Boosting multivariate Gaussian models for probabilistic temperature forecasts

Thorsten Simon[1], Nikolaus Umlauf[2], Georg J. Mayr[1], Achim Zeileis[2]

[1] Institute of Atmospheric and Cryospheric Sciences, University of Innsbruck, Austria
[2] Department of Statistics, University of Innsbruck, Austria

E-mail for correspondence: thorsten.simon@uibk.ac.at

Abstract: Many weather prediction tasks are multivariate problems, e.g., predicting several quantities (such as temperature and precipitation) for a particular time or predicting a single quantity over time. In the latter case, a state-of-the-art method is to fit several marginal prediction models and then combine them using ensemble copula coupling (ECC). As an alternative approach, we propose to fit a single multivariate Gaussian model where all parameters (means, variances, and correlations) can be expressed by additive models. For estimation of the resulting large number of parameters a gradient boosting algorithm is employed. Results for a case study show equal performance with respect to marginal predictive distributions and better performance with respect to the full multivariate distribution in comparison to nonhomogeneous Gaussian regressions (NGRs) combined with ECC.

Keywords: boosting; additive models; multivariate Gaussian; weather prediction.

1 Introduction

To obtain calibrated weather forecasts for several lead times, e.g., predicting temperature 12 h, 36 h, 60 h, ... in advance, the output of numerical weather prediction (NWP) ensemble systems is often postprocessed using statistical models. One popular choice for temperature forecasts is nonhomogeneous Gaussian regression combined with ensemble copula coupling (NGR-ECC, see Schefzik *et al.*, 2013). In a first step (NGR) linear models are employed for the location and scale parameter of a Gaussian distribution for several lead times separately. In a second

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

step (ECC) the predicted quantiles are reordered according to the raw ensemble output, in order to preserve the covariance structure of the NWP ensemble.

This study aims at extending NGR in the following way: Estimating predictive distributions for several lead times and their correlations simultaneously. The location, scale and correlation parameters of a multivariate Gaussian (MVN) will be expressed by GAM-type additive predictors η. Estimating multivariate distributions with these specifications is a complex task for dimensions higher than 2 (Klein *et al.*, 2015). Gradient boosting can offer an attractive solution to fit a MVN with additive predictors for all parameters as only first derivatives of the log-likelihood with respect the predictors are required. Another benefit of boosting is that selection and shrinkage of coefficients can be obtained.

2 Methods

The log-likelihood of the multivariate Gaussian for a k-dimensional observation $\mathbf{y} = (y_1, y_2, \ldots, y_k)^{\mathrm{T}}$ can be parameterized as follows,

$$l(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathbf{y}) = -\frac{k}{2}\log(2\pi) - \frac{1}{2}\log(|\boldsymbol{\Sigma}|) - \frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu}),$$

where $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_k)^{\mathrm{T}}$ denotes the vector of the mean parameters and $\boldsymbol{\Sigma}$ denotes the covariance matrix. The latter can be decomposed into $\boldsymbol{\Sigma} = \mathbf{D}\boldsymbol{\Omega}\mathbf{D}$, where $\mathbf{D}$ is a diagonal matrix with the standard deviations $\sigma_1, \sigma_2, \ldots, \sigma_k$ on the diagonal, and $\boldsymbol{\Omega}$ is the correlation matrix with the elements ρ_{ij}.

The parameters μ_i, σ_i and ρ_{ij} are linked to their predictors $\eta_{\mu i}$, $\eta_{\sigma i}$ and $\eta_{\rho ij}$ by the identity, the log and the rhogit function, respectively. The partial derivatives with respect to the predictors of μ_i and σ_i are,

$$\frac{\partial l}{\partial \eta_{\mu i}} = \sum_{j=1}^{k} \varsigma_{ij}(y_j - \mu_j) \quad \text{and} \quad \frac{\partial l}{\partial \eta_{\sigma i}} = -1 + \tilde{y}_i \sum_{j=1}^{k} \omega_{ij}\tilde{y}_j,$$

where ς_{ij} and ω_{ij} denote the elements of the inverse covariance matrix $\boldsymbol{\Sigma}^{-1}$ and inverse correlation matrix $\boldsymbol{\Omega}^{-1}$, respectively. Additionally, $\tilde{y}_i = (y_i - \mu_i)/\sigma_i$. The partial derivative with respect to the predictor of ρ_{ij} is

$$\frac{\partial l}{\partial \eta_{\rho ij}} = \left[-\frac{1}{2}\omega_{ij} + \frac{1}{2}\left(\sum_{m=1}^{k}\omega_{im}\tilde{y}_m\right)\left(\sum_{m=1}^{k}\omega_{jm}\tilde{y}_m\right)\right] \times \left(1+\eta_{\rho ij}^2\right)^{-\frac{3}{2}}.$$

