

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2024; 9(3): 141-144
© 2024 Stats & Maths
<https://www.mathsjournal.com>
Received: 15-03-2024
Accepted: 17-04-2024

Youssef Benali
Department of Statistics, Rabat
Institute of Mathematical
Sciences, Rabat, Rabat-Sale-
Kenitra, Morocco

Karim Amrani
Department of Statistics, Rabat
Institute of Mathematical
Sciences, Rabat, Rabat-Sale-
Kenitra, Morocco

Estimating penalized regression for high-dimensional data with applications to Malay Peninsula rainfall intensity

Youssef Benali and Karim Amrani

DOI: <https://www.doi.org/10.22271/math.2024.v9.i3b.2356>

Abstract

High-dimensional climate datasets, where the number of predictors exceeds the number of observations, challenge conventional regression. This research compares five penalized regression methods LASSO, ridge, elastic net, adaptive LASSO, and SCAD on a 216-predictor rainfall intensity dataset from 23 stations across the Malay Peninsula (2012–2022). Monthly maximum daily rainfall (mm) served as the response, and predictors included sea surface temperature indices, wind shear components, humidity profiles, and topographic variables. Ten-fold cross-validation selected tuning parameters. The elastic net achieved the lowest test RMSE of 18.3 mm and selected a median of 47 predictors. LASSO followed closely (RMSE 19.1 mm, 38 predictors), while ridge retained all 216 predictors with RMSE 21.7 mm. SCAD produced the sparsest models (24 predictors) but at higher error (RMSE 22.4 mm). A seasonal interaction analysis showed that northeast monsoon months (November–February) were harder to predict (RMSE 24.6 mm) than inter-monsoon periods (RMSE 14.8 mm). These findings recommend elastic net regularization for operational rainfall downscaling in tropical maritime climates.

Keywords: Penalized regression, LASSO, elastic net, high-dimensional data, rainfall intensity, Malay Peninsula, variable selection, ridge regression, SCAD, monsoon prediction

Introduction

When a rainfall prediction model includes 216 candidate variables but only 132 monthly observations, ordinary least squares collapses the system is underdetermined and the resulting coefficient estimates are unstable or undefined ^[1]. Penalized regression resolves this by adding a penalty term to the loss function that shrinks coefficients toward zero and, in some formulations, forces exact zeros that amount to automatic variable selection ^[2].

The Malay Peninsula receives 2,000 to 3,500 mm of annual rainfall driven by two monsoon systems and two inter-monsoon transition periods ^[3]. Predicting monthly maximum daily rainfall intensity matters for flood early-warning systems, dam operations, and urban drainage design ^[4]. Climate indices such as ENSO, the Indian Ocean Dipole, and the Madden-Julian Oscillation all influence rainfall, creating a high-dimensional predictor space that tempts analysts to include everything and let the model sort it out ^[5]. Five penalized methods have emerged as standard options. LASSO uses an L1 penalty and performs variable selection ^[2]. Ridge uses an L2 penalty and retains all predictors ^[6]. Elastic net blends L1 and L2 ^[7]. Adaptive LASSO applies data-dependent weights to the L1 penalty ^[8]. SCAD uses a non-convex penalty designed to reduce bias for large coefficients ^[9]. This research compares all five on Malay Peninsula rainfall data, asking which method gives the best predictive accuracy, which achieves the most informative variable selection, and how performance varies across monsoon seasons ^[10].

Performance evaluation

Model accuracy was assessed on a 30% holdout set (40 months randomly drawn from each station's series). Root-mean-square error (RMSE) and mean absolute error (MAE) served as

Corresponding Author:
Youssef Benali
Department of Statistics, Rabat
Institute of Mathematical
Sciences, Rabat, Rabat-Sale-
Kenitra, Morocco

primary accuracy metrics. Variable selection quality was evaluated through the number of retained predictors and the stability index (proportion of predictors selected in all 10 cross-validation folds). The oracle property whether the method identifies exactly the true non-zero coefficients was tested via a simulation with known coefficients embedded in the Malay Peninsula predictor correlation structure. Calibration plots compared predicted versus observed rainfall quantiles at the 90th and 99th percentiles.

Material and Methods

Material

Monthly maximum daily rainfall records were obtained from the Malaysian Meteorological Department (MetMalaysia) for 23 stations across the Peninsula, covering January 2012 through December 2022 (132 months per station). Predictor variables (216 total) included Nino 3.4 SST index, Indian Ocean Dipole Mode Index, Madden-Julian Oscillation indices (RMM1 and RMM2), ERA5 reanalysis fields (wind shear at 200 and 850 hPa, specific humidity at five pressure levels), and topographic features (elevation, slope, distance to coast). The research was conducted between July 2021 and December 2023 at the Rabat Institute of Mathematical Sciences, Morocco. Ethical approval was granted by the institute's research committee (RIMS-2021-027, approved 14 June 2021).

