The Journal of Open Source Software

PyNM: a Lightweight Python implementation of Normative Modeling

Annabelle Harvey[12] **and Guillaume Dumas**[13]

1 Centre de Recherche du CHU Sainte-Justine, Université de Montréal, QC, Canada **2** Centre de Recherche de l'Institut Universitaire de Gériatrie de Montréal, Université de Montréal, QC, Canada **3** Mila - Quebec AI Institute, Université de Montréal, QC, Canada

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor:

Submitted: 11 March 2022
Published: unpublished

Summary

The majority of studies in neuroimaging and psychiatry are focussed on case-control analysis (Marquand et al., 2019). However, case-control relies on well defined groups which is more the exception than the rule in biology. Psychiatric conditions are diagnosed based on symptoms alone, which makes for heterogeneity at the biological level (Marquand et al., 2016). Relying on mean differences obscures this heterogeneity and the resulting loss of information can produce unreliable results or misleading conclusions (Loth et al., 2021).

Normative Modeling is an emerging alternative to case-control analyses that seeks to parse heterogeneity by looking at how individuals deviate from the normal trajectory. Analogous to normative growth charts, normative models map the mean and variance of a trait for a given population against a set of explanatory variables (usually including age). Statistical inferences at the level of the individual participant can then be obtained with respect to the normative range (Marquand et al., 2019). This framework can detect patterns of abnormality that might not be consistent across the population and recasts disease as an extreme of the normal range.

PyNM is a lightweight python implementation of Normative Modeling making it approachable and easy to adopt. The package provides:

- Python API and a command-line interface for wide accessibility
- Automatic dataset splitting and cross-validation
- Five models from various back-ends in a unified interface that cover a broad range of common use cases
- Solutions for very large datasets and heteroskedastic data
- Integrated plotting and evaluation functions to quickly check the validity of the model fit and results
- Comprehensive and interactive tutorials

Statement of need

The basic idea underpinning Normative Modeling is to fit a model on the controls (or a subset of them) of a dataset, and then apply it to the rest of the participants. The difference between the model's prediction and the ground truth for the unseen participants relative to the variance around the prediction quantifies their deviation from the normal. While simple in concept, implementing Normative Modeling requires some care in managing the dataset and choosing an appropriate model.

In principle, any model that estimates both the mean and variance of the predictive distribution could be used for Normative Modeling. However, in practice we impose more constraints. First and foremost, the assumptions of the model must be met by the data. Second, we want to

distinguish between epistemic and aleatoric uncertainty. Epistemic or systematic uncertainty stems from how information about the distribution is collected, whereas aleatoric uncertainty is intrinsic to the distribution and represents the true variation of the population (Xu et al., 2021).

To the author's knowledge, PCNtoolkit (Marquand, 2020) is the only other available package for Normative Modeling. It implements methods that have been applied in a range of psychiatry and neuroimaging studies including (Kia et al., 2020) and (Fraza et al., 2021), and is accompanied by thorough tutorials and a framework for Normative Modeling in computational psychiatry (Rutherford et al., 2021). While it includes features that make it an obvious choice for advanced users in many cases, is not as approachable to beginners and does not implement several key models.

PyNM is intended to take users through their first steps in Normative Modeling to using advanced models on complex datasets. Crucially, it manages the dataset and has interactive tutorials – making it quick for new users to try the method either on their own data or on provided simulated data. The tutorials motivate the use of each model and highlight their limitations to help clarify which model is appropriate for what data, and built in plotting and evaluation functions (Figure 1) make it simple to check the validity of the model output. The package includes five models from various backends in a unified interface, including a wrapper for GAMLSS (Rigby & Stasinopoulos, 2005) from R that is otherwise not yet available in python, and the selected models cover many settings including big data and heteroskedasticity.

Earlier versions of PyNM code were used in the following publications:

- Lefebvre et al. (2018)
- Maruani et al. (2019)
- Bethlehem et al. (2020)

Usage Example

```
from pynm.pynm import PyNM

# Load data
# df contains columns 'score','group','age','sex','site'
df = pd.read_csv('data.csv')

# Initialize pynm w/ data and confounds
m = PyNM(df,'score','group', confounds = ['age','c(sex)','c(site)'])

# Run models
m.loess_normative_model()
m.centiles_normative_model()
m.gp_normative_model()
m.gamlss_normative_model()

# Collect output
data = m.data
```

