Submitted to Statistical Methods in Imaging Conference 2024: Theory and Methods Track

https://socr.umich.edu/docs/uploads/2024/Shen_StatisticalFoundationsOfInvarianceInDNN_SMI_2024.pdf

Statistical Foundations of Invariance and Equivariance in Deep Artificial Neural Network Learning - Appendices

Authors: Yueyang Shen (University of Michigan), Yupeng Zhang (University of Wisconsin) and Ivo D. Dinov (University of Michigan).

Conference: American Statistical Association (AmStats) Statistical Methods in Imaging (SMI) 2024

I Biomedical and Imaging Applications

Numerous biomedical datasets contain underlying invariances, which are particularly important to model given the often limited sample sizes. For example, approaches for modeling tensors containing spatiotemporal measurements should be invariant to translation along spatial and temporal dimensions and invariant to spatial rotation. Similarly, molecular measurements such as gene expression should be invariant to permutation of genes. There are likely many other invariances specific to different types of biomedical data, which require domain expertise to elucidate and formulate mathematically. Often, translational research projects need to formulate these as *group* invariances using a framework similar to the invariance concepts described in the introduction.

Two types of validation datasets are commonly used to build, fine-tune, and validate such mathematical representation and computational algorithms. These may include both real and simulated data. Appropriate, rigorous and scalable simulations for validating DL predictions require a setup providing a mechanism to learn a function $h_\theta : Domain(X) \longrightarrow Range(Y)$ parameterized by $\theta \in \Theta$, based on an synthetic (or simulated) training samples $(x, y) \sim p(x, y)$. The joint (model) probability distribution of the process $p(x, y)$ can be constructed to meet specific solution goals, e.g., load certain effects, introduce multivariate relations, control the level of noise, etc. Then a sample from the model distribution can be drawn to represent an empirical sample $D = \{(x_i, y_i)\}_{i=1}^{N}$. The parameter simulator attempts to model a (closely related) distribution $q(x, y|\psi)$ by adjusting q and the corresponding parameters ψ so that $q(x, y|\psi) \sim p(x, y)$. The

deep network modeling goal is typically to automatically learn the parameters of the simulator ψ and optimize the objective function L over another random (validation) dataset $D' = \{(x'_j, y'_j)\}_{j=1}^{N'}$.

An alternative to using targeted simulations, one can use available high-dimensional biomedical informatics datasets to test the new techniques, fine-tune the algorithms, evaluate the computational performance, and quantify the level of DL invariance. Examples of such case-studies include the UK Biobank [1] and MIMIC-III [2]. Figure 1 shows an example of using a $10K \times 7K$ data tensor of the UKBB neuroimaging and clinical archive to illustrate the high-dimensional *learning-compression-clustering* process via uniform manifold approximation and projection (UMAP) [3, 4]. The red and green brain images in the second image shows separation of *patients* and *controls*. This clear split between the two phenotypes indicates good within cohort similarity (within group) and consistent between cohort separation (between-group).

The *UK Biobank (UKBB) archive* represents another rich national health resource containing census-like multisource healthcare data. The archive offers challenges related to supervised and unsupervised learning, model-based and model-free inference, DL and ensemble based techniques [5]. The dataset comprises 500K subjects with $4K$ clinical features and $3K$ derived neuroimaging biomarkers. For validating newly proposed DNN techniques, one can capitalize on an ensemble approach, e.g., compressive big data analytics (CBDA) technique [6], and the deep neural network capabilities of the SuperLearner [7].

The *MIMIC-III archive* also includes complex clinical data with deidentified health information of over $60K$ hospitalizations. It includes $800K$ nursing notes (unstructured data) and over $5K$ structured data elements, e.g., demographics, vital signs, laboratory tests, medications, length of hospitalization, etc. [2, 4].

