Journal of Econometrics 222 (2021) 579–600

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Max-linear regression models with regularization☆

Qiurong Cui, Yuqing Xu, Zhengjun Zhang *, Vincent Chan

Department of Statistics, University of Wisconsin-Madison, WI 53706, United States of America

ARTICLE INFO

Article history:
Available online 1 August 2020

Keywords:
Max-linear regression
Regularization
EM algorithm
Econometric model
Business statistics

ABSTRACT

Motivated by the newly developed max-linear competing copula factor models and max-stable nonlinear time series models, we propose a new class of max-linear regression models to take advantages of easy interpretable features embedded in linear regression models. It can be seen that linear relation is a special case of max-linear relation. We develop an EM algorithm based maximum likelihood estimation procedure. The consistency and asymptotics of the estimators for parameters are proved. To advance max-linear models to deal with high dimensional predictors, we adopt the common strategy of regularization in the high dimensional regression literature. We demonstrate the broad applicability of max-linear models using simulation examples and real applications in econometric and business modeling. The results, in terms of predictability, show a significant improvement compared with solely using regular regression models and other existing machine learning methods. The results enhance our understanding of the relationship between the response variable and the predictors, and among the predictors as well.

1. Introduction

In econometric modeling and many other applied and theoretical areas, regression is a fundamental statistical method for studying the relationships among variables, and has received tremendous attention over the past decades. More specifically, we are given a set of observations $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$, where the Y_i's are the dependent (response) variables and the $\mathbf{X}_i$'s are the independent (predictor) variables. The relationship of dependent and independent variables may be written as $Y_i = f(\mathbf{X}_i, \epsilon_i)$ (inseparable) or $Y_i = f(\mathbf{X}_i) + \epsilon_i$ (separable), where ϵ_i's are random noises or measurement errors. In linear regression, independent variables are connected with linear relationships and linked to the response. However, in many cases, a linear function is inadequate to describe the relations between a response and its explanatory features. It is often assumed that in the separable form, f is piece-wise linear, non-linear, or nonparametric. The literature is rather rich, and we do not intend to detail all developments in this area as our focus is within the domain of competing contributions of structured linear relations. Interested readers are referred to Bates and Watts (1988), Wahba (1990), Wand and Jones (1994), Fan and Gijbles (1996), Tanizaki and Mariano (1998), Patton and Timmermann (2007), Henderson and Parmeter (2015), Fan et al. (2019), Linton and Xiao (2019), amongst many others.

In probability theory, stability refers to the property that a linear combination of independent copies of a random variable has the same distribution as the variable, up to location and scale parameters. Distributions with such property

☆ We thank the editor Oliver Linton and one anonymous referee for their insightful comments and suggestions that substantially improve the quality and the presentation of the paper.

* Corresponding author.
E-mail address: zjz@stat.wisc.edu (Z. Zhang).

https://doi.org/10.1016/j.jeconom.2020.07.017
0304-4076/

are often referred as "stable distributions", and a number of mathematical results can be derived for such distributions. Beyond linearity, max-stable law (Pancheva, 1989) refers to that the maximum of independent copies of a random variable has the same distribution as the variable, up to location and scale parameters. The stable law and the max-stable law motivated use of max-linear structure instead of linear operator in many aspects including factor models, e.g., Zhang (2009) and Cui and Zhang (2018), nonlinear correlation and dependence, e.g., Zhang (2008), Zhang et al. (2011, 2017), and other dependence models (Zhang et al., 2016; Zhang and Zhu, 2016). In the time series literature, max-linear models have been successfully developed for asymptotically (in)dependent random variables across time, e.g., Heffernan et al. (2007), Naveau et al. (2011) amongst others. For resource competing measures, we refer to Malinowski et al. (2016).