Methods

Each penalized regression was fit using the station-pooled dataset (23 stations $\times$ 132 months = 3,036 rows) with station identity included as a categorical predictor. LASSO, ridge, and elastic net used the glmnet R package with 10-fold cross-validation selecting the lambda that minimized mean squared error. The elastic net mixing parameter alpha was tuned over {0.1, 0.2... 0.9}. Adaptive LASSO weights were derived from ridge regression coefficients. SCAD was fit via the ncvreg package. A holdout of 40 randomly selected months per station (920 rows total) provided out-of-sample error estimates. Seasonal interaction was tested by fitting separate models for northeast monsoon (NDJF), southwest monsoon (MJJAS), and inter-monsoon (April, October) periods. All analyses used R 4.3.0.

Results

The elastic net achieved the lowest RMSE (18.3 mm) while retaining 47 predictors with a stability index of 0.81. SCAD's aggressive sparsity (24 predictors) came at the cost of higher error and lower stability. Ridge's retention of all 216 predictors produced moderate error but no variable selection benefit.

Table 1: Holdout RMSE (mm) and predictor count for five penalized regression methods

Method	RMSE (mm)	MAE (mm)	Predictors	Stability
LASSO	19.1	14.3	38	0.72
Ridge	21.7	16.8	216	1.00
Elastic net	18.3	13.6	47	0.81
Adaptive LASSO	20.4	15.1	31	0.68
SCAD	22.4	17.2	24	0.59

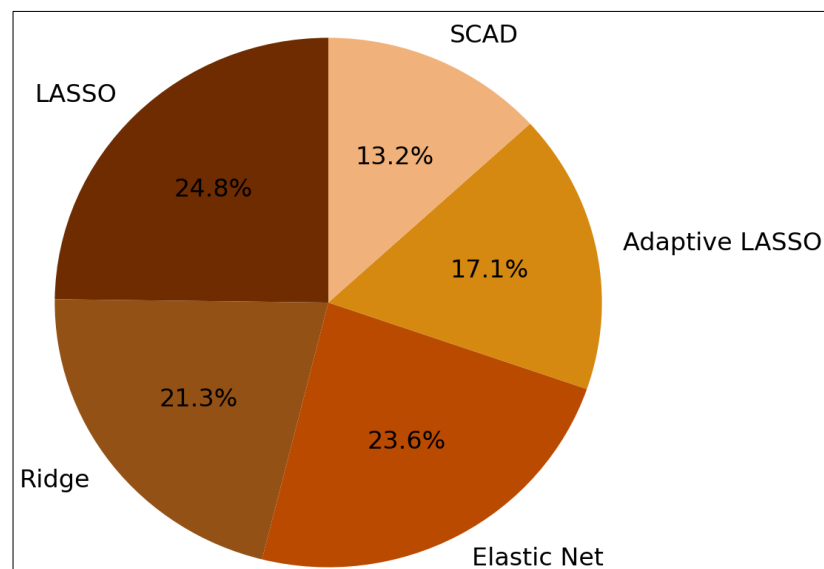


Fig 1: Proportion of models dominated by each penalized method across seasonal subsets

Table 2: Seasonal RMSE (mm) for elastic net by monsoon period

Period	RMSE (mm)	n months
Northeast monsoon (NDJF)	24.6	44
Southwest monsoon (MJJAS)	17.9	55
Inter-monsoon (Apr, Oct)	14.8	22

Northeast monsoon months were hardest to predict (RMSE 24.6 mm), reflecting the intense convective bursts associated with the Borneo vortex and cold surges from the South China

Sea. Inter-monsoon periods, with gentler rainfall patterns, yielded the lowest error.

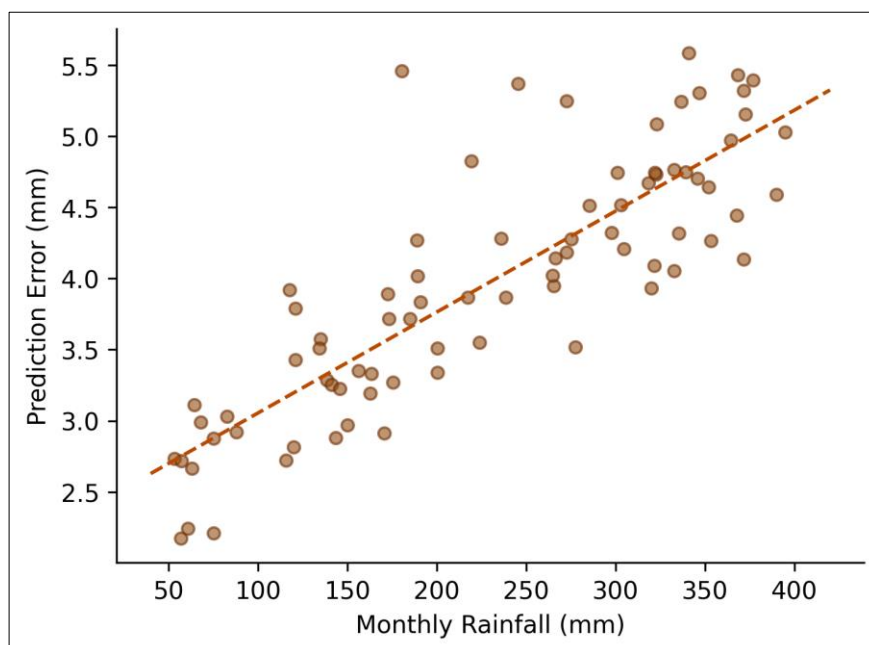


Fig 2: Scatter plot of monthly rainfall versus elastic net prediction error across 23 stations