The *Allen Institute Mouse Brain Data* provides molecular and anatomical measurements of the circuitry of the mouse central nervous system [8]. This archive represents multimodal data of diverse quantitative types referenced to the spatiotemporal Allen Mouse Brain Atlas, including more than $600K$ high quality single-cell and single-nucleus samples assayed by six molecular modalities[9, 10]. The brain connectivity tensor includes $49K$ spatial voxels, each with expression data for $18K$ genes, and axon projection strength between each pair of voxels. The underlying spatial and graphical structure of this tensor implies that

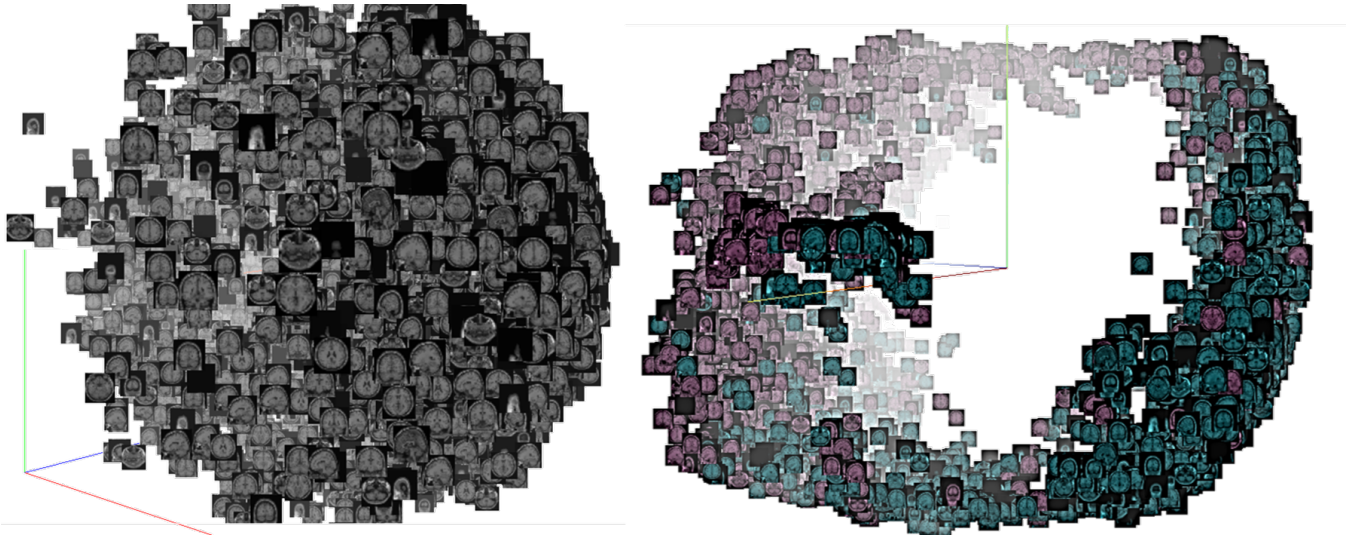

Figure 1: 3D rendering demonstrating the initial (left) and final iteration (right) of a UMAP clustering (400 epochs) of the high-dimensional tensor of the UKBB neuroimaging-clinical data [3, 4].

functions computed over the voxels (such as those learned by a neural network) should be invariant to changes that preserve the spatial and molecular similarities of voxels.

Another direct application of DL-invariance in biomedical applications involves the complex-time (kime) representation, which directly connects fundamental laws of physics, data science, and artificial intelligence [11, 12]. This method provides a prism through which we can explore data-driven decision-making based on translating statistical-physics concepts such as observables, states, wavefunctions, and likelihoods, to their DL counterparts – features, data, inference functions, and probabilities. Spacekime-interpretation [13, 14] of longitudinal data allows explicating the meaning of random variability in observational data as multiple independent sampling of complex-time phases from the compactly supported phase domain. For instance, assume we sample $1,000$ observations, at a fixed $(x = space, t = time)$ location. This allows us to walk through, instantiate, and peer into the (known or unknown) process distribution $1,000$ times. In this process, the kime-phase is linked to the projection of spacekime events into the 4D Minkowski spacetime. At the same time, the kime-phase is also coupled to the random sampling mechanism that collapses the wavefunction into a specific value (scalar or tensor observation) during the data collection procedure.