In this paper, we consider max-linear model replacing the linear combination in the linear regression model by a max-linear structure. More specifically, in a max-linear regression model, when independent variables are included in a certain group, they are connected with linear relationship within groups forming factors, and the max relationship between groups forming competing factors. This max-linear structure allows non-linear relationship between dependent variables and independent variables using competing structures between different groups of features while keeping linear relationship within groups to achieve model simplicity and interpretability. The max-linear model is closely related to the max-linear competing factor models in Cui and Zhang (2018) which studies the dependence among random variables with latent/hidden idiosyncratic factors and random loading coefficients. Replacing the random loading coefficients in the max-linear competing factor models by functions of linear combinations of predictors introduces a new class of models, i.e., max-linear regression models or just max-linear models. As a result, the inference procedure for the new model has to be developed. In this paper, we develop estimation procedures not only for regular (low or fixed) dimension settings but also for high-dimensional settings.

This paper makes the following contributions to the econometric and general literature.

- Compared with the existing nonlinear models with/without Gaussian errors, e.g., models in the references mentioned earlier, the new model opens a new direction of non-linearly structured regression models using a simple max structure, yet encompassing simple linear models as its special cases. The simple max structure enables us to easily interpret the contribution of each predictor to the response. It makes modeling non-Gaussian response variables possible through Gaussian measurement errors without assuming any other specific non-Gaussian error distributions. For example, the max-linear models referred earlier in time series settings assume innovations to be extreme value type random variables. Tanizaki and Mariano (1998) directly use non-Gaussian distributed random error term. In terms of nonlinearity, the competing factor models in Cui and Zhang (2018) do not involve predictors (independent variables), but rather focus on the copula structure, i.e., not yet in a multiple regression setting. Other models that deal with nonlinearity, e.g., Bates and Watts (1988), Wahba (1990), Wand and Jones (1994), Fan and Gijbles (1996), Tanizaki and Mariano (1998), Patton and Timmermann (2007), Henderson and Parmeter (2015), Fan et al. (2019), Linton and Xiao (2019), also have specific structure assumptions. We use max structure to introduce non-Gaussianity and asymmetry while the distribution family of random errors is still normal. The max structure keeps a relatively explicit and simple format of the parametric linear regression model comparing with other work on asymmetric and non-Gaussian models and introduces the idea of factor competition into regression. Due to its great interpretability, the max linear structure can be widely applicable to many research problems, e.g., logistic regressions in classifications, competing risks in survival analysis, amongst many research and application areas.
- The model is proven more suitable than black-box machine learning models such as random forest in some of the econometric examples, yet still maintains the interpretability of simple linear models, as it is a natural and explicit extension of linear models but has far better prediction power and flexibility due to the max operator.
- The max linear regression framework bridges separability and inseparability in econometric models. On one hand, within each group/factor, we have a simple linear combination of predictors and an error term, i.e., a separable form of $f(\mathbf{X}_i)+\epsilon_i$, while the max linear regression model enforces an inseparable model which can be generally written as $Y_i = f(\mathbf{X}_i, \epsilon_i)$, but it degenerates to linear model (separable) when only one competing factor actually dominates. On the other hand, if we assume that the error terms in all competing factors to be identical, the max-linear regression model again turns to a separable model. These features will be further discussed in latter sections.

The rest of the paper is organized as follows. In Section 2, we introduce the max-linear model. In Section 3, we apply the Expectation–Maximization (EM) algorithm to find maximum likelihood estimator and provide theory for the consistency and asymptotic normality of the estimator. We then introduce penalization to invite model sparsity and also provide unbiased prediction estimator. In Section 4, we carry out simulation studies and compare our model to other state-of-the-art models, e.g., group lasso, random forest, and general linear regression. In Section 5, we apply our method to financial and economic datasets and the FIFA players' market values dataset with comparison to other methods. Technical proofs of eleven lemmas and two main theorems are provided in Appendix (online supplementary file).

