutional wasserstein distances: Efficient optimal transportation on geometric domains," *ACM Trans. Graph.*, vol. 34, no. 4, July 2015.

[22] J. Wasserthal et al., "Tractseg - fast and accurate white matter tract segmentation," *NeuroImage*, vol. 183, pp. 239–253, 2018.

[23] V. Arsigny et al., "A Log-Euclidean framework for statistics on diffeomorphisms," in *MICCAI*, 2006, vol. 4190 of *LNCS*, pp. 924–931.

7. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available in open access by the HCP (`http://www.humanconnectomeproject.org/`). Ethical approval was not required as confirmed by the license attached with the open access data.