PREDICTING THE PRINCIPAL COMPONENTS OF CORTICAL MORPHOLOGICAL VARIABLES

MANUSCRIPT

V. B. B. Mello[1], **F. H. de Moraes**[1,2], **B. Mota**[1]
[1] Instituto de Física, Universidade Federal do Rio de Janeiro (UFRJ), Rio de Janeiro, Brazil
[2] Brain Connectivity Unit, Instituto D'Or de Pesquisa e Ensino (IDOR), Rio de Janeiro, Brazil
Contact: vbraga@pos.if.ufrj.br

July 7, 2022

ABSTRACT

The generating mechanism for the gyrification of the mammalian cerebral cortex remains a central open question in neuroscience. Although many models have been proposed over the years, very few were able to provide empirically testable predictions. In this paper, we assume a model in which the cortex folds for all species of mammals according to a simple mechanism of effective free energy minimization of a growing self-avoiding surface subjected to inhomogeneous bulk stresses to derive a new set of morphological variables to express cortical morphology. In terms of these new variables, we seek to understand the variance present in two morphometric datasets: a human MRI harmonized multi-site dataset comprised by 3324 healthy controls (CTL) from 4 to 96 years old and a collection of different mammalian cortices with morphological measurements extracted manually. This is done using a standard Principal Component Analysis (PCA) of the cortical morphometric space. We prove there is a remarkable coincidence (typically less than $8°$) between the resulting principal components vectors in each datasets and the directions corresponding to the new variables. This shows that the new, theoretically-derived variables are a set of natural and independent morphometrics with which to express cortical shape and size.

Keywords: PCA, cortical folding model, cortical folding universal rule.

Introduction

Why the cerebral cortex folds in some mammals is a fundamental question in neuroscience that still lacks a broadly accepted mechanism. It is reasonable to think about the cortical folding as a characteristic fixed by natural selection because it allows a dramatic increases in cortical size and complexity, while preserving the cranial size, axonal elongation and metabolic cost [1,2]. However, the evolutionary argument is not sufficient to unravel the full process that results in the surface folding. Empirical evidence points towards an integrative explanation [3], unifying bio-mechanical, genetic and cellular signalling mechanism.

From a purely biomechanical point of view, there is a variety of models proposed to explain the physics of cortical folding: buckling from a constrained expansion [4,5], warping due to differential tangential expansion [6] or emerging from axonal elongation [7]. Despite the qualitative compatibility with the cortical morphology, all these theories failed to provide a set of empirical testable predictions to further investigate their validity.

Following the biomechanical approach, Mota & Herculano-Houzel [8] proposed a folding model hypothesizing that the cortex is a crumpled elastic surface in thermal equilibrium (like a crumpled paper ball) [9]. The model is based on the physics of minimization of effective free energy of a growing self-avoiding surface subject to inhomogeneous bulk stresses. It predicts a power-law relationship between cortical thickness (T), exposed (A_E), and total area (A_T) (Equation 1)

$$T^{\frac{1}{2}} A_T \;=\; k A_E^{\alpha} \tag{1}$$

for which the index α is a universal constant with a theoretical value of 1.25 and k is a dimensionless constant not specified by the model. The most immediate consequence of this model is that the usual set of geometrical variables $\{T^2, A_T, A_E\}$ used to describe the cortex should not be independent, but rather related in a highly restrictive equation. Representing these variables in the $\{\log T^2, \log A_T, \log A_E\}$ space one can define

$$K \equiv \log k \;=\; \underbrace{\frac{1}{4}\log T^2 + \log A_T - \frac{5}{4}\log A_E}_{a(x-x_0)+b(y-y_0)+c(z-z_0)\,=\,0} \tag{2}$$

In this log-space, Equation (1) can be re-interpreted as an assertion that actual cortices must lie at or near a $K =$ constant plane, what is called here the theoretical cortical plane. The cortical morphology that in principle could be compatible with any set of values $\{T^2, A_T, A_E\}$, is thus restricted to a plane perpendicular to the vector $\vec{K} \;=\; (1/4, 1, -5/4)$ in the log-space. Taking a step further, as Equation 1 is a a product of power-laws relating intrinsic and extrinsic measures of cortical areas, the model also suggests that the cortex is self-similar over some range of length scales, i.e, an approximation of a real fractal.

This simple statistical physics model was validated by experimental data. Firstly, it was found empirically that for 63 different mammalian species cortical folding scales universally according equation 1 with $\alpha_{\text{exp}} \;=\; 1.31 \pm 0.01$ (when considering a heterogeneous, i.e, multiple acquisition protocols dataset) [8]. The work was extended in [10], where it is shown that the universal law also holds within a single species dataset, but also provide an evidence that a deviation from the universal rule could be used to identify diseases. Then, in [11], it is shown that the white matter volume and white/gray matter ratio in mammalian species [12] are a consequence of the universal scaling of cortical folding. The first evidence of the self-similarity of the cortical surface was observed in [13] with an experimental verification of the universal scaling law across lobes of the same cortex. More recently, the theoretical framework of the model was expanded by [14], which derives a more natural orthogonal set of independent morphological variables for the same morphometric space defined by $\{\log T^2, \log A_T, \log A_E\}$. We start with the nearly invariant K as defined above; since k is dimensionless K is not affected by an isometric scaling $\log T^2 \to \log \lambda T^2$, $\log A_T \to \log \lambda A_T$ and $\log A_E \to \log \lambda A_E$, in which λ is a positive real number; thus one can define both a direction $\vec{I} \;=\; \{1, 1, 1\}$ that correspond to this isometric scaling, and a corresponding new variable $I = \log T^2 + \log A_T + \log A_E$ that captures all information about cortical size. The third and last vector to both $\vec{K}$ and $\vec{I}$ ($\vec{S} \;=\; \vec{I} \times \vec{K}$), and is associated with the variable $S \;=\; -\frac{9}{4}\log T^2 + \frac{3}{2}\log A_T + \frac{3}{4}\log A_E$ that captures most of the variation in shape across actual cortices.

These new variables thus naturally decompose morphology in terms of size, variant and invariant shape components, in a way that was shown to better discriminate between both typical and atypical structural changes, like aging and diseases [14]. Furthermore, we know the original variable set cannot be correlationally independent over the datasets studied, because of the constraint imposed by Equation (1). The question then becomes, what about the new set? If they can be shown to be independent, their usefulness as discriminators would be greatly enhanced, and this would furthermore hint at the existence of the underlying independent degrees of freedom with which evolution can shape a cerebral cortex.

In passing, we emphasize that although this new set of variables were derived from a specific model, their use to express and analyze morphology is model-agnostic, insofar as their distributional properties such as the near invariance of K and their statistical independence (shown in the present work) can be regarded as empirical facts. Obversely, however, these facts do provide a stringent test for other putative models for cortical folding mechanisms.