2. Max-linear regression

We begin by reviewing the max-linear competing factor model in Cui and Zhang (2018). Suppose Z_1, Z_2 and Z_3 are independent and identically distributed Fréchet random variables with cumulative distribution function

$$P(X \le x) = \exp\{-(\frac{x-m}{s})^{-\alpha}\} \text{ if } x > m,$$

where $\alpha > 0$ is a shape parameter. m and $s > 0$ are location and scale parameters. Further suppose a_{i1}, a_{i2}, a_{i3} are positive coefficients. The association between Y_i's are implicitly derived from the max-linear representation

$$Y_i = \max(a_{i1}Z_1,\ a_{i2}Z_2,\ a_{i3}Z_3), \qquad i = 1, 2, \ldots, n, \tag{2.1}$$

where n is the dimension of random vector Y. Taking logarithm transformation of both sides of (2.1), we get

$$\log Y_i = \max(\log a_{i1} + \log Z_1,\ \log a_{i2} + \log Z_2,\ \log a_{i3} + \log Z_3), \tag{2.2}$$

for all $i = 1, 2, \ldots, n$. Further assume that a_{i1}, a_{i2} and a_{i3} depend only on all predictor variables $\mathbf{X}$, then

$$\log Y = \max\left(f_1(\mathbf{X}) + \log Z_1,\ f_2(\mathbf{X}) + \log Z_2,\ f_3(\mathbf{X}) + \log Z_3\right), \tag{2.3}$$

where f_1, f_2 and f_3 are measurable functions. In Eq. (2.3), three factors are competing with each other, and each factor is connected to predictor variables $\mathbf{X}$. This observation motivates us to apply max structure in modeling the relationship between the dependent variable and independent variables, especially when independent variables have grouped structures. Simplifying f_j as a linear combination of the predictor variables leads to the max-linear regression model.

Now we formally specify the max-linear regression model as follows.

Assume that the response variable Y_i is max-linearly associated with L linearly combined competing factors, i.e., the maximum of linear combinations of each group of predictors.

$$Y_i = \max\left\{\log\alpha_1 + \mathbf{X}_{i\cdot}^{(1)}\beta_1 + \epsilon_{i1}, \ldots, \log\alpha_L + \mathbf{X}_{i\cdot}^{(L)}\beta_L + \epsilon_{iL}\right\},\ i = 1, \ldots, n, \tag{2.4}$$

with the following assumptions and notations:

1. $(\mathbf{X}^{(l)})_{n\times p_l}$ is the predictor matrix corresponding to the lth factor, $1 \le l \le L$. $\beta_l \in \mathbb{R}^{p_l}$ is the vector of coefficient for the lth factor. Let $\boldsymbol{\beta} = (\beta_1^T, \ldots, \beta_L^T)^T$, and $\mathbf{X} = (\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(L)})_{n\times\sum_{l=1}^{L} p_l}$. Denote $\mathbf{X}_{i\cdot}^{(l)}$ as the row vector from $(\mathbf{X}^{(l)})_{n\times p_l}$, referring to the predictor variables in factor l for observation i. The factor matrices may have overlapping columns, but should not be completely identical, or else unidentifiability will be introduced.
2. $\epsilon^{(l)} = (\epsilon_{1l}, \ldots, \epsilon_{nl})^T \in \mathbb{R}^n$ denotes the vector of idiosyncratic errors for the lth factor, normally distributed, and independent for $l = 1, \ldots, L$. In Section 4.3, we provide numerical results showing that our estimation is not particularly sensitive to dependent idiosyncratic noises.
3. $\log(\alpha_l), l = 1, \ldots, L$, is the intercept for the lth factor, controlling importance level of the lth factor. This idea can be better visualized when we take the exponential function of Eq. (2.4).

$$Z_i = \exp(Y_i) = \max\left\{\alpha_1 \cdot \exp(\mathbf{X}_{i\cdot}^{(1)}\beta_1)\cdot\xi_{i1}, \ldots, \alpha_L \cdot \exp(\mathbf{X}_{i\cdot}^{(L)}\beta_L)\cdot\xi_{iL}\right\}, \tag{2.5}$$

where $\xi^{(l)} = (\xi_{1l}, \ldots, \xi_{nl})^T = \exp(\epsilon^{(l)}) \in \mathbb{R}_+^n$ has a log-normal distribution. Denote $\widetilde{\beta}_l = (\log(\alpha_l), \beta_l^T)^T$.