In order to fit the model a gradient boosting algorithm is applied as implemented by Umlauf *et al.* (2017). The algorithm is an iterative procedure. The number of iterations $m_{\max}$ has to be defined in advance. In each step the coefficients of the term which would contribute most to maximizing the log-likelihood are updated by the proportion ν of the local estimate of the coefficients. Thus, the boosting algorithm results in $m_{\max}$ distinct sets of coefficients. The optimal set of coefficients is selected by out-of-sample validation. A generic description of gradient boosting is given by Mayr *et al.* (2012).

3 Application in weather prediction

A case study is presented for predicting temperature in Innsbruck, Austria (47.260°N, 11.357°E). Six lead times are considered, 12 h, 36 h, $\dots$, 132 h in advance. Data is on hand from January 2011 to December 2016 leading to a sample size of roughly 2150. The ensemble predictions of the European Centre for Medium-Range Weather Forecasts (ECMWF) serve as NWP input. The additive predictors are declared as follows,

$$\eta_{\mu i} = \alpha_{i,0} + \alpha_{i,1} * \mathrm{mean}(\mathtt{ens}_i) + f_{i,cc}(\mathtt{yearday}),$$

$$\eta_{\sigma i} = \beta_{i,0} + \beta_{i,1} * \log(\mathrm{sd}(\mathtt{ens}_i)) + g_{i,cc}(\mathtt{yearday}),$$

$$\eta_{\rho ij} = \gamma_{ij,0} + \gamma_{ij,1} * \mathrm{cor}(\mathtt{ens}_i, \mathtt{ens}_j),$$

where indices $i, j \in \{1, 2, \dots, 6\}$ refer to the lead times 12 h, 36 h, $\dots$, 132 h, respectively. $\mathtt{ens}_i$ denotes the raw ECMWF ensemble temperature forecast, and $f_{i,cc}(\mathtt{yearday})$ and $g_{i,cc}(\mathtt{yearday})$ are cyclic smooth functions modeled by splines to account for annual cycles. The model is trained on the period 2011–2015. Validation on the data of year 2016 leads to an optimal set of coefficients after $m_{\mathrm{opt}} = 5000$ iterations where $\nu = 0.05$.

Figure 1 displays the fitted nonlinear functions $f_{i,cc}(\cdot)$ and $g_{i,cc}(\cdot)$, which contribute to the additive predictors $\eta_{\mu i}$ and $\eta_{\sigma i}$, respectively.

The coefficients describing $f_{i,cc}(\cdot)$ were not selected within the first 5000 iterations. Thus, $f_{i,cc}(\cdot)$ remains flat. This suggests that the bias between the model temperature and observations is constant throughout the year.

$g_{i,cc}(\cdot)$ (Fig. 1, right) contributes to the predictor of σ_i on the log-scale. $g_{i,cc}(\cdot)$ reveals an annual cycle with two peaks. One occurs in January and one in June/July. The fitted effects for all lead times vary only slightly among eachother.

However, the main focus of this study is to model the correlation structure of temperature between the lead times, 12 h, 36 h, $\dots$, 132 h. Figure 2 summarizes the distribution of the fitted correlations. All parameters on the first diagonal next to the main diagonal of Ω vary around 0.74. The parameters on the second diagonal vary around 0.46. Thus, the fitted correlation matrixes exhibit a structure similar to a symmetric Toeplitz matrix or or even close to the correlation matrix of a AR-process.

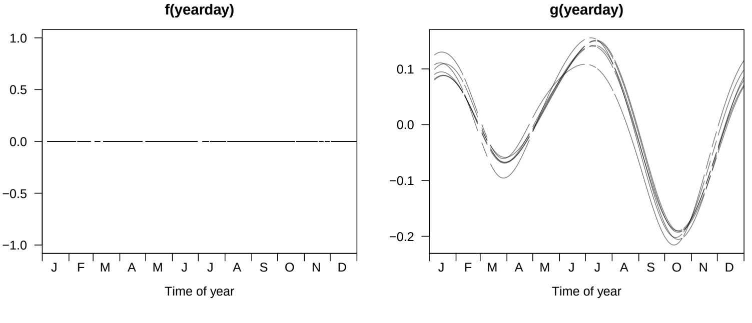

FIGURE 1. Nonlinear effects for all lead times. **Left:** $f_{i,cc}(\mathtt{yearday})$ contributing to $\eta_{\mu i}$. **Right:** $g_{i,cc}(\mathtt{yearday})$ contributing to $\eta_{\sigma i}$.

