DTU Electrical Engineering
Department of Electrical Engineering

DTU

Regularized Nonhomogeneous Regression for Predictor Selection in Ensemble Post-Processing

Jakob W. Messner (jwmm@elektro.dtu.dk), Georg J. Mayr, and Achim Zeileis

Introduction

Ensemble Forecasts:

- often biased and uncalibrated

→ statistical post-processing

Nonhomogeneous Gaussian Regression (NGR; Gneiting et al., 2005):

- predictive Gaussian distribution (temperature T)
- mean is a function of the ensemble mean (m)
- variance is a function of the ensemble variance (s^2)

$$T \sim N(\mu, \sigma^2)$$
$$\mu = \beta_0 + \beta_1 m$$
$$\log(\sigma) = \gamma_0 + \gamma_1 \log(s)$$

- coefficients β_0, β_1, γ_0, γ_1 are estimated by maximizing the log-likelihood:

$$\sum \log\left[\frac{1}{\sigma}\Phi\left(\frac{T-\mu}{\sigma}\right)\right] \quad (1)$$

Predictor variables:

- usually only temperature ensemble forecasts (m, s)
- further potential predictor variables:
  - ensemble predictions of other variables (e.g., pressure, cloud cover, ...)
  - predictions from other numerical models or weather centers
  - current observations
  - transformations and interactions,
  - ...
- extend NGR for multiple inputs $x_1, x_2, \ldots, x_J, z_1, z_2, \ldots, z_K$:

$$\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_J x_J$$
$$\log(\sigma) = \gamma_0 + \gamma_1 z_1 + \gamma_2 z_2 + \ldots + \gamma_K z_K$$

Problem:

- too many inputs can lead to overfitting and decreased forecast performance
- how to select best set of predictor variables?

→ **automatic predictor selection**

Data

- 18UTC – 06UTC 2 meter minimum temperatures in Vienna
- ECMWF +18–30 hours ensemble forecasts 2011 – 2015
- removed seasonality of forecasts and observations with standardized anomalies (see also poster X4.204)
- means, maxima, and mimima of forecasts over regarded time window
- last available observation

→ 307 potential input variables

- training: 2011–2014, testing: 2015

Regularized Regression

Two different approaches to prevent overfitting:

Gradient boosting (Messner et al., 2017):

- alternative iterative optimization algorithm to maximize (1)
- initialize all coefficients with zero
- in each iteration slightly **update only the one coefficient that improves the current fit most**

→ if not run until convergence, **only important inputs have non-zero coefficients**

- select optimum stopping iteration by cross validation

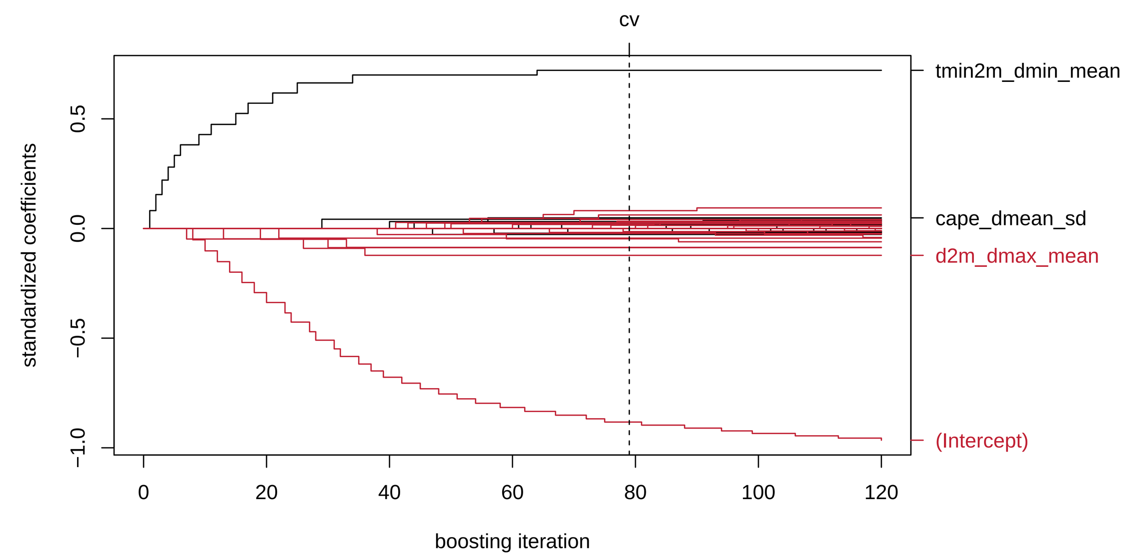

Figure 1: Boosting coefficients for different stopping iterations. Coefficients for μ are shown as black lines and for $\log(\sigma)$ as red lines. The optimum stopping iteration from cross validation is shown as dashed vertical line. The most important coefficients are labeled.