To test the independence between the new variable set $\{K, S, I\}$ proposed in [14], we performed standard Principal Component Analysis (PCA) to obtain the independent components of variance in the $\{\log T^2, \log A_t, \log A_e\}$ space, for two different cases: across healthy human subjects and across multiple species. These vectors were then directly compared to the theoretical vectors $\vec{K}$, $\vec{S}$ and $\vec{I}$ associated with the new variable set. Finally, a simulated dataset was generated according equation 1 to address the sensitiveness of the PCA technique to the dataset's true principal directions.

Materials and Methods

The morphological data used in this analysis were all previously published. Human participants and their respective MRI-derived metrics were either published by Wang et al. [10, 13] or from open-access databases, as AHEAD [15] and

AOMIC (PIOP01 and PIOP02) [16]. The data of different mammalian cortex from multiple clades were all provided by [8].

For the human MRI dataset, the total number of subjects is 3324 healthy controls (CTL) from 4 to 96 years old. The structural images from AHEAD, AOMIC ID1000, PIOP1, and PIOP2 were processed in FreeSurfer v6.0.0 [17] without manual intervention at the surfaces [18]. The FreeSurfer localGI pipeline generates the external surface and calculates each vertex's local Gyrification Index (localGI) [19]. Values of Average Cortical Thickness, Total Area, Exposed Area were extracted with Cortical Folding Analysis Tool [20]. We defined as ROI the whole hemisphere, frontal, temporal, occipital, and lateral lobes (based on FreeSurfer definition of lobes). The lobes' area measurements were corrected by their integrated Gaussian Curvature, removing the partition size effect and allowing a direct comparison between lobes and hemisphere's cortical folding. In order to perform a joint analysis, the morphological variables were harmonized following the prescription given in [21]. The data harmonization procedure is based on the time evolution of the basic morphological variables T, A_T or A_E, modeled as an exponential decay with the age t_{age}. A joint linear fit was performed in the log-linear scale, maintaining the same angular coefficient among all different samples but different linear coefficient. The linear coefficients were then subtracted to harmonize all datasets.

Across healthy aging, the value of the morphological variables naturally changes. This variability can significantly impact the PCA results and the posterior interpretation. To avoid this problem, we used two techniques: (i) restrict the analysis to the age range between 20-40 years (adulthood) as it would be the best representation of a stationary fully developed cortex; (ii) take advantage from the harmonization linear regression, removing the age dependence using each subject age and the angular coefficient fitted, a procedure called here de-ageing. This allow to perform the analysis to the full age range (20-60 years) in which the model was experimentally validated [21]; In the supplementary material 2, one finds the replication of the analysis for all age ranges on the dataset (with and without the de-ageing).

Regarding the mammalian dataset from multiple clades, this work used the extracted morphological variables made available by the authors at [8]. This dataset also presents heterogeneity in acquisition, compiling Mota & Herculano-Houzel's group data (refered here as dataset A) with previously published cortical morphological measurements (dataset B). The details about each acquisition protocols can be found in the original publications. This variability can also significantly impact the PCA results due to systematic uncertainties from different experiments. To avoid this problem, we applied the analysis using dataset A and B separately. In the supplementary material 3, one finds the replication of the analysis for the datasets combined.

Data analysis

Principal Component Analysis

Principal Component Analysis (PCA) is a standard procedure in multivariate data analysis, usually used to dimensionality reduce a dataset [22]. With a direct analogy to the calculation of a rigid body principal axes of inertia, it creates a new set of uncorrelated variables, the principal components, solving an eigenvalue/eigenvector problem for the covariance matrix of a distribution or set of points. In this work, the PCA was performed numerically using Python SciPy library [23]. Across this article, the vectors $\hat{v} = (v_1, v_2, v_3)$, $\hat{u} = (u_1, u_2, u_3)$ and $\hat{w} = (w_1, w_2, w_3)$ represent the normalized principal components ordered by the size of explained variance.

Considering the $\{\log T^2, \log A_T, \log A_E\}$ space, the universal scaling law strongly suggests that the vector $\vec{K} = (1/2, 1, 5/4)$ (Equation 2), must be the principal component with the lowest explained variance. It is a near-invariant quantity obtained by isolating k in Equation 1 (it defines the theoretical cortical plane), and is associated with conserved viscoelastic properties of cerebral cortical matter (k is a function of these properties on the thermodynamic model proposed to explain the universal cortical folding [8, 9]). For humans, for whom in addition to a nearly invariant K one also finds that the average cortical thickness varies little [13], one would expect only small variations along the $\log T^2$ direction. Thus, the component with the most variance should lie along the $\vec{n}_1 = (0, 5/4, 1)$ direction. The remaining PC is directly found by the cross product $\vec{n}_2 = \vec{K} \times \vec{n}_1$. The normalized vectors that will be used in comparison with the PCA are denote here as $\hat{n}_1 = (0, 0.781, 0.625)$, $\hat{n}_2 = (-1, 0, 0)$, and $\hat{K} = (-0.154, -0.617, 0.771)$.

In terms of comparative neuroanatomy, there is substantial variation in cortical thickness and areas. Following the interpretation in [14], we predict that the principal components in this case are the normalized directions defined by the independent morphological variables K, S and I given by $\hat{I} = (0.577, 0.577, 0.577)$, $\hat{S} = (-0.802, 0.534, 0.267)$, and $\hat{K} = (-0.154, -0.617, 0.771)$. If the variables $\{K, S, I\}$ are actually a natural set of morphological variables, capturing different and orthogonal aspects of the cortex, their respective vectors should be the principal components in the $\{\log T^2, \log A_T, \log A_E\}$ space.

Statistical Significance of the PCA results

To better interpret the results above, we used Monte Carlo simulations as a tool to address the power of the analysis to recover the true principal components vectors and address the statistical significance of the results. The data was generated over $N_{\text{sim}} = 1000$ iterations using a Monte Carlo method to sample points near the theoretical cortical plane defined by equations 2. The number of points generated and their variance were chosen to match the available data on the human MRI dataset. In this simulated datasets, we applied the same PCA procedure used to analyze the real data, calculating the distribution for the angle θ_{diff} between the real principal component $\vec{K}$ and the component in the PCA with the least explained variance for each simulated dataset.

Data and code availability

All codes and processed dataset are available at `https://github.com/bragamello/metaBIO`. AHEAD and AOMIC (PIOP01 and PIOP02) data is available at `https://zenodo.org/record/5750619`.