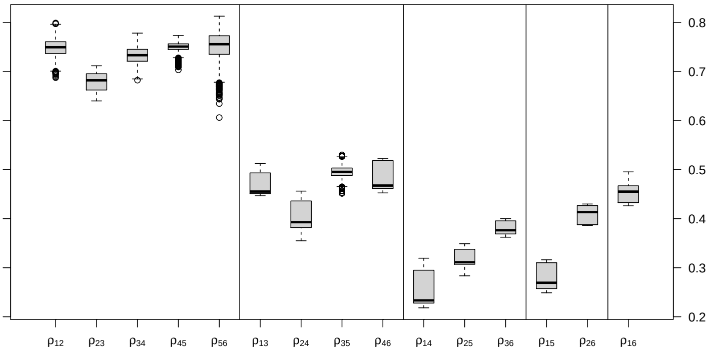

FIGURE 2. Fitted correlation coefficients ρ_{ij} for all days in 2016. Each box-and-whisker plot indicates the distribution of one correlation parameter over all sample cases.

The range indicated by the box-and-whisker plots (Fig. 2) suggests that the values of the intercepts $\gamma_{ij,0}$ are more important for determining the structure of the correlation matrix than the coefficients of the linear terms, $\gamma_{ij,1}$. Thus, only a small part of the correlation structure modeled by the numerical ensemble can be retained by the statistical model.

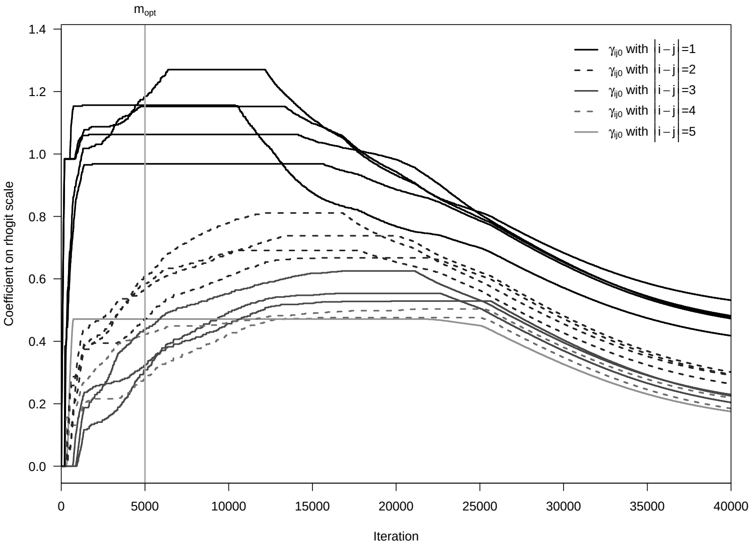

FIGURE 3. Boosting paths for the coefficients of the predictors of the correlation parameters on the rhogit scale.

Figure 3 illustrates how the values of the coefficients $\gamma_{ij,0}$ develop during the iterative boosting procedure. The intercepts $\gamma_{ij,0}$ are selected before the $\gamma_{ij,1}$ are selected. However, after 15000–25000 iterations the values of the intercepts start dropping, which might be caused by an overfitting of the location parameter μ_i.

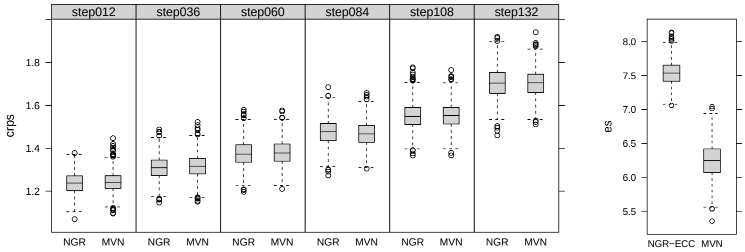

FIGURE 4. Out-of-sample scoring. **Left:** continuous rank probability score (CRPS) for univariate NGR models and marginal predictive distributions of the boosted MVN model. **Right:** energy score (ES) of the NGR-ECC and the boosted MVN.