Remarks:

1. The rationale behind having linear relations within groups and nonlinear relations only between groups has two folds. (1) Technically, the max relation as nonlinear relation between groups allows factors to compete with each other, which is motivated by the recent development of competing risk models, e.g., Cui and Zhang (2018). (2) Practically, our new model is motivated by a risk analysis point of view. Consider a stock in the equity market, its risk of loss can be due to its operational errors, a market recession, or a global economic contagion effect, to name a few. The risk of loss is therefore the maximum of the three (or more) risks, i.e., whichever is the largest. Naturally, each risk may be modeled as a function of predictors, i.e., economic indicators used in Section 5. A linear structure has greater interpretability than any other model structure and has been a driving force in the econometric literature and many other research fields. It is adopted in this paper. Of course, a linear relationship has its own limitation, and other relations within groups can be a choice, which will also increase the model complexity.
2. We note that the max operator rules out possible problems caused by some outlying values in features $\mathbf{X}$. As such, it increases the robustness of the regression model.
3. Assumption of independence across groups is needed for establishing an explicit complete likelihood function in estimation via EM algorithm (Section (3.1)). However, cross-group dependence of idiosyncratic noises is desirable in many cases. In order to take this type of dependence into consideration, in latter sections, our numerical results (Section 4) show that our estimation is not particularly sensitive to dependent idiosyncratic noises.

The max-linear structure introduces a non-Gaussian distribution for the response variable Y, with $E(Y|\mathbf{X})$ a non-linear function of $\mathbf{X}$ as we can see later in Eq. (3.7). It is obvious that linear regression is a subset of max-linear models when the number of factors reduces to 1. It is also clear that Model (2.4) is inseparable overall but separable within each factor.

The max-linear model exploits the structure of competing factors and has some natural exemplification. For instance, to evaluate soccer players' market values, one would have to take into account of different factors (sets of attributes) and the factor that plays a dominant role might be different depending on the player's position: attacking and skills win

over other factors for the pioneer's, while defense skills win over for defenders. Another example is in online advertising, biding prices not only take into account of Ads quality but also publishing websites' quality in order to penalize bad sites; therefore features are naturally grouped into advertiser related and publisher related. A wise biding algorithm would inspect the competition between advertiser factor and publisher factor.

3. Estimation and prediction

3.1. Estimation via EM algorithm

Ordinary least squares estimators obviously cannot be applied to max-linear models. Similar to a mixture of distributions, given the latent dominant factor index, the max-linear structure simplifies to a simple linear structure. This kind of estimation problem is canonically solved via Expectation–Maximization (EM) based maximum likelihood estimation procedure. We now briefly review the EM algorithm.

Dempster et al. (1977) firstly introduced the EM algorithm. Starting from the basic identity

$$p(\mathbf{k}|\theta, \mathbf{y}) = \frac{p(\mathbf{y}, \mathbf{k}|\theta)}{p(\mathbf{y}|\theta)},$$

where θ is the vector of unknown parameters of interest, $\mathbf{k}$ is the vector of augmented data, and $\mathbf{y}$ is the vector of observed data, the complete-data likelihood $L^C(\theta|\mathbf{y}, \mathbf{k})$ and observed-data likelihood follow

$$\log L(\theta|\mathbf{y}) = E_{\theta_0}[\log L^C(\theta|\mathbf{y}, \mathbf{k})] - E_{\theta_0}[\log p(\mathbf{k}|\theta, \mathbf{y})],$$

where the expectation is taken with respect to $p(\mathbf{k}|\theta_0, \mathbf{y})$ with θ_0 as the true parameter.