Results

Performing the PCA in the humans harmonized dataset, one obtains the results summarized in Figure 1. Considering subjects in the age range between 20-40 years, the vectors found are $\hat{v} = (0.056, 0.704, 0.707)$, $\hat{u} = (-0.996, 0.089, -0.009)$, and $\hat{w} = (-0.070, -0.704, 0.707)$, deviating, respectively, $7.2°$, $6.4°$, and $7.9°$ from the normalized theoretical predictions $\hat{n}_1 = (0, 0.781, 0.625)$, $\hat{n}_2 = (-1, 0, 0)$, and $\hat{K} = (-0.154, -0.617, 0.771)$. Using the de-aged dataset, one finds the vectors $\hat{v} = (0.077, 0.703, 0.707)$, $\hat{u} = (-0.994, 0.110, -0.0004)$, and $\hat{w} = (-0.078, -0.703, 0.707)$, deviating, respectively, $7.9°$, $6.9°$, and $7.5°$ from the normalized theoretical predictions. These results do not change substantially when considering more restricted age range over the human lifespan. See appendix 2.

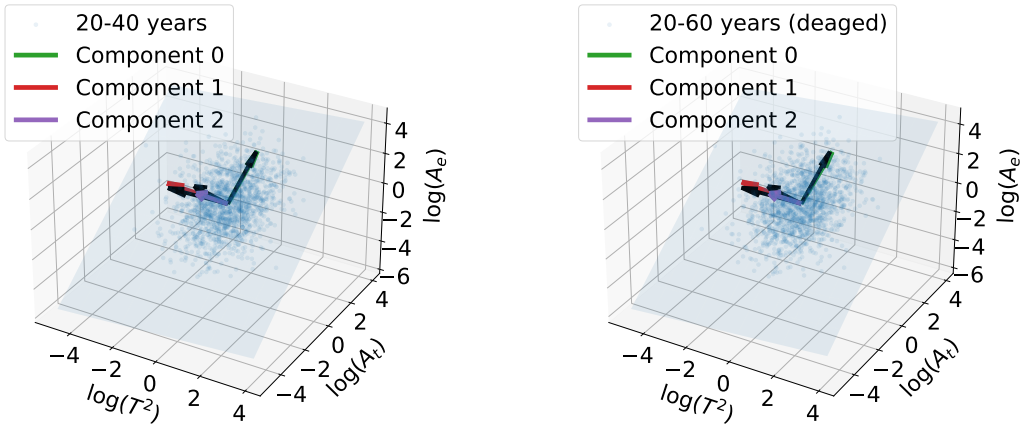

Figure 1: Cortical plane for the human MRI dataset and the vectors reconstructed by PCA. On the left panel, we considered only subjects in the age range 20-40 year. On the right panel the de-ageing procedure was applied allowing to analyze the full age range in which the model was experimentally validated. The results of the two approaches are consistent. The arrows in black represents the principal components predicted by the model. The angular difference between theoretical components and the experimental ones are $\theta_{\text{diff}} < 8°$. The results are remarkably in agreements with the theoretical predictions.

In terms of the comparative neuroanatomy dataset, the PCA result is shown in Figure 2. Considering dataset A, the vectors found are $\hat{v} = (0.490, 0.604, 0.628)$, $\hat{u} = (-0.863, 0.437, 0.253)$, and $\hat{w} = (-0.121, -0.666, 0.736)$, deviating, respectively, $6.0°$, $6.7°$, and $4.0°$ from the normalized theoretical predictions $\hat{I} = (0.577, 0.577, 0.577)$, $\hat{S} = (-0.802, 0.534, 0.267)$, and $\hat{K} = (-0.154, -0.617, 0.771)$. Considering dataset B, the vectors found are $\hat{v} = (0.499, 0.602, 0.623)$, $\hat{u} = (-0.859, 0.439, 0.264)$, and $\hat{w} = (-0.114, -0.666, 0.736)$, deviating, respectively, $5.4°$, $6.4°$, and $4.1°$ from the predictions. As expected, the $\hat{K}$ is very close to being the component with the least variance, reflecting the confining of data points at or close to the theoretical plane. The combination of both datasets is discussed in appendix 3 in which is shown that this PCA method can be used to detect subtle systematic differences

between different datasets, and underscores the need to properly harmonize said datasets before performing joint analyzes.

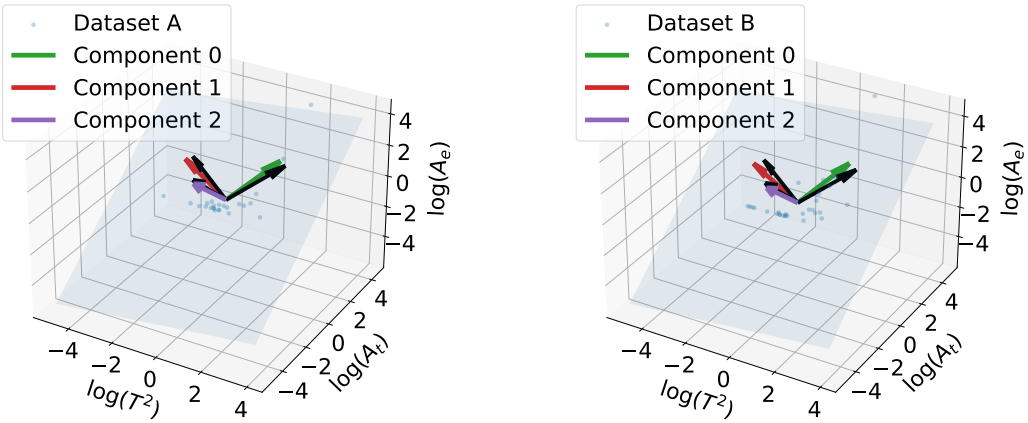

Figure 2: Cortical plane considering different mammalian species dataset and the vectors reconstructed by PCA. The arrows in black represents the principal components predicted by the model, the directions of the natural variables K, S and I. To avoid systematic uncertainties from different experiments we applied the analysis in dataset A (left panel) and B (right panel) separately. The angular difference between theoretical components and the experimental ones are $\theta_{\text{diff}} < 7°$. The results are remarkably in agreements with the predictions and consistent between distinct experiments, suggesting that these variables are indeed a natural mathematical language to describe the cortex, encapsulating orthogonal information about the morphology.

In order to address the power of the PCA to recover the true principal components vectors, we ran the analysis using $N_{\text{sim}} = 1000$ simulated datasets (i.e with known third PC direction). Figure 3 show the distribution of the angle θ_{diff} measuring the difference between the theoretical direction and the ones reconstructed by the method. We found the deviation between the real $\hat{K}$ direction simulated and the one reconstructed by PCA was $\theta_{\text{diff}} = (6.2° \pm 0.8°)$ in agreement with the values found when using real data.

Also, this result was extended by analyzing the vectorial components of the third PC $\hat{K}$ reconstructed by PCA as shown in Figure 4.