Figure 4 compares the performance of the proposed method to the performance obtained by the state-of-the-art method NGR-ECC. The NGR models are also fitted via boosting. The marginal predictive distributions of the boosted MVN model are compared to NGR models fitted for every single lead time separately via the continuous ranked probability score (CRPS, Gneiting and Raftery, 2007). The left panel of Figure 4 reveals that the two models perform equally well with respect to their marginal distributions. The multivariate performance of the models is assessed in terms of the energy score (ES, Gneiting and Raftery, 2007). The boosted MVN outperforms the NGR-ECC in our case (Figure 4, right).

4 Conclusions

This study suggests to fit multivariate Gaussian distributions via gradient boosting, where additive predictors can be defined for all location, scale and correlation parameters. A case study in the field of weather forecasting shows promising results.

Further investigations are needed to fully understand the potential of boosting MVN with additive predictors. There are alternative ways to parameterize the correlation matrix, i.e., modeling the parameters of its inverse or its Cholesky decomposition (Pourahmadi, 2011).

In the present case one could assume an AR-process among the response variable as they are temporally ordered. This kind of parameterization is implicated by the findings in this study (cf. Fig. 2). However, more research is needed to examine whether a parsimonious or flexible parameterization is superior in this kind of application.

Depending on the problem changing the parameterization can have an effect on the required iterations until convergence of the boosting algorithm, and could yield different results when a shrunken version of the model is selected.

Acknowledgments: We acknowledge the funding by the Austrian Research Promotion Agency (FFG) project LightningPredict (Grant No. 846620).

References

Gneiting, T. and Raftery, A.E. (2007). Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Association*, **102**, no. 477, 359 – 378.

Klein, N., Kneib, T., Klasen, S. and Lang, S. (2015). Bayesian structured additive distributional regression for multivariate responses. *Journal of the Royal Statistical Society, Series C*, **64**, Part 4, 569 – 591.

Mayr, A., Fenske, N., Hofner, B., Kneib, T. and Schmid, M. (2012). Generalized additive models for location, scale and shape for high dimensional data—a flexible approach based on boosting. *Journal of the Royal Statistical Society, Series C*, **61**, Part 3, 403 – 427.

Pourahmadi, M. (2011). Covariance estimation: The GLM and regularization perspectives. *Statistical Science*, **26**, no. 3, 369 – 387.

Schefzik, R., Thorarinsdottir, T.L., and Gneiting, T. (2013). Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical Science*, **28**, no. 4, 616 – 640.

Umlauf, N., Klein, N., and Zeileis, A. (2017). BAMLSS: Bayesian additive models for location, scale and shape (and beyond). *Working Papers, Faculty of Economics and Statistics*, University of Innsbruck.

Generalization of the Whittle likelihood for nonparametric spectral density estimation

Claudia Kirch[1], Matthew Edwards[2], Alexander Meier[1], Renate Meyer[2]

[1] Otto-von-Guericke University, Magdeburg, Germany
[2] University of Auckland, New Zealand

E-mail for correspondence: `meyer@stat.auckland.ac.nz`

Abstract: Most nonparametric Bayesian approaches use Whittle's likelihood to estimate the spectral density as the main nonparametric characteristic of stationary time series, as e.g. Choudhuri et al. (2004) and Rosen et al (2012). But as shown in Contreras-Cristan et al. (2006), the loss of efficiency of the nonparametric approach using Whittle's likelihood can be substantial. We show that the Whittle likelihood can be regarded as a special case of a nonparametrically corrected parametric likelihood which gives rise to a robust and more efficient Bayesian nonparametric spectral density estimate based on a generalized Whittle likelihood. Its frequentist properties are investigated in a simulation study. Applications to LIGO gravitational wave data and the El Niño Southern Oscillation phenomenon will be described.

Keywords: Bayesian nonparametrics; stationary time series; spectral density estimation; Bernstein polynomial prior; gravitational waves.

1 Introduction

Most Bayesian nonparametric approaches to time series analysis are based on Whittle's likelihood approximation (Whittle, 1957), as e.g. Choudhuri et al. (2004) and Rosen et al. (2012). We will show that the Whittle likelihood can be regarded as the likelihood of a parametric working model, namely iid Gaussian, which has been nonparametrically corrected in the frequency domain. Borrowing an idea from a periodogram bootstrap for time series in Kreiss and Paparoditis (2003), we propose a generalization of the Whittle likelihood that uses a more realistic parametric working model, e.g. an AR(p) model, again nonparametrically corrected in the frequency domain and suggest a Bayesian semi-parametric

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.