Denote $Q(\theta|\theta_0, \mathbf{y}) = E_{\theta_0}[\log L^C(\theta|\mathbf{y}, \mathbf{k})]$. To maximize $\log L(\theta|\mathbf{y})$, we only need to iteratively increase $Q(\theta|\theta_0, \mathbf{y})$ because $E_{\theta_0}[\log p(\mathbf{k}|\theta, \mathbf{y})]$ naturally gets smaller when $p(\mathbf{k}|\theta, \mathbf{y}) \neq p(\mathbf{k}|\theta_0, \mathbf{y})$ by Jensen's inequality.

The max-linear regression formulation has a natural form of augmentation. By surfacing the latent variable of the dominant factor, we can create the complete likelihood and estimate the parameters via EM approach. Let K be the latent variable such that $K = l$ iff $Y = \mu_l + \epsilon_l$, where $\mu_l = \log \alpha_l + x_l^T \beta_l$, x_l is a vector of covariates from factor l. We now start with a few distributional facts of the newly introduced class of distributions generated by the max-linear regression model. The probability density function of Y, with cumulative distribution function

$$P(Y \leq y) = \prod_{l=1}^{L} P(\mu_l + \epsilon_l \leq y) = \prod_{l=1}^{L} \Phi(\frac{y - \mu_l}{\sigma_l}),$$

can be obtained by taking derivative and is

$$p(y) = \sum_{l=1}^{L} \frac{1}{\sigma_l} \phi(\frac{y - \mu_l}{\sigma_l}) \prod_{k \neq l} \Phi(\frac{y - \mu_k}{\sigma_k}). \tag{3.1}$$

The joint distribution of (Y, K) is

$$P(Y \leq y, K = l) = P(\mu_k + \epsilon_k \leq \mu_l + \epsilon_l \leq y, \forall k \neq l) = \int_{-\infty}^{y} \frac{1}{\sigma_l} \phi(\frac{z - \mu_l}{\sigma_l}) \prod_{k \neq l} \Phi(\frac{z - \mu_k}{\sigma_k}) dz.$$

Therefore the joint probability density function of the mixed variables is

$$p(y, K = l) = \frac{1}{\sigma_l} \phi(\frac{y - \mu_l}{\sigma_l}) \prod_{k \neq l} \Phi(\frac{y - \mu_k}{\sigma_k}). \tag{3.2}$$

Notice that Eq. (3.1) is a sum of Eq. (3.2). Therefore the complete likelihood of Y, K becomes

$$p(K, Y; X, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}) = \prod_{i=1}^{n} \prod_{l=1}^{L} \left[\frac{1}{\sigma_l} \phi(\frac{Y_i - \log \alpha_l - \mathbf{X}_{i\cdot}^{(l)} \beta_l}{\sigma_l}) \prod_{k \neq l} \Phi(\frac{Y_i - \log \alpha_k - \mathbf{X}_{i\cdot}^{(k)} \beta_k}{\sigma_k}) \right]^{\mathcal{I}(K_i = l)},$$

with the log-likelihood as

$$L(K, Y; X, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\sigma}) = \sum_{i=1}^{n} \sum_{l=1}^{L} \mathcal{I}(K_i = l) \log \left[\frac{1}{\sigma_l} \phi(\frac{Y_i - \log \alpha_l - \mathbf{X}_{i\cdot}^{(l)} \beta_l}{\sigma_l}) \cdot \prod_{k \neq l} \Phi(\frac{Y_i - \log \alpha_k - \mathbf{X}_{i\cdot}^{(k)} \beta_k}{\sigma_k}) \right].$$

The competing structure easily invites large number of predictors within factors, and the estimation becomes less efficient especially when the number of predictors is greater than that of the observations. In Section 3.3 we explore one way of penalization to enforce sparse models.