Discussion

There is already some evidence that the cortical surface behaves as an elastic crumpled surface [8]. Within the framework of this theory, a set of orthogonal morphological variables was proposed as a more natural way of expressing a given cortical shape in the $\{\log T^2, \log A_T, \log A_E\}$ morphometric space [14]. The result presented here show that the directions defined by these new variables largely coincide with the principal components of variance for comparative neuroanatomy data of several dozen mammalian species. The new variables are thus shown to be almost uncorrelated with one another, and can be properly regarded as a set of both natural and independent morphometrics, greatly enhancing their potential analytical usefulness.

The degree of coincidence between morphological variables and principal components was then quantified using a simulated dataset, generated with principal components given by the exponents in equation 1 and the same variance as present in the analyzed dataset. It was found that the deviation between the real $\hat{K}$ direction simulated and the one reconstructed by PCA was $\theta_{\text{diff}} = (6.2° \pm 0.8°)$ (see Figure 3), a value consistent with the one observed in the experimental data.

A deeper explanation about the systematic discrepancy observed between the vectorial components of the theoretically predicted vectors and the ones obtained using the PCA is given in Appendix 1

Considering the analysis of different mammalian species, the result is even stronger than the one obtained for humans. Not only it is in agreement with the theoretical prediction but also it suggests that the independent set of morphological variables K, S and I are indeed a natural mathematical language to describe the cortex, encapsulating orthogonal information about the morphology. Furthermore, the fact that it is possible to understand the cortical morphology using

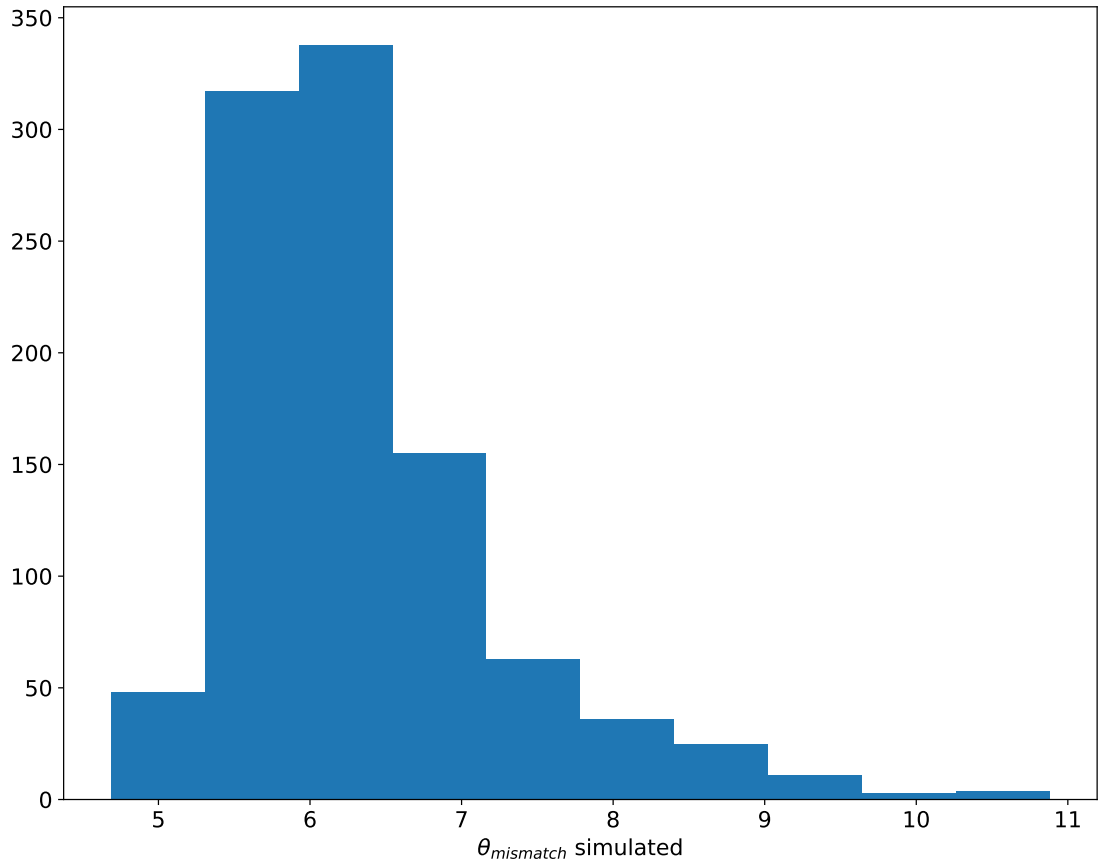

Figure 3: Angular difference θ_{diff} between the real principal component $\hat{K}$ and the one reconstructed by the PCA with the smallest explained variance. The deviation found is $\theta_{\text{diff}} = (6.2° \pm 0.8°)$.

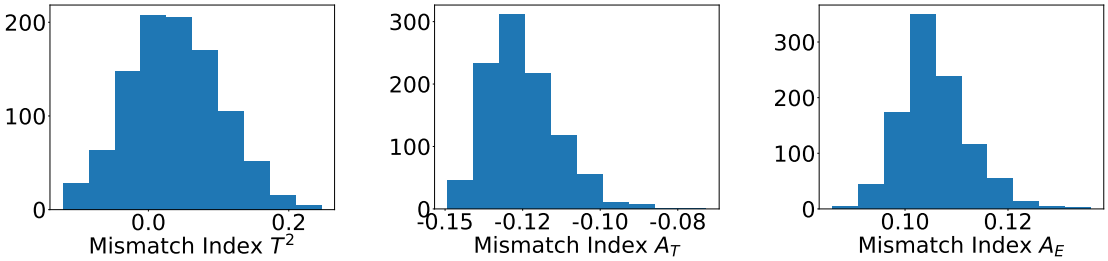

Figure 4: Difference between the expected vectorial components of $\hat{K}$ and the values reconstructed by PCA. The simulation shows that not only PCA has a systematic shift from the generated values, but also does not reach a precision below decimal.

a unique mathematical description advocates in favor of incorporating compared neuroanatomy insights in the study of human diseases.

Acknowledgements

This work was funded by the Instituto Serrapilheira, National Council for Scientific and Technological Development (CNPq), and intramural grants from Instituto D'Or de Pesquisa e Ensino (IDOR). Bruno Mota is supported by Instituto Serrapilheira (grant Serra-1709-16981) and CNPq (PQ 2017 312837/2017-8). We are thankful to the research team and volunteers for participating in this research project. We specially thanks Dr. Yujiang Wang for the careful reading of this paper and the insightful comments, which were really valuable for reaching this final version. The authors have no competing interests.

Author Contributions

V.B.B.M, F.H.P.M., and B.M. designed research; V.B.B.M, F.H.P.M. and B.M. performed research; V.B.B.M analyzed data; and V.B.B.M, F.H.P.M., and B.M. wrote the paper. All authors reviewed the manuscript.