For example, in spacetime functional neuroimaging studies, 1D voxel-based fMRI BOLD intensities are represented as time-series [15] . In spacekime, the same fMRI signal can be transformed to 2D kime-surfaces whose topology, curvature, and metric properties significantly augment the classical auto-correlation characterization of the corresponding 1D fMRI time-courses, Figure **??**. The richer manifold structure of the fMRI kime-surface representation provides enormous potential to build invariant DL networks that represent event-related fMRI equivalence classes, not subject-specific or noise-related characteristics in the neuroimaging data. Symbolically, let X be the fMRI state space corresponding to all possible resting and stimulation conditions and $G \subseteq \{g : X \to X\}$ be a group of transformations acting on X, e.g., affine transformations scaling/shifting the fMRI intensities or rotating/shearing the fMRI tensors. Two fMRI volumes $x_1, x_2 \in X$ obtained under the same conditions (controlled stimulus experiment) following the same probability distribution would be *equivalent* with respect to G, if $\exists g_2 \in G$, s.t. $x_2 = g_2(x_1)$. The *orbit* of $x_1 \in X$ includes the *equivalence class* of all $x \sim x_1$, i.e., $X(x_1) = \{g(x_1) : g \in G\}$. If θ is a parameter representing the fMRI stimuli (e.g., rest, audio, visual, motor, taste, spell), the space of the corresponding density functions $F = \{f(\cdot|\theta) : \forall \theta \in \Theta\}$, which model the fMRI intensities, is *invariant* under the action of the group G when $\forall g \in G,\ \forall \theta \in \Theta,\ \exists \theta_g \in \Theta$, such that $x_2 \equiv g(x_1)$ has density $f_{x_2} = f_{x_1}(x_2|\theta_g)$, where $x \sim f_x \equiv f_x(x|\theta)$. To estimate the parameter given some observed data, e.g., compute $a = a(\theta|x) \in A$, we can optimize a loss function $L(\theta, a)$. *Invariance* of L with respect to G would imply that $\forall g \in G,\ \forall a \in A$, $\exists a_g \in A$ such that $L(\theta, a) = L(\theta_g, a_g),\ \forall \theta \in \Theta$. For instance, to identify brain locations involved in processing a motor-task (e.g., finger-tapping), we can build a DL network that picks only 3D voxels (locations) where the distribution of the fMRI intensities during motor *stimulation* ($f(x|\theta_s)$), ON-state, is different from the corresponding *resting-state* fMRI distribution ($f(x|\theta_r)$), OFF-state. The key element here is to build the DL network on the fMRI voxel-indexed kime-surfaces, not as traditionally done using the raw 1D time-series. *DL modeling of the kime-surfaces may encode some of the intrinsic topological structure embedded in the fMRI kime-manifold.*

The spacekime representation generalizes longitudinal processes such as classical time-series, defined over the positive reals, to kime-surfaces, defined over the complex plane. This abstraction provides a fertile

ground for development of innovative deep learning techniques that capitalize on the space-kime structure of the enriched state-space. Additionally, a statistical formulation of spacekime analytics in a Bayesian inference framework [16] provides a direct realization of classical random sampling as generation of independent kime-phases from the phase state distribution [13]. This framework facilitates the approximation of the prior-predictive distribution and the calculation of the posterior-predictive distribution. Applying these kime-derived posterior distributions to examine DL estimate-invariance in longitudinal datasets is expected to increase prediction accuracy, improve extrapolating forecasts, and increase the precision of likelihood approximations (e.g., improve statistical inference).

References

[1] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *Plos med*, 12(3):e1001779, 2015.

[2] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[3] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[4] Ivo D Dinov. *Data Science and Predictive Analytics*. Springer, 2018.

[5] Yiwang Zhou, Lu Zhao, Nina Zhou, Yi Zhao, Simeone Marino, Tuo Wang, Hanbo Sun, Arthur W Toga, and Ivo D Dinov. Predictive big data analytics using the uk biobank data. *Scientific reports*, 9(1):1–10, 2019.

[6] Simeone Marino, Yi Zhao, Nina Zhou, Yiwang Zhou, Arthur W. Toga, Lu Zhao, Yingsi Jian, Yichen Yang, Yehu Chen, Qiucheng Wu, Jessica Wild, Brandon Cummings, and Ivo D. Dinov. Compressive big data analytics: An ensemble meta-algorithm for high-dimensional multisource datasets. *PLOS ONE*, 15(8):1–21, 08 2020.

[7] Mark J Van der Laan, Eric C Polley, and Alan E Hubbard. Super learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007.

[8] Quanxin Wang, Song-Lin Ding, Yang Li, Josh Royall, David Feng, Phil Lesnar, Nile Graddis, Maitham Naeemi, Benjamin Facer, Anh Ho, et al. The allen mouse brain common coordinate framework: a 3d reference atlas. *Cell*, 181(4):936–953, 2020.

