The Journal of Open Source Software

FSML – A Modern Fortran Statistics and Machine Learning Library

Sebastian Gerhard Mutz [1]

1 School of Geographical and Earth Sciences, University of Glasgow

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Open Journals

Reviewers:

- @openjournals

Submitted: 01 January 1970
Published: unpublished

Summary

FSML is a modern Fortran statistics and machine learning library suitable for contemporary research problems and teaching. It includes procedures for basic statistics, hypothesis tests, linear and non-linear methods, and statistical distribution functions.

Statement of Need

The advances in computing technology over the past two decades have expanded the practical scope of statistics and allowed the widespread use of machine learning (ML). This also transformed research practices and enhanced predictive modelling across many disciplines, including Earth sciences (Boateng & Mutz, 2023), operational weather forecasting (Lang et al., 2024), and more.

Fortran is a well-established general purpose programming language that is commonly adopted in science due to its stability, reliability, performance, and array functionality. It is widely used for parallelised high-performance computing and numerical modelling (e.g., Giorgetta et al., 2018). The same strenghts make it suitable for computationally demanding ML procedures and data-driven predictions. Furthermore, it is more energy-efficient than other high-level programming languages (Pereira et al., 2021), which is another factor to consider as the widespread adoption of computationally demanding ML techniques increases electricity consumption (Jia, 2024), adds more stress on Earth's climate and environments, and creates new challenges as a consequence (e.g., Dodge et al., 2022; Freitag et al., 2021). Despite Fortran's long history in data-driven prediction and ML (e.g., Breiman, 2001; Gutmann et al., 2022; Tomassetti et al., 2009), it has not been as widely adopted in these fields as other languages and lacks well documented, accessible toolkits for statistics and classic ML. While projects like Neural-Fortran(Curcic, 2019), ATHENA(Taylor, 2024), and FStats cover some important procedures for deep-learning and classic statistics, the Fortran statistics and ML ecosystem remains relatively small. This potentially deters from the use of Fortran, which is already perceived as less accessible than other popular languages due to 1) the lack of familiarity with modern Fortran features, which is exacerbated by stagnating adoption of Fortran at universities, and 2) shortcomings that are currently being addressed by Fortran-lang community projects (Kedward et al., 2022).

FSML (Fortran Statistics and Machine Learning) purposefully integrates these projects: It uses stdlib for linear algebra, leverages fpm for easier building and distribution, and is developed to support compilation with the interactive LFortran compiler in addition to GFortran. As such, it builds on recent community efforts and addresses two needs:

1. It adds to the modern Fortran statistics and ML software ecosytem; a richer ecosystem makes Fortran a more attractive choice as a robust, high-performance, energy-efficient option.

2. The use of fpm, the support of free open-source compilers, the extensive documentation, and its permissive license (MIT) faciliate its early adoption and integration into various statistics and ML projects by students, early career researchers, and teachers. It can thus help counter the stagnating adoption of Fortran.

Software Description

Scope

FSML consists of a set of accessible and well-documented statistics and ML procedures, suitable for many contemporary research problems and teaching. These procedures are categorised into five thematic modules:

- DST: Statistical distribution functions (e.g., the probability density, cumulative distribution, and quantile functions of the Student's t and generalised Pareto distributions).
- STS: Basic statistics for describing and understanding data (e.g., mean, variance, correlation).
- TST: Parametric and non-parametric hypothesis tests (e.g., analysis of variance, Mann–Whitney U).
- LIN: Statistical procedures relying heavily on linear algebra (e.g., principal component analysis, ridge regression, linear discriminant analysis).
- NLP: Non-linear and algorithmic procedures (e.g., k-means clustering).