References

[1] Georg F. Striedter, Shyam Srinivasan, and Edwin S. Monuki. Cortical folding: When, where, how, and why? *Annual Review of Neuroscience*, 38(1):291–307, 2015. PMID: 25897870.

[2] George F. Striedter. *Principles of brain evolution.* Principles of brain evolution. Sinauer Associates, Sunderland, MA, US, 2005.

[3] Cristina Llinares-Benadero and Víctor Borrell. Deconstructing cortical folding: genetic, cellular and mechanical determinants. *Nature Reviews. Neuroscience*, 20(3):161–176, March 2019.

[4] Tuomas Tallinen, Jun Young Chung, John S. Biggins, and L. Mahadevan. Gyrification from constrained cortical expansion. *Proceedings of the National Academy of Sciences of the United States of America*, 111(35):12667–12672, Sep 2014.

[5] Tuomas Tallinen, Jun Young Chung, François Rousseau, Nadine Girard, Julien Lefèvre, and L. Mahadevan. On the growth and form of cortical convolutions. *Nature Physics*, 12(6):588–593, Jun 2016.

[6] Lisa Ronan, Natalie Voets, Catarina Rua, Aaron Alexander-Bloch, Morgan Hough, Clare Mackay, Tim J. Crow, Anthony James, Jay N. Giedd, and Paul C. Fletcher. Differential tangential expansion as a mechanism for cortical gyrification. *Cerebral cortex (New York, N.Y. : 1991)*, 24(8):2219–2228, Aug 2014.

[7] Maria A. Holland, Kyle E. Miller, and Ellen Kuhl. Emerging brain morphologies from axonal elongation. *Annals of Biomedical Engineering*, 43(7):1640–1653, Jul 2015.

[8] Bruno Mota and Suzana Herculano-Houzel. Cortical folding scales universally with surface area and thickness, not number of neurons. *Science (New York, N.Y.)*, 349(6243):74–77, July 2015.

[9] Yacov Kantor, Mehran Kardar, and David R. Nelson. Statistical mechanics of tethered surfaces. *Phys. Rev. Lett.*, 57:791–794, Aug 1986.

[10] Yujiang Wang, Joe Necus, Marcus Kaiser, and Bruno Mota. Universality in human cortical folding in health and disease. *Proceedings of the National Academy of Sciences of the United States of America*, 113(45):12820–12825, November 2016.

[11] Bruno Mota, Sandra E. Dos Santos, Lissa Ventura-Antunes, Débora Jardim-Messeder, Kleber Neves, Rodrigo S. Kazu, Stephen Noctor, Kelly Lambert, Mads F. Bertelsen, Paul R. Manger, Chet C. Sherwood, Jon H. Kaas, and Suzana Herculano-Houzel. White matter volume and white/gray matter ratio in mammalian species as a consequence of the universal scaling of cortical folding. *Proceedings of the National Academy of Sciences*, 116(30):15253–15261, 2019.

[12] Kechen Zhang and Terrence J. Sejnowski. A universal scaling law between gray matter and white matter of cerebral cortex. *Proceedings of the National Academy of Sciences*, 97(10):5621–5626, 2000.

[13] Yujiang Wang, Joe Necus, Luis Peraza Rodriguez, Peter Neal Taylor, and Bruno Mota. Human cortical folding across regions within individual brains follows universal scaling law. *Communications Biology*, 2(1):191, May 2019.

[14] Yujiang Wang, Karoline Leiberg, Tobias Ludwig, Bethany Little, Joe H Necus, Gavin Winston, Sjoerd B Vos, Jane de Tisi, John S Duncan, Peter N Taylor, and Bruno Mota. Independent components of human brain morphology. *NeuroImage*, 226:117546, February 2021.

[15] Anneke Alkemade, Martijn J Mulder, Josephine M Groot, Bethany R Isaacs, Nikita van Berendonk, Nicky Lute, Scott JS Isherwood, Pierre-Louis Bazin, and Birte U Forstmann. The Amsterdam Ultra-high field adult lifespan database (AHEAD): A freely available multimodal 7 Tesla submillimeter magnetic resonance imaging database. *NeuroImage*, 221:117200, November 2020.

[16] Lukas Snoek, Maite M. van der Miesen, Tinka Beemsterboer, Andries van der Leij, Annemarie Eigenhuis, and H. Steven Scholte. The Amsterdam Open MRI Collection, a set of multimodal MRI datasets for individual difference analyses. *Scientific Data*, 8(1):85, March 2021.

[17] Bruce Fischl. FreeSurfer. *NeuroImage*, 62(2):774–781, August 2012.

[18] Christopher S. McCarthy, Avinash Ramprashad, Carlie Thompson, Jo-Anna Botti, Ioana L. Coman, and Wendy R. Kates. A comparison of FreeSurfer-generated data with and without manual intervention. *Frontiers in Neuroscience*, 9, 2015.

[19] M. Schaer, M. B. Cuadra, L. Tamarit, F. Lazeyras, S. Eliez, and J. Thiran. A Surface-Based Approach to Quantify Local Cortical Gyrification. *IEEE Transactions on Medical Imaging*, 27(2):161–170, February 2008.

[20] Yujiang Wang, Tobias Ludwig, and Bruno Mota. Analysis pipeline to extract cortical morphology measures from Freesurfer for cortical folding analysis. https://zenodo.org/record/3608675, October 2019.

[21] F. H. P. de Moraes, V. B. B. Mello, F. Tovar-Moll, and B. Mota. Establishing a baseline for human cortical folding morphological variables: a multicenter study. *Frontiers in Neuroscience*, 2022 (accepted for publication).

[22] Ian T. Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

[23] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

1 Limitations on the PCA

The PCA is not able to reconstruct the vectorial components with the required precision, reflecting exactly on the distribution shown in Figure 3. Also, this conclusion extends to the vectorial components of $\vec{K}$ reconstructed by PCA as shown in Figure 4. The analysis not only has a systematic shift from the simulated values, but also does not reach a precision below decimal.

The explanation for this impossibility becomes clearer when one realizes that the PCA is a way to reconstruct all exponents in equation 1. In other words, assuming a relation in the form of equation 3

$$\log k \;=\; \beta \log T + \gamma \log A_T - \alpha \log A_E \tag{3}$$

PCA is performing a joint fit of the exponents α, β and γ and is used to test if these values are compatible with the theoretical ones, $\alpha = 5/4$, $\beta = 1/2$ and $\gamma = 1$.

However, by doing the same task with a standard maximum likelihood approach, it becomes clearer that the available data is not sufficient to determine all this coefficients together with the required precision. Firstly, we start fitting the universal scaling relation in the particular case of lissencephalic mammals for which the relation scales as $\log T \;=\; \gamma \log k + \frac{(\alpha-\gamma)}{\beta} \log A_E$. Figure 5 show the result of fitting this equation for the lissencephalic available data.