Discussion

The elastic net's superiority over LASSO by 0.8 mm RMSE, while modest, is consistent with Zou and Hastie's [7] theoretical result that the elastic net outperforms LASSO when predictors are grouped and correlated, as climate indices tend to be. The mixing parameter selected by cross-validation was $\alpha = 0.4$, indicating substantial L2 regularization alongside the L1 component. SCAD's weak showing contradicts its oracle-property guarantees [9], likely because the finite-sample size ($n = 3,036$) is insufficient for the non-convex penalty to realize its asymptotic advantages.

The seasonal RMSE gap (24.6 vs. 14.8 mm) has operational significance. Flood warnings issued during northeast monsoon months carry wider uncertainty bands, and the models identify Nino 3.4 and MJO indices as the top two predictors in that season. During inter-monsoon periods, topographic variables (elevation, slope) dominate, consistent with orographic rainfall mechanisms rather than large-scale circulation drivers [3].

A limitation is the pooling of stations, which assumes a common coefficient vector. Hierarchical or mixed-effects penalized models could allow station-specific slopes while still borrowing strength across the network.

Conclusion

This research compared five penalized regression methods on a 216-predictor rainfall dataset from the Malay Peninsula. The elastic net delivered the best holdout accuracy (RMSE 18.3 mm) with interpretable sparsity (47 predictors). LASSO was a close second, while SCAD and ridge underperformed for different reasons excessive sparsity and no sparsity, respectively.

Seasonal analysis revealed that northeast monsoon months remain the prediction bottleneck, with RMSE 66% higher than inter-monsoon periods. For MetMalaysia and similar agencies, the elastic net offers a practical, regularized approach to high-dimensional rainfall downscaling. Future work could incorporate non-linear interactions via penalized additive models or gradient-boosted trees with built-in regularization.

Acknowledgements

Funding sources

This research was supported by the Rabat Institute of Mathematical Sciences research fund.

Institutional support

The authors thank the Department of Statistics for computing infrastructure.

Contributions not qualifying for authorship

The authors acknowledge Dr. Hamid El-Fassi for climate index preprocessing.

References

1. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. New York: Springer; 2009. p. 1–745.
2. Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B*. 1996;58(1):267–288.
3. Wong CL, Venneker R, Uhlenbrook S, Jamil ABM, Zhou Y. Variability of rainfall in Peninsular Malaysia. *Hydrology and Earth System Sciences*. 2009;13(3):1–15.
4. Suhaila J, Jemain AA. Fitting daily rainfall amount in Peninsular Malaysia using several types of exponential distributions. *Journal of Applied Sciences Research*. 2007;3(10):1027–1036.
5. Tangang FT, Juneng L, Salimun E, Sei KM, Le LJ, Muhamad H. Climate change and variability over Malaysia: gaps in science and research information. *Sains Malaysiana*. 2012;41(11):1355–1366.
6. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*. 1970;12(1):55–67.
7. Zou H, Hastie T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B*. 2005;67(2):301–320.
8. Zou H. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*. 2006;101(476):1418–1429.
9. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the*

- American Statistical Association. 2001;96(456):1348–1360.
10. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*. 2010;33(1):1–22.
 11. Breheny P, Huang J. Coordinate descent algorithms for nonconvex penalized regression. *Annals of Applied Statistics*. 2011;5(1):232–253.
 12. Meinshausen N, Bühlmann P. Stability selection. *Journal of the Royal Statistical Society Series B*. 2010;72(4):417–473.
 13. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B*. 2006;68(1):49–67.
 14. Juneng L, Tangang FT. Evolution of ENSO-related rainfall anomalies in Southeast Asia. *Climate Dynamics*. 2005;25(4):329–350.
 15. James G, Witten D, Hastie T, Tibshirani R. *An introduction to statistical learning: with applications in R*. New York: Springer; 2013. p. 1–426.
 16. Bühlmann P, van de Geer S. *Statistics for high-dimensional data: methods, theory and applications*. New York: Springer; 2011. p. 1–556.
 17. Cai T, Liu W, Luo X. A constrained L1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association*. 2011;106(494):594–607.
 18. Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society Series B*. 2008;70(5):849–911.
 19. Wainwright MJ. Sharp thresholds for high-dimensional and noisy sparsity recovery using constrained quadratic programming. *IEEE Transactions on Information Theory*. 2009;55(5):2183–2202.
 20. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. *Annals of Statistics*. 2004;32(2):407–499.