SB3C2023
Summer Biomechanics, Bioengineering and Biotransport Conference
June 4-8, Vail, CO, USA

ENHANCING CORRUPT CARDIOVASCULAR FLOW DATA WITH MACHINE LEARNING

Hunor Csala (1,2), Amirhossein Arzani (1,2)

(1) Department of Mechanical Engineering, University of Utah, Salt Lake City, UT, USA
(2) Scientific Computing and Imaging Institute, University of Utah, Salt Lake City, UT, USA

INTRODUCTION

Obtaining clean, high resolution velocity measurements of blood flow inside small arteries such as cerebral vasculature is challenging, as the existing experimental techniques have several limitations. Time-resolved three-dimensional phase contrast magnetic resonance imaging (4D flow MRI) is a popular approach in research settings, however it is constrained by low spatio-temporal resolution, noise, and other artifacts [1]. Particle image velocimetry (PIV) often considered the gold-standard in experimental fluid dynamics also suffers from some limitations. Therefore, handling corrupt blood flow data is key challenge towards developing more accurate and robust cardiovascular flow models. There are well developed algorithms in the machine learning community that can tackle similar issues, such as data imputation, denoising or outlier detection. Existing methods have been less frequently used and leveraged for complex real-world fluid flow problems [2], such as cardiovascular flows [3]. Specifically, we do not understand which one of these approaches commonly used in the machine learning community perform better for hemodynamics data. This study investigates and compares several techniques for filling in missing values and denoising unsteady blood flow data in an image-based 3D intracranial aneurysm model.

METHODS

Voxel-based data mimicking some of 4D flow MRI data features was created from computational fluid dynamics (CFD) simulation results. Pulsatile blood flow inside an internal carotid artery (ICA) aneurysm was simulated using SimVascular, a finite element numerical solver. The Reynolds number was 555, based on the maximum systolic inlet velocity and inlet diameter. A population averaged inlet waveform was used, 1000 snapshots were saved through one cardiac cycle. The unstructured mesh consisted of 6.6M elements, which was resampled to a voxelized grid with a uniform spatial resolution of 0.5 mm, resulting in 27000 voxels. Two types of data corruption were investigated here, missing data and Gaussian noise. These were artificially added to the voxelized data. Voxels were randomly removed in space and time for the missing data case and Gaussian noise was added for the noisy case. The process of creating corrupt synthetic voxel-based data is illustrated in Figure 1. For the missing data case, the fraction of missing data was varied between 10% and 90%. For the denoising case, the fraction of noisy data was varied between 10% and 70%. The standard deviation of the added noise was 10% of the maximum velocity value.

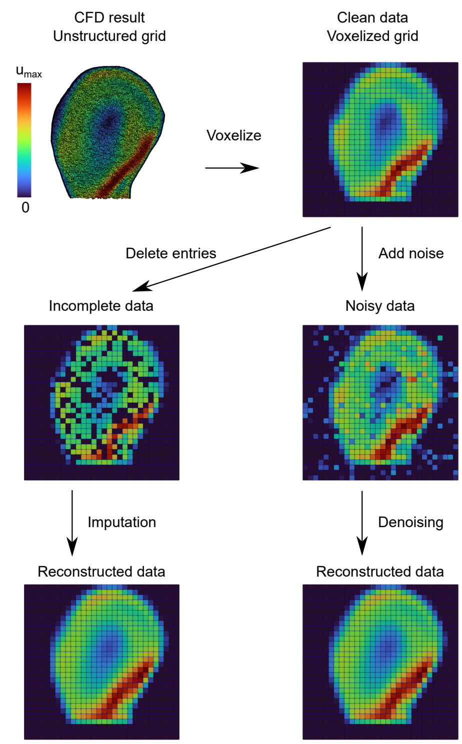

Figure 1: Process of creating corrupt voxel-based data from CFD results.

Imputation is the process of filling in missing data based on some statistics of the observable dataset. Four methods were used for data imputation: probabilistic principal component analysis (PPCA), iterative singular value decomposition (itSVD), softImpute, and an autoencoder. PPCA is a probabilistic version of PCA using the expectation-maximization (EM) algorithm. itSVD and softImpute are two techniques based on iterative low-rank SVD decompositions and soft-thresholding. The autoencoder approach is a fully-connected deep neural network that has a bottle-neck layer in the middle, whose dimensionality is much smaller than the input and output, therefore the network learns a low-dimensional embedding of the data.

Three algorithms were used for denoising: robust principal component analysis (RPCA), Noise2Noise autoencoder (N2N), and a conventional denoising autoencoder (DAE). RPCA tries to separate the data into a low-rank matrix containing the noise-free data and a sparse noise matrix. N2N is a neural network based denoising algorithm that does not require clean training data. That is, the input and the target data are both noisy data representing two samples from the same noise distribution. The DAE approach does denoising in a supervised fashion. It was trained with 7 simulations with different mean inlet velocities (Reynolds number ranging between 380 and 650, based on the systolic maximum inlet velocity) on the same geometry, then tested on unseen velocity data ($\mathrm{Re}_{max} = 555$). The noise and missing data characteristics (fraction of data missing or noisy and noise standard deviation) were the same during training and testing. It is important to note that the matrix completion (itSVD, softImpute) and PCA-based methods (PPCA, RPCA) do not require training data.