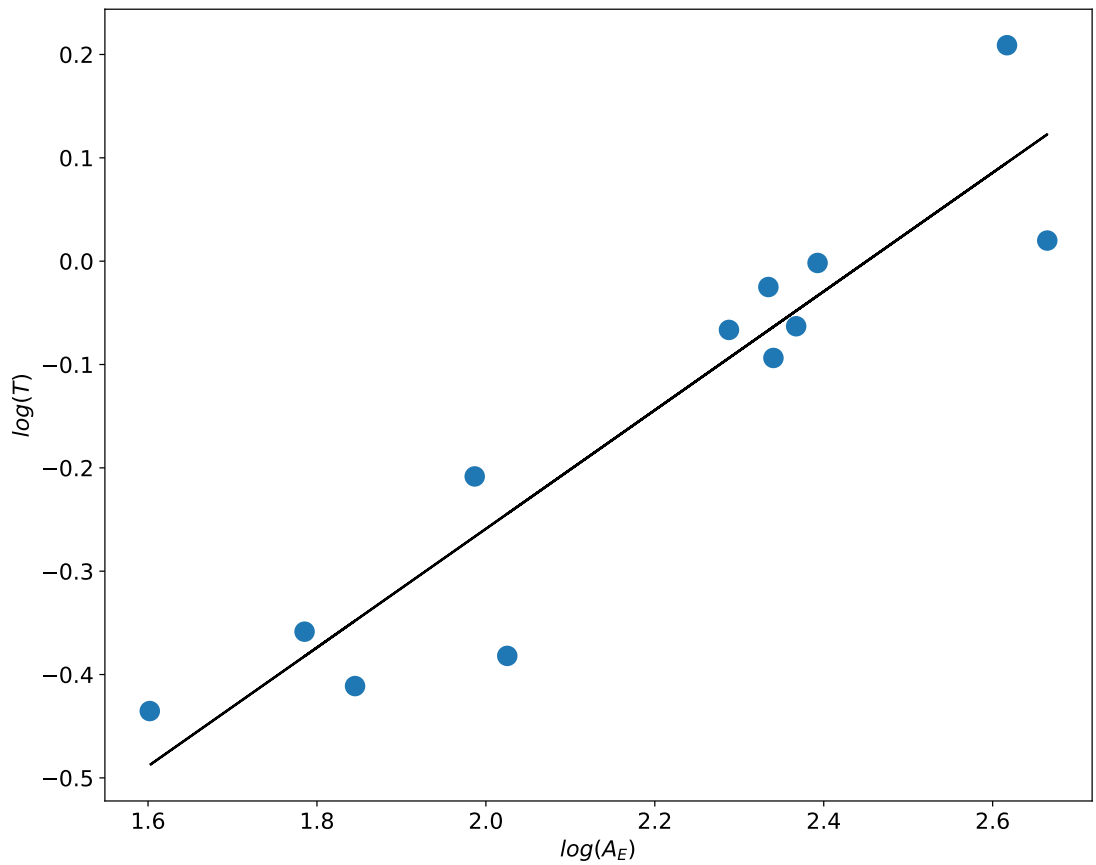

Figure 5: Fitting the universal scaling relation in the particular case of lissencephalic mammals. The angular coefficient $\frac{(\alpha-\gamma)}{\beta} = (0.57 \pm 0.07)$ provides a constraint on the possible values for the exponents α, β and γ in equation 3 and is in agreement with the theoretical predictions $\alpha = 5/4$, $\beta = 1/2$ and $\gamma = 1$.

One may note that angular coefficient $\frac{(\alpha-\gamma)}{\beta} = (0.57 \pm 0.07)$ is compatible with the theoretical predictions, providing a constraint on the possible values for the exponents. Using this constraint, it is possible to fit the data from gyrencefalic

species. Here the human MRI dataset was used because it has significantly more data. Figure 6 shows the log-likelihood as function of the exponents α and β.

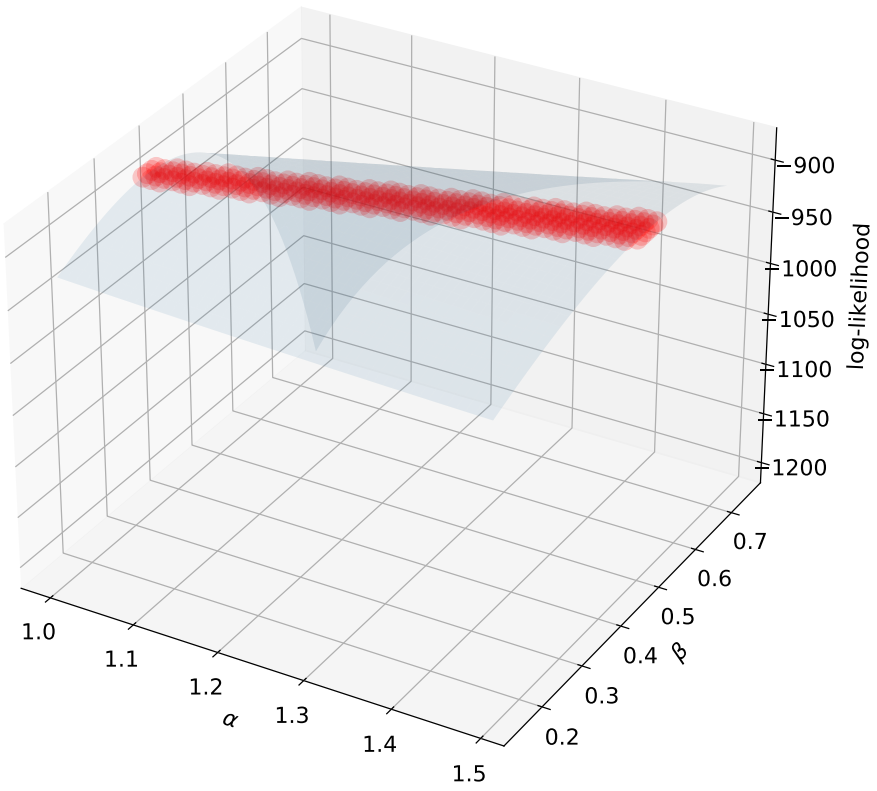

Figure 6: Log-likelihood function with the constraint established by lissencephalic data. Instead of a function peaked around a localized region, it is observed a large range of equivalent peaks, represented in red, that it is compatible with the theoretical values. However, it does not provide a method to precise determine the coefficients.

Instead of a function peaked around a localized region, one obtains a large range of equivalent peaks that, although it is compatible with the theoretical values, fitting these values with high precision is not possible.