LASSO regularization:

- maximize penalized likelihood:

$$\sum \log\left[\frac{1}{\sigma}\Phi\left(\frac{T-\mu}{\sigma}\right)\right] + \lambda\left(\sum_{j=1}^{J}|\beta_j| + \sum_{k=1}^{K}|\gamma_k|\right)$$

- **penalizes absolute coefficient** values

→ **coefficients** of **unimportant** variables are shrunk to zero

- select optimum penalization parameter λ by cross validation

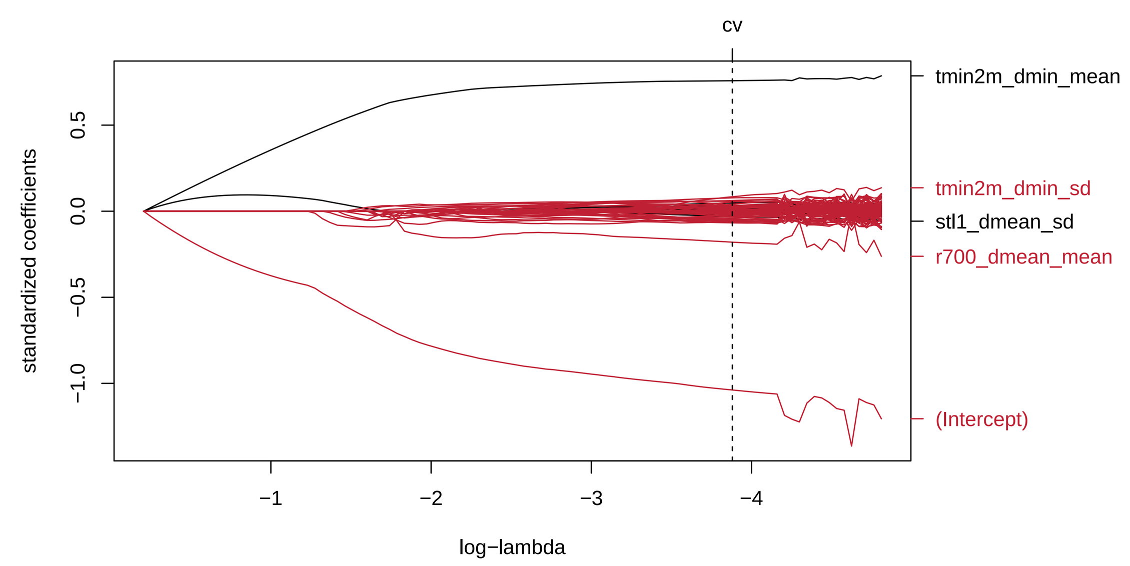

Figure 2: Same as Figure 1 but for LASSO with different values of λ.

Results

Selected predictor variables:

boosting		LASSO	
μ	$\log(\sigma)$	μ	$\log(\sigma)$
tmin2m_dmin_mean	d2m_dmax_mean	tmin2m_dmin_mean	r700_dmean_mean
cape_dmean_sd	r700_dmax_mean	stl1_dmean_sd	tmin2m_dmin_sd
stl1_dmin_mean	d700_dmin_sd	r850_dmax_sd	w500_dmin_sd
q1000_dmax_mean	fg10m_dmean_sd	d2m_dmax_mean	w850_dmean_sd
...	...	...	...
total #: 12	total #: 17	total #: 17	total #: 61

Table 1: Selected input variables by boosting and LASSO. Variable names have syntax *name_aggregation_statistic*. *dmin*, *dmin*, and *dmean* denote the minimum, maximum, and mean of the forecasts between +18 and +30 respectively. *mean* and *sd* are the ensemble mean and log-standard deviation respectively.

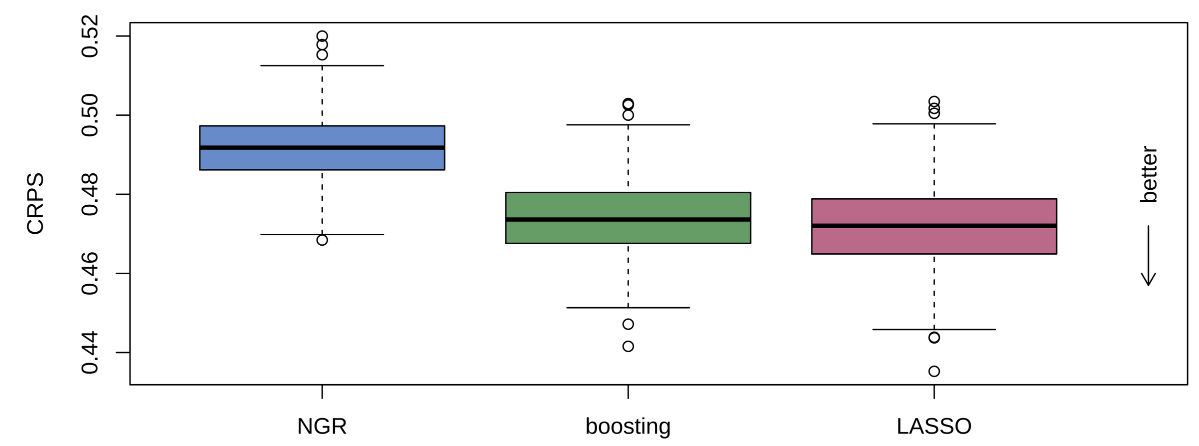

Figure 3: Continuous ranked probability score (CRPS) of NGR (only minimum temperature ensemble as input), gradient boosting, and LASSO regularization

Summary

Regularized nonhomogeneous regression:

- automatically selects best set of variables

→ clearly improved forecast performance

- boosting and LASSO select different variable sets
- highly correlated inputs → similar performance
- LASSO: computationally more efficient
- boosting: more flexible

CRAN R-package crch:

- gradient boosting already implemented
- coordinate descent algorithm for LASSO paths coming soon

References:

Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, **133 (5)**, 1098–1118.

Messner, J. W., G. J. Mayr, and A. Zeileis, 2017: Nonhomogeneous boosting for predictor selection in ensemble postprocessing. *Monthly Weather Review*, **145 (1)**, 137–147.