RESULTS

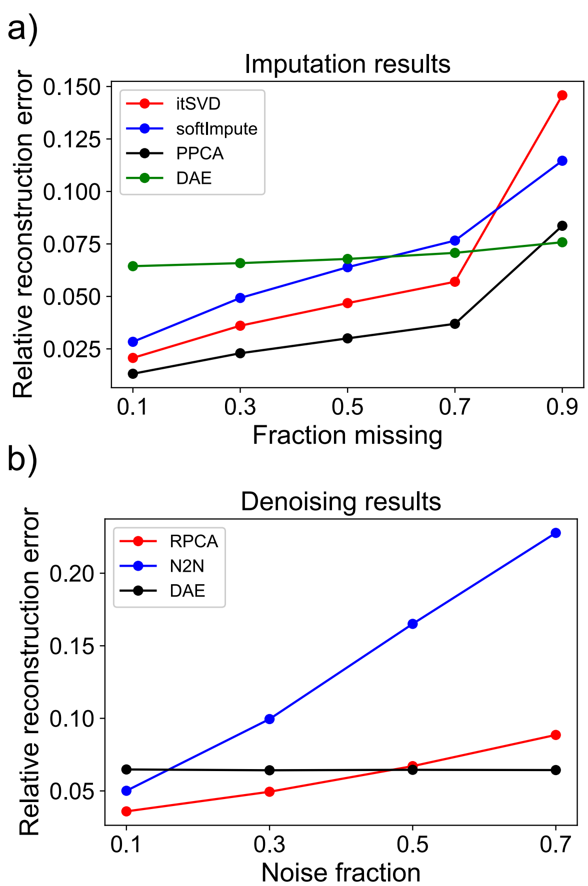

Figure 2: a) Imputation relative reconstruction error with different percentage of data missing. b) Denoising relative reconstruction error with different percentage of data being noisy.

Imputation and denoising results are shown in Figure 2, top and bottom panel, respectively. The relative reconstruction error ε is defined as

$$\epsilon = \frac{\|\mathbf{X} - \mathbf{X}_{rec}\|}{\|\mathbf{X}\|} \,, \tag{1}$$

where $\mathbf{X}$ is the clean ground-truth data, $\mathbf{X}_{rec}$ is the filled in or denoised data, and the Frobenius norm is used. In case of imputation, the two matrix completion methods (itSVD and softImpute) have similar results. PPCA has the lowest error in almost all cases. As the percentage of missing data grows the error increases, and a sharp rise can be seen at 90% missing data for these three methods. The autoencoder is much less sensitive to the amount of data missing, probably due to the nature of supervised learning, and its error is only slightly increasing as the fraction of missing data increases. Overall, all methods perform well and can fill in the missing entries in the aneurysm velocity data, as the relative reconstruction error is below 10% for most cases, and even for the highest missing fraction of 0.9 the error is less than 15% for all four techniques. This suggests that blood flow inside an aneurysm can be described well by an intrinsic low-dimensional model that can be inferred from a few data points.

For denoising, N2N has clearly inferior performance to the other methods. RPCA achieves good performance in the low-noise setting, while the DAE outperforms RPCA in the high-noise scenario. Again, the autoencoder errors does not seem to be depending too much on the fraction of noisy data. RPCA and DAE achieve an error smaller than 10% for all noise fractions, while for N2N the error exceeds 20% at 70% of noisy data.

DISCUSSION

We investigated several machine learning algorithms for handling corrupt blood flow data inside an ICA aneurysm. Matrix completion, PCA-based, and neural network-based techniques proved to be able to enhance blood flow data corrupted with randomly missing voxel entries and added Gaussian noise. Matrix factorization methods seem to be more useful in the low-corruption case, while for highly corrupted datasets neural networks perform better. Overall, these machine learning algorithms have the potential to significantly improve hemodynamic data quality and enable the creation of robust cardiovascular models.

Our findings suggest that neural networks are more valuable when dealing with highly corrupted datasets with available data for training. On the other hand, matrix completion and PCA-based methods can achieve excellent results with mildly corrupted data, without the need for clean training data. However, there is a serious limitation to the matrix completion and PCA methods. These methods assume that the missing data is random in space and time, and the noise is Gaussian noise. Without these assumptions, these methods could fail. On the other hand, if there is enough training data available, the autoencoder based methods can achieve good results, no matter the pattern of missing data or the characteristic of the noise. Obviously it is desired that the training and testing data have the same type of corruption. Creating models that can handle multiple noise or missing data characteristics is a future challenge that needs to be investigated.

Our future work includes extending this setup to real experimental data. Additionally, we are working on other types of data corruptions such as aliasing in 4D flow MRI, which could be treated using outlier detection algorithms.

ACKNOWLEDGEMENTS

This work was supported by NSF grant No. 2246916.

REFERENCES

[1] Fathi MF et al. *Computer Methods and Programs in Biomedicine* 197 (2020). ISSN: 0169-2607. DOI: `https://doi.org/10.1016/j.cmpb.2020.105729`.

[2] Scherl I et al. *Phys. Rev. Fluids* 5 (5 2020). DOI: `10.1103/PhysRevFluids.5.054401`.

[3] Arzani A et al. *Journal of the Royal Society Interface* 18.175 (2021).