2 PCA applied to the human harmonized dataset divided by ages

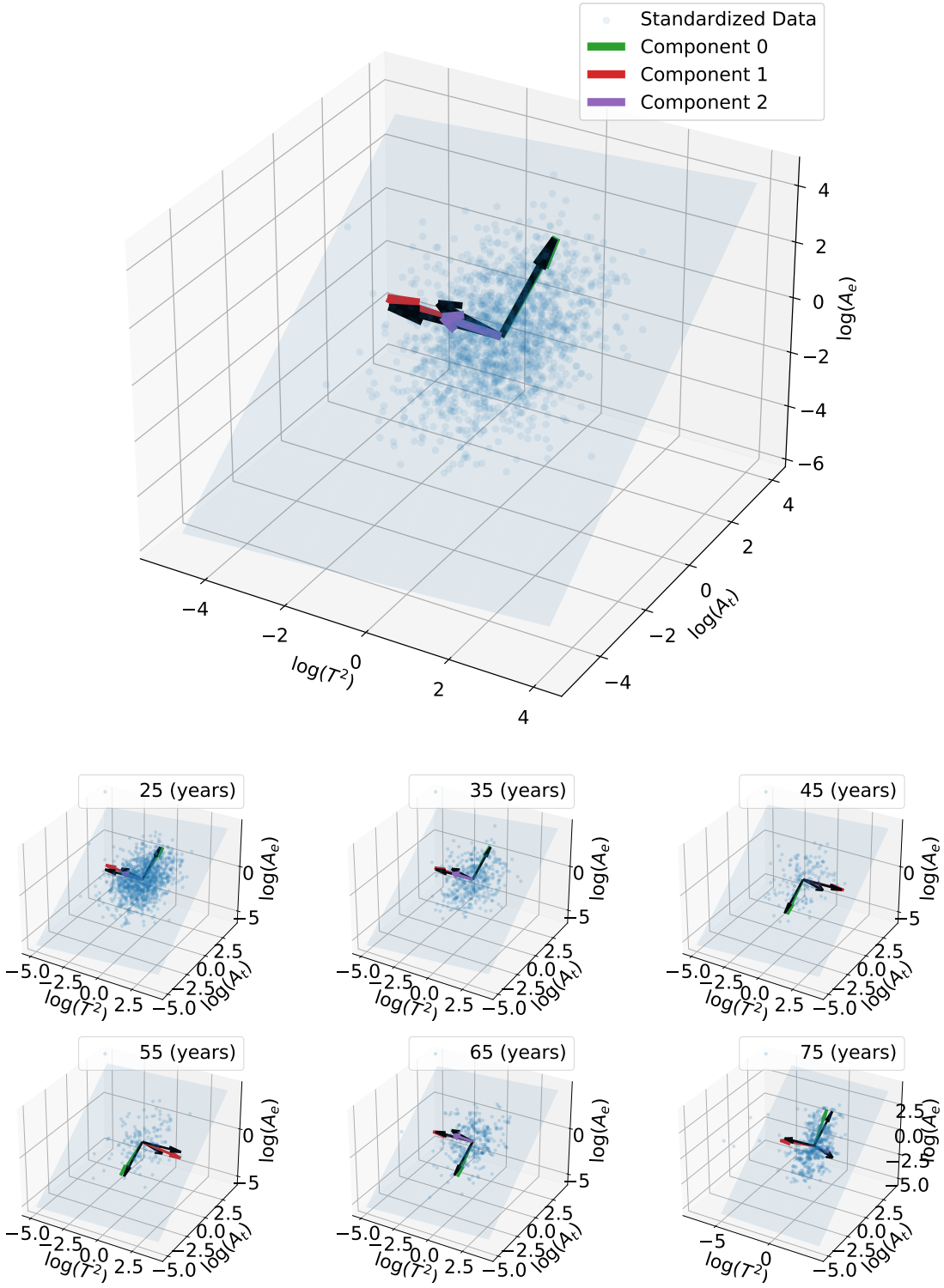

Figure 7: Cortical plane for the human MRI harmonized dataset and the vectors reconstructed by PCA for different age ranges. The arrows in black represents the principal components predicted by the model. A angular difference of 180° on the vectors reflects the ambiguity of selecting the vector normal to the plane thus adding no information.

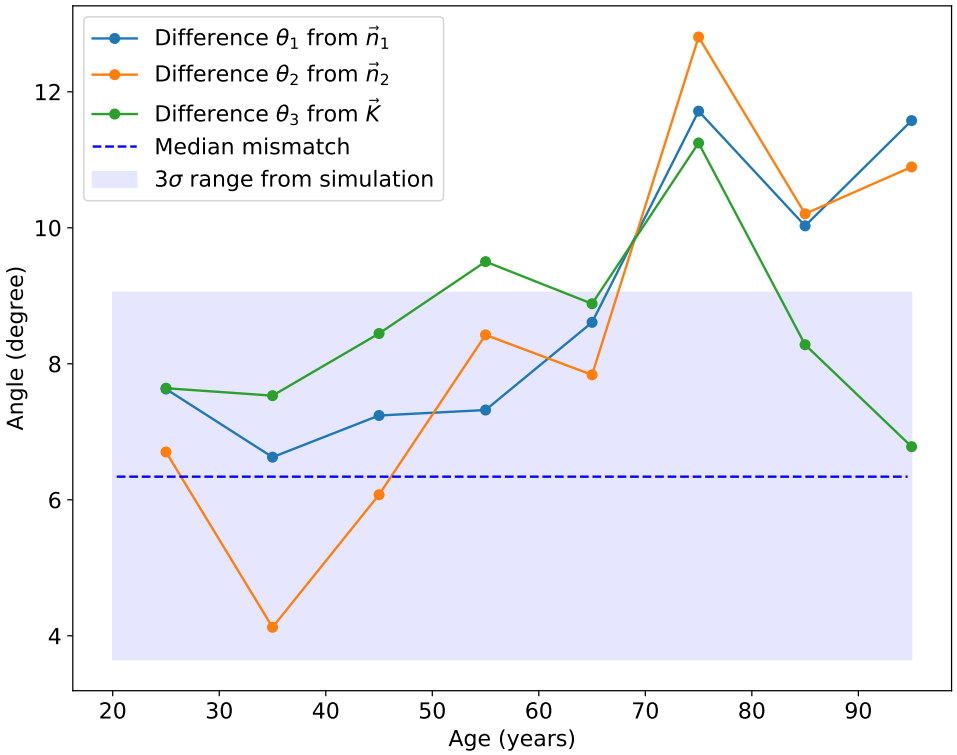

Figure 8: Angular difference between the vectors reconstructed by PCA and the theoretical predictions across healthy aging. The 3σ region was obtained using the simulations. The results are remarkably in agreements with the predictions when considering ages below 65 years. Above this age there is a clear change in the directions with the greatest variability. This deviation from the theoretical predictions correlated with age can be further exploited in future works in order to build a theoretical model for the morphology of healthy cortical aging.