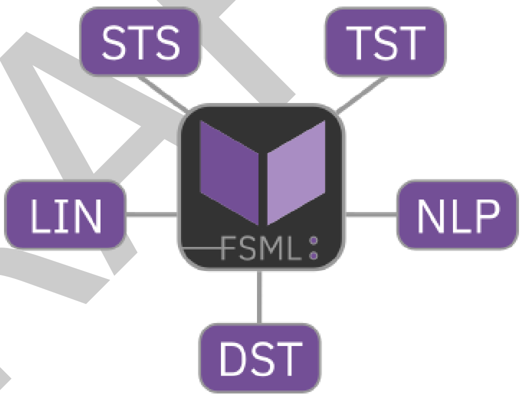

Figure 1: FSML has five thematic modules: Basic statistics (STS), hypothesis tests (TST), linear procedures (LIN), non-linear procedures (NLP), and statistical distribution functions (DST).

FSML has minimal requirements. It uses Fortran 2008 features, Fortran-lang stdlib for linear algebra, and fpm for easy building and distribution.

Documentation

The FSML handbook is hosted on fsml.mutz.science and can be re-generated from its source files. It includes a detailed, example-rich documentation of the covered procedures, as well as installation instructions and information for contributors.

Examples

The examples below demonstrate the use of zFSML interfaces, using double precision (dp):

- statistical distribution functions:

```fortran
! exponential distribution probability density function
! with x=0.8 and lambda=0.5
fx = fsml_exp_pdf(0.8_dp, lambda=0.5_dp)
! generalised Pareto cumulative distribution function
! with modified shape (xi) and location (mu) parameters
fx = fsml_gpd_cdf(1.9_dp, xi=1.2_dp, mu=0.6_dp)
```

- sample statistics and dependency measures:

```fortran
! mean of vector x
mean = fsml_mean(x)
! sample standard deviation of vector x
std = fsml_std(x, ddf=1.0_dp)
! Pearson correlation coefficient for vectors x1 and x2
pcc = fsml_pcc(x1, x2)
```

- hypothesis tests:

```fortran
! two-sample t-test for unequal variances (Welch's t-test);
! returns test statistic (t), degrees of freedom (df), and p
call fsml_ttest_2sample(x1, x2, t, df, p, eq_var=.false.)
! one-way ANOVA on a rank-2 array (x2d);
! returns f-statistic (f), degrees of freedom (df1, df2) and p
call fsml_anova_1way(x2d, f, df1, df2, p)
```

- multiple linear ridge regression:

```fortran
! ridge regression for 100 data points, 5 variables, and lambda=0.2;
! returns y intercept (b0), regression coefficients (b), and R^2 (rsq)
call fsml_ridge(x, y, 100, 5, 0.2_dp, b0, b, rsq)
```

FSML's repository and handbook includes examples for every public interface.

Past and Ongoing FSML Projects

The FSML procedures for clustering and linear discriminant analysis were reworked from the code used for climate pattern detection and explanation (Mutz et al., 2018; Mutz & Ehlers, 2019). FSML's empirical orthogonal functions and analysis of variance were used in Mutz (2025). FSML's distribution functions are currently used for modelling climate extremes.

Acknowledgements

I gratefully acknowledge the Fortran-lang community efforts that this project integrates (fpm, stdlib, and LFortran), as well as the always helpful discussions the with the same community on Fortran-lang discourse and GitHub. I also extend my gratitude to Herbert Peck.

References

Boateng, D., & Mutz, S. G. (2023). pyESDv1.0.1: An open-source python framework for empirical-statistical downscaling of climate information. *Geoscientific Model Development*, *16*(22), 6479–6514. https://doi.org/10.5194/gmd-16-6479-2023

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Curcic, M. (2019). A parallel fortran framework for neural networks and deep learning. *SIGPLAN Fortran Forum*, *38*(1), 4–21. https://doi.org/10.1145/3323057.3323059

Dodge, J., Prewitt, T., Tachet des Combes, R., Odmark, E., Schwartz, R., Strubell, E., Luccioni, A. S., Smith, N. A., DeCario, N., & Buchanan, W. (2022). Measuring the carbon intensity of AI in cloud instances. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1877–1894. https://doi.org/10.1145/3531146.3533234