[9] Zizhen Yao, Hanqing Liu, and Xie et al. An integrated transcriptomic and epigenomic atlas of mouse primary motor cortex cell types. *bioRxiv*, 2020.

[10] Pascal Grange, Jason W Bohland, Benjamin W Okaty, Ken Sugino, Hemant Bokil, Sacha B Nelson, Lydia Ng, Michael Hawrylycz, and Partha P Mitra. Cell-type–based model explaining coexpression patterns of genes in the brain. *Proceedings of the National Academy of Sciences*, 111(14):5397–5402, 2014.

[11] Zhang Y Liu Y Guo Y Shen Y Deng D Qiu Y Dinov ID. Zhang, R. Kimesurface representation and tensor linear modeling of longitudinal data. *Neural Comput Applic*, page in press, 2022.

[12] Ivo D. Dinov and Milen Velchev Velev. *Data Science: Time Complexity, Inferential Uncertainty, and Spacekime Analytics*. De Gruyter, 2021.

[13] I.D. Dinov and M.V. Velev. *Data Science: Time Complexity, Inferential Uncertainty, and Spacekime Analytics*. De Gruyter STEM. Walter de Gruyter GmbH, 2021.

[14] Ivo Dinov. Data science, time complexity, and spacekime analytics. *Bulletin of the American Physical Society*, 2021.

[15] S.H. Faro and F.B. Mohamed. *BOLD fMRI: A Guide to Functional Imaging for Neuroscientists*. SpringerLink : Bücher. Springer New York, 2010.

[16] D.C. Knill and W. Richards. *Perception as Bayesian Inference*. Cambridge University Press, 1996.

[17] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pages 2990–2999. PMLR, 2016.

[18] Rui Wang, Robin Walters, and Rose Yu. Approximately equivariant networks for imperfectly symmetric dynamics. In *International Conference on Machine Learning*, pages 23078–23091. PMLR, 2022.

[19] Persi Diaconis. Finite forms of de finetti's theorem on exchangeability. *Synthese*, 36:271–281, 1977.

Appendix A. Symmetry breaking for GCNN

Let f be the input signal to the layer and ψ be the parameter for the convolution kernel, the group convolution theorem states that ***the linear layer in MLP connections is equivariant iff and only if (iff) it is a group convolution with the form***

$$[f *_G \psi](g) = \sum_{h \in G} f(h)\psi(g^{-1}h) \tag{1}$$

where $f(h)$ is the Haar measure, which assigns weights to the group elements for averaging. This can be explicated as

$$\begin{aligned}\Big[[L_u f] *_G \psi\Big](g) &= \sum_{h \in G} f(u^{-1}h)\psi(g^{-1}h) = \sum_{h \in G} f(h)\psi(g^{-1}uh) = \sum_{h \in G} f(h)\psi((u^{-1}g)^{-1}h) \\ &= \Big[L_u[f *_G \psi]\Big](g)\end{aligned} \tag{2}$$

Namely, $\Big[[L_u f] *_{\mathbb{Z}^2} \psi\Big](g) = \Big[L_u[f *_G \psi]\Big](g)$ is the equivariance condition.
Example: For translational group $\mathbb{Z}^2$, this can be explicated as [17]

$$\Big[[L_t f] *_{\mathbb{Z}^2} \psi\Big](x) = \sum_{y \in \mathbb{Z}^2} f(y-t)\psi(y-x) = \sum_{y \in \mathbb{Z}^2} f(y)\psi(y-(x-t)) = \Big[L_t[f *_{\mathbb{Z}^2} \psi]\Big](x) \tag{3}$$

Another example is the scaling symmetry $\mathbb{Z}_{>0}$. *Symmetry breaking*: The way to perform symmetry breaking is by coupling the equivariant kernel with group dependent weight $w_l(h)$ [18]

$$[f *_G \psi](g) = \sum_{h \in G} f(h) \sum_{l=1}^{L} w_l(h)\psi_l(g^{-1}h) \tag{4}$$

We prove this is non-perfect equivariance. We illustrate this using the translational CNN example

$$[[L_t f] * \psi](x) = \sum_{y \in \mathbb{Z}^2} f(y-t) \sum_{l=1}^{L} w_l(y)\psi_l(y-x) = \sum_{y \in \mathbb{Z}^2} f(y) \sum_{l=1}^{L} w_l(y+t)\,\psi_l(y-(x-t)) \tag{5}$$

On the other hand

$$\Big[L_t[f * \psi]\Big](x) = \sum_{y \in \mathbb{Z}^2} f(y)\psi((x-t), y) = \sum_{y \in \mathbb{Z}^2} f(y) \sum_{l=1}^{L} w_l(y)\,\psi_l(y-(x-t)) \tag{6}$$

where $\psi(g,h) = \sum_{l=1}^{L} w_l(h)\psi_l(g^{-1}h)$, $g = x-t, y = h$. The discrepancy between these two terms presents the **equivariant gap**.