The Journal of Open Source Software

Figures

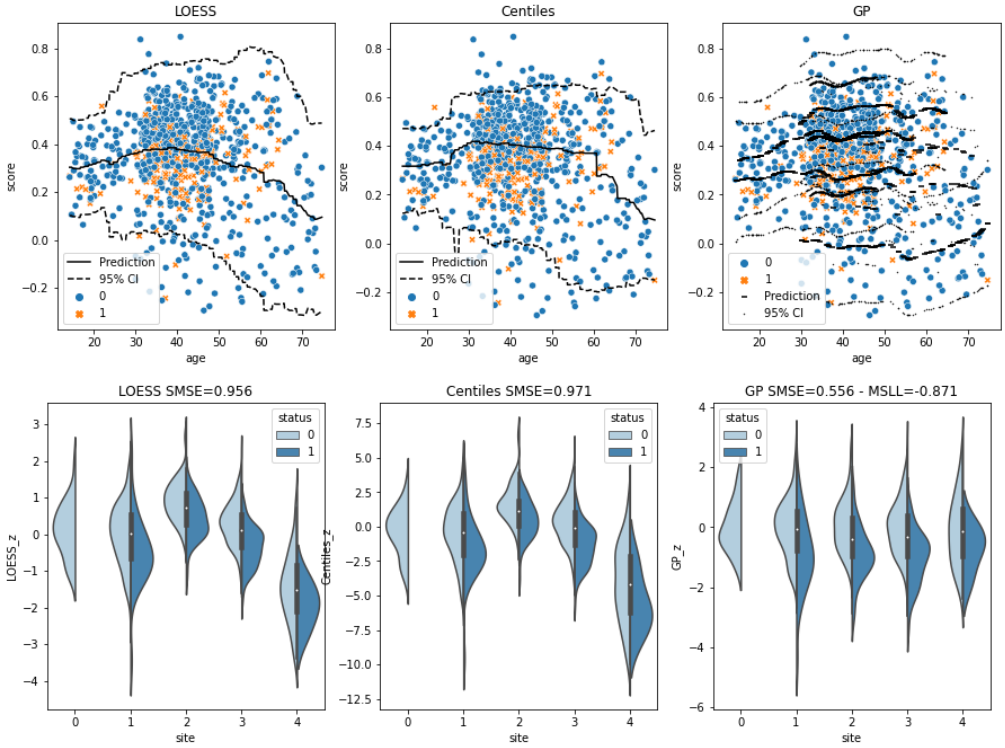

Figure 1: Output of built in plotting function for model fit and residuals.

Acknowledgements

The development of this code has benefited from useful discussions with Andre Marquand, Thomas Wolfers, Eva Loth, Jumana Amad, Richard Bethelem, and Michael Lombardo.

Fundings: This work is supported by IVADO, FRQS, CFI, MITACS, and Compute Canada.

References

Bethlehem, R. A. I., Seidlitz, J., Romero-Garcia, R., Trakoshis, S., Dumas, G., & Lombardo, M. V. (2020). A normative modelling approach reveals age-atypical cortical thickness in a subgroup of males with autism spectrum disorder. *Communications Biology*, *3*(1), 486. https://doi.org/10.1038/s42003-020-01212-9

Fraza, C. J., Dinga, R., Beckmann, C. F., & Marquand, A. F. (2021). Warped bayesian linear regression for normative modelling of big data. *NeuroImage*, *245*, 118715. https://doi.org/https://doi.org/10.1016/j.neuroimage.2021.118715

Kia, S. M., Huijsdens, H., Dinga, R., Wolfers, T., Mennes, M., Andreassen, O. A., Westlye, L. T., Beckmann, C. F., & Marquand, A. F. (2020). Hierarchical bayesian regression for multi-site normative modeling of neuroimaging data. In A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, & L. Joskowicz (Eds.), *Medical image computing and computer assisted intervention – MICCAI 2020* (pp. 699–709). Springer International Publishing. ISBN: 978-3-030-59728-3

Lefebvre, A., Delorme, R., Delanoë, C., Amsellem, F., Beggiato, A., Germanaud, D., Bourgeron, T., Toro, R., & Dumas, G. (2018). Alpha waves as a neuromarker of autism spectrum disorder: The challenge of reproducibility and heterogeneity. *Frontiers in Neuroscience*, *12*. https://doi.org/10.3389/fnins.2018.00662

Loth, E., Ahmad, J., Chatham, C., López, B., Carter, B., Crawley, D., Oakley, B., Hayward, H., Cooke, J., San José Cáceres, A., Bzdok, D., Jones, E., Charman, T., Beckmann, C., Bourgeron, T., Toro, R., Buitelaar, J., Murphy, D., & Dumas, G. (2021). The meaning of significant mean group differences for biomarker discovery. *PLOS Computational Biology*, *17*(11), 1–16. https://doi.org/10.1371/journal.pcbi.1009477

Marquand, A. F. (2020). PCNtoolkit: Predictive clinical neuroscience toolkit. In *GitHub repository*. GitHub. https://github.com/amarquand/PCNtoolkit

Marquand, A. F., Kia, S. M., Zabihi, M., Wolfers, T., Buitelaar, J. K., & Beckmann, C. F. (2019). Conceptualizing mental disorders as deviations from normative functioning. *Molecular Psychiatry*, *24*(10), 1415–1424. https://doi.org/10.1038/s41380-019-0441-1

Marquand, A. F., Rezek, I., Buitelaar, J., & Beckmann, C. F. (2016). Understanding heterogeneity in clinical cohorts using normative models: Beyond case-control studies. *Biological Psychiatry*, *80*(7), 552–561. https://doi.org/10.1016/j.biopsych.2015.12.023

Maruani, A., Dumas, G., Beggiato, A., Traut, N., Peyre, H., Cohen-Freoua, A., Amsellem, F., Elmaleh, M., Germanaud, D., Launay, J.-M., Bourgeron, T., Toro, R., & Delorme, R. (2019). Morning plasma melatonin differences in autism: Beyond the impact of pineal gland volume. *Frontiers in Psychiatry*, *10*. https://doi.org/10.3389/fpsyt.2019.00011

Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape,(with discussion). *Applied Statistics*, *54.3*, 507–554.

Rutherford, S., Kia, S. M., Wolfers, T., Fraza, C., Zabihi, M., Dinga, R., Berthet, P., Worker, A., Verdi, S., Ruhe, H. G., Beckmann, C. F., & Marquand, A. F. (2021). The normative modeling framework for computational psychiatry. *bioRxiv*. https://doi.org/10.1101/2021.08.08.455583

Xu, B., Kuplicki, R., Sen, S., & Paulus, M. P. (2021). The pitfalls of using gaussian process regression for normative modeling. *PLOS ONE*, *16*(9), 1–14. https://doi.org/10.1371/journal.pone.0252108