Freitag, C., Berners-Lee, M., Widdicks, K., Knowles, B., Blair, G. S., & Friday, A. (2021). The real climate and transformative impact of ICT: A critique of estimates, trends, and regulations. *Patterns*, *2*(9), 100340. https://doi.org/10.1016/j.patter.2021.100340

Giorgetta, M. A., Brokopf, R., Crueger, T., Esch, M., Fiedler, S., Helmert, J., Hohenegger, C., Kornblueh, L., Köhler, M., Manzini, E., Mauritsen, T., Nam, C., Raddatz, T., Rast, S., Reinert, D., Sakradzija, M., Schmidt, H., Schneck, R., Schnur, R., … Stevens, B. (2018). ICON-a, the atmosphere component of the ICON earth system model: I. Model description. *Journal of Advances in Modeling Earth Systems*, *10*(7), 1613–1637. https://doi.org/10.1029/2017MS001242

Gutmann, E. D., Hamman, Joseph. J., Clark, M. P., Eidhammer, T., Wood, A. W., & Arnold, J. R. (2022). En-GARD: A statistical downscaling framework to produce and test large ensembles of climate projections. *Journal of Hydrometeorology*, *23*(10), 1545–1561. https://doi.org/10.1175/JHM-D-21-0142.1

Jia, Y. (2024). Analysis of the impact of artificial intelligence on electricity consumption. *2024 3rd International Conference on Artificial Intelligence, Internet of Things and Cloud Computing Technology (AIoTC)*, 57–60. https://doi.org/10.1109/AIoTC63215.2024.10748289

Kedward, L. J., Aradi, B., Čertík, O., Curcic, M., Ehlert, S., Engel, P., Goswami, R., Hirsch, M., Lozada-Blanco, A., Magnin, V., Markus, A., Pagone, E., Pribec, I., Richardson, B., Snyder, H., Urban, J., & Vandenplas, J. (2022). The state of fortran. *Computing in Science & Engineering*, *24*(2), 63–72. https://doi.org/10.1109/MCSE.2022.3159862

Lang, S., Alexe, M., Chantry, M., Dramsch, J., Pinault, F., Raoult, B., Clare, M. C., Lessig, C., Maier-Gerber, M., Magnusson, L., & others. (2024). AIFS-ECMWF's data-driven forecasting system. *arXiv Preprint arXiv:2406.01465*. https://doi.org/10.48550/arXiv.2406.01465

Mutz, S. G. (2025). The effect of high-mountain asia topography on northern hemisphere atmospheric flow. *EGU General Assembly 2025, Vienna, Austria, 27 Apr–2 May 2025, EGU25-7283*. https://doi.org/10.5194/egusphere-egu25-7283

Mutz, S. G., & Ehlers, T. A. (2019). Detection and explanation of spatiotemporal patterns in late cenozoic palaeoclimate change relevant to earth surface processes. *Earth Surface Dynamics*, *7*(3), 663–679. https://doi.org/10.5194/esurf-7-663-2019

Mutz, S. G., Ehlers, T. A., Werner, M., Lohmann, G., Stepanek, C., & Li, J. (2018). Estimates of late cenozoic climate change relevant to earth surface processes in tectonically active orogens. *Earth Surface Dynamics*, *6*(2), 271–301. https://doi.org/10.5194/esurf-6-271-2018

Pereira, R., Couto, M., Ribeiro, F., Rua, R., Cunha, J., Fernandes, J. P., & Saraiva, J. (2021). Ranking programming languages by energy efficiency. *Science of Computer Programming*, *205*, 102609. https://doi.org/10.1016/j.scico.2021.102609

Taylor, N. T. (2024). ATHENA: A fortran package for neural networks. *Journal of Open Source Software*, *9*(99), 6492. https://doi.org/10.21105/joss.06492

Tomassetti, B., Verdecchia, M., & Giorgi, F. (2009). NN5: A neural network based approach for the downscaling of precipitation fields – model description and preliminary results. *Journal of Hydrology*, *367*(1), 14–26. https://doi.org/10.1016/j.jhydrol.2008.12.017