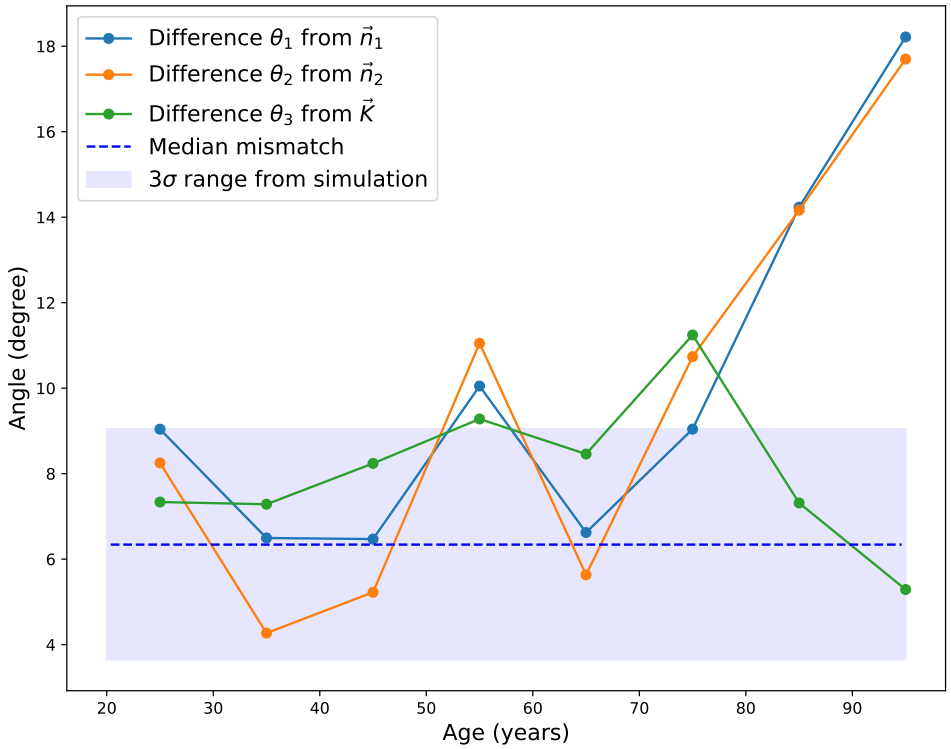

Figure 9: Angular difference between the vectors reconstructed by PCA and the theoretical predictions across healthy aging using the de-aged dataset. The 3σ region was obtained using the simulations. The results are remarkably in agreements with the predictions when considering ages below 65 years. As expected, without the age dependency, the results are constant until the deviation from the model after 65 years. Above this age there is a clear change in the directions with the greatest variability. This deviation from the theoretical predictions correlated with age can be further exploited in future works in order to build a theoretical model for the morphology of the healthy cortical aging.

3 PCA applied to the combined mammalian dataset

The PCA applied to the combined mammalian dataset as shown in figure 10 presents larger deviation from the theoretical. The vectors found are $\hat{v} = (0.417, 0.632, 0.653)$, $\hat{u} = (-0.904, 0.364, 0.224)$, and $\hat{w} = (-0.096, -0.684, 0.723)$, deviating, respectively, $10.6°$, $11.7°$, and $5.8°$ from the normalized theoretical predictions $\hat{I} = (0.577, 0.577, 0.577)$, $\hat{S} = (-0.802, 0.534, 0.267)$, and $\hat{K} = (-0.154, -0.617, 0.771)$. There is an statistically significant increase on the angular difference from $\hat{S}$ and $\hat{I}$. In comparison with the result presented in the main text, in which both datasets had consistent findings when analyzed separately, there is an indicative of a relative uncounted systematic error between the datasets. This illustrates how this PCA method can be used to detect subtle systematic differences between different datasets, and underscores the need to properly harmonize said datasets before performing joint analyzes. The suggestion of an unaccounted systematic error between datasets A and B should be further investigated by analyzing the experimental errors of data acquisition, having the potential to elucidate the experimental value for the index $\alpha = 1.32$.

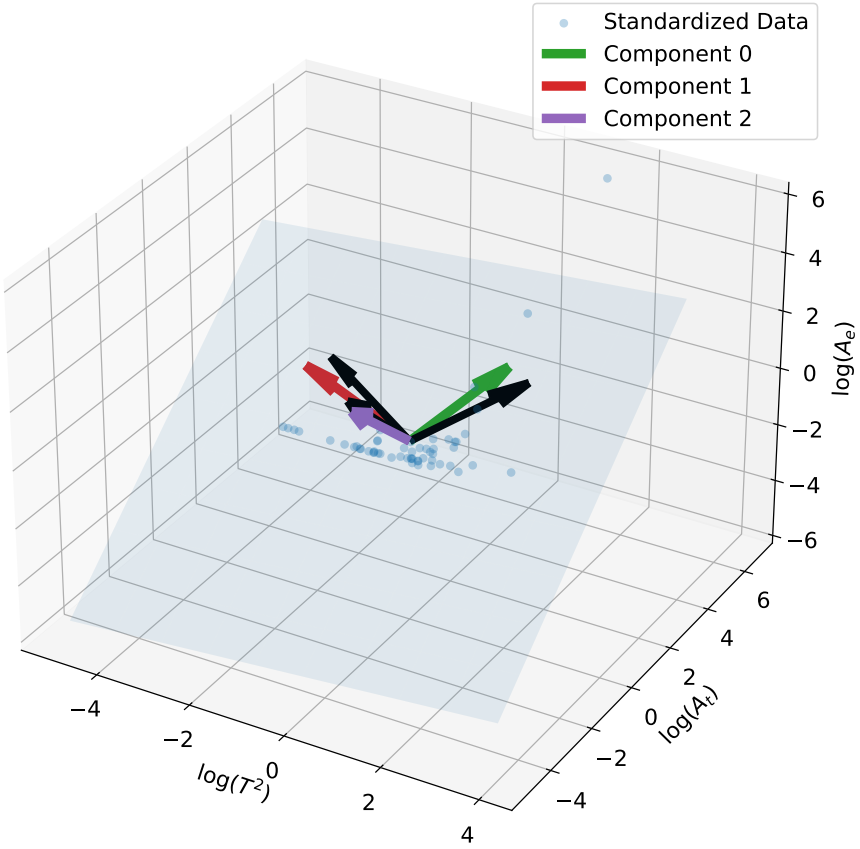

Figure 10: Cortical plane considering different mammalian species dataset and the vectors reconstructed by PCA. The arrows in black represents the principal components predicted by the model, the directions of the natural variables K, S and I. In this analysis the datasets A and B were used together. There is an increase in the angular difference between theoretical components $\hat{S}$ and $\hat{I}$ and the experimental ones. This result is interpreted as an indicative of a relative uncounted systematic error between the datasets.