Appendix B. The toy coin toss

Consider two coin toss with 4 possible outcomes $p_1 = p(X_1 = 0, X_2 = 0), p_2 = P(X_1 = 0, X_2 = 1), p_3 = P(X_1 = 1, X_2 = 0), p_4 = P(X_1 = 1, X_2 = 1)$. Consider the 4-simplex that is parametrized by (p_1, p_2, p_3, p_4). Exchangeability implies that $p_2 = p_3$. Thus, the exchangeability 4-simplex can be parametrized by $(t, p, p, 1 - 2p - t)$. Independence implies that $p_1 = (p_1 + p_2)(p_1 + p_3), p_2 = (p_1 + p_2)(p_2 + p_4), p_3 = (p_1 + p_3)(p_3 + p_4), p_4 = (p_2 + p_4)(p_3 + p_4)$, All the solution lies on the curve $(p^2, p - p^2, p - p^2, (1 - p)^2)$ and one can argue any adjustment does not satisfy the condition. Finally, for admitting a finite DeFinetti representation, $p(X_i = 0 \mid p) = p, p(X_i = 1 \mid p) = 1 - p$. Thus, $p_1 = \int_{[0,1]} p^2 d\mu(p) = \mathbb{E}[p^2]$. Similarly, the 4-tuple set should be $D = (\mathbb{E}_\mu[p^2], \mathbb{E}_\mu[p] - \mathbb{E}_\mu[p^2], \mathbb{E}_\mu[p] - \mathbb{E}_\mu[p^2], 1 - 2\mathbb{E}_\mu[p] + \mathbb{E}_\mu[p^2])$. We show that the ray from $(0, 0, 0, 1)$[1] to D must cross some $P' = (p'^2, p' - p'^2, p' - p'^2, (1 - p')^2)$ with (See Figure 2)

$$\frac{p'^2}{\mathbb{E}_\mu[p^2]} = \frac{p' - p'^2}{\mathbb{E}_\mu[p] - \mathbb{E}_\mu[p^2]} = \frac{(1 - p')^2 - 1}{1 - 2\mathbb{E}_\mu[p] + \mathbb{E}_\mu[p^2] - 1} = \lambda, \lambda \geq 1 \tag{7}$$

The equation boils down to $p'^2 = \lambda \mathbb{E}_\mu[p^2], p' = \lambda \mathbb{E}_\mu[p]$. Canceling p', $\lambda = \frac{\mathbb{E}_\mu[p^2]}{(\mathbb{E}_\mu[p])^2} \geq 1$ from the fact that variance is non-negative. To facilitate the computation and visualization, we use random simulation of the Barycentric coordinates within the simplex and filter numerically the points that matches the three conditions discussed above.

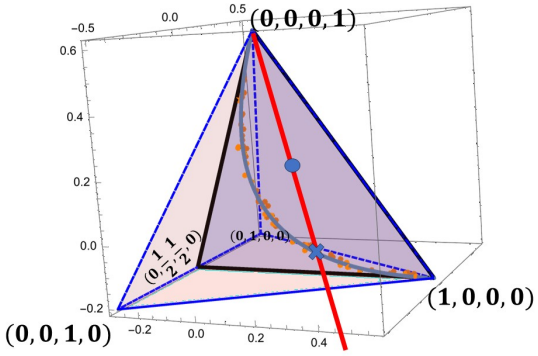

Figure 2: 3D coordinate converted from Barycentric coordinates. Yellow scatter: Numerical scatters satisfying the independence condition with the ground truth curve that contains all the region satisfying independence. The region to the right of the grey curve admits De Finetti representation. The black triangle covered region satisfy exchangeability. The blue blob is D and the cross is P' (Equation 7). Consistent with[19].

1. The (1,0,0,0) scenario is similar.