In order to take the E-steps in the EM algorithm, the conditional distribution of the latent variable given the observed is derived as follows. Denote $\eta_l(y)$ as the conditional probability mass function of K given y, then

$$\eta_l(y) = P(K = l|y) = \frac{\frac{1}{\sigma_l}\phi(\frac{y-\mu_l}{\sigma_l})\prod_{k\neq l}\Phi(\frac{y-\mu_k}{\sigma_k})}{\sum_{l=1}^{L}\frac{1}{\sigma_l}\phi(\frac{y-\mu_l}{\sigma_l})\prod_{k\neq l}\Phi(\frac{y-\mu_k}{\sigma_k})}, \tag{3.3}$$

which makes the analytical form of $Q(\theta|\theta^t)$ in the E-step in the EM algorithm available, where θ^t is the estimated θ at step t.

$$\begin{aligned}
Q(\theta|\theta^t) &= \sum_{i=1}^{n}\sum_{l=1}^{L}\eta_{il}^{(t)}\log\left[\frac{1}{\sigma_l}\phi\left(\frac{Y_i-\log\alpha_l-\mathbf{X}_{i\cdot}^{(l)}\beta_l}{\sigma_l}\right)\prod_{k\neq l}\Phi\left(\frac{Y_i-\log\alpha_k-\mathbf{X}_{i\cdot}^{(k)}\beta_k}{\sigma_k}\right)\right] \\
&= \sum_{i=1}^{n}\sum_{l=1}^{L}\left\{\eta_{il}^{(t)}\log\left[\frac{1}{\sigma_l}\phi\left(\frac{Y_i-\log\alpha_l-\mathbf{X}_{i\cdot}^{(l)}\beta_l}{\sigma_l}\right)\right]+\right. \\
&\quad \left.(1-\eta_{il}^{(t)})\log\left[\Phi\left(\frac{Y_i-\log\alpha_l-\mathbf{X}_{i\cdot}^{(l)}\beta_l}{\sigma_l}\right)\right]\right\} \\
&= -\frac{n}{2}\log(2\pi)+\sum_{i=1}^{n}\sum_{l=1}^{L}\left\{-\eta_{il}^{(t)}\log\sigma_l-\frac{\eta_{il}\left(Y_i-\log\alpha_l-\mathbf{X}_{i\cdot}^{(l)}\beta_l\right)^2}{2\sigma_l^2}\right. \\
&\quad \left.+(1-\eta_{il}^{(t)})\log\left[\Phi\left(\frac{Y_i-\log\alpha_l-\mathbf{X}_{i\cdot}^{(l)}\beta_l}{\sigma_l}\right)\right]\right\},
\end{aligned}$$

where $\eta_{il}^{(t)}$ is the short for $\eta_l(Y_i, \mathbf{X}_{i\cdot}, \theta^t)$, i.e., the conditional probability of the lth factor dominating for observation i conditional on the response variable and parameter estimates at the step t. This conditional probability shows the result of competing of factors: each $\eta_{il}^{(t)}$ is the winning probability of factor l in the competition for the ith sample at the tth iteration. The first equation simply takes the conditional expectation of the complete log-likelihood function with respect to $p(K|\theta^t, Y)$.

The M-step of the EM algorithm maximizes $Q(\theta|\theta^t)$, which can be written as the summation of $Q_l(\theta|\theta^t)$ over factors in Eq. (3.4),

$$\begin{aligned}
Q_l(\theta|\theta^t) &= -\frac{n}{2L}\log(2\pi)+\sum_{i=1}^{n}\left\{-\eta_{il}\log\sigma_l-\frac{\eta_{il}\left(Y_i-\log\alpha_l-\mathbf{X}_{i\cdot}^{(l)}\beta_l\right)^2}{2\sigma_l^2}\right. \\
&\quad \left.+(1-\eta_{il})\log\left[\Phi\left(\frac{Y_i-\log\alpha_l-\mathbf{X}_{i\cdot}^{(l)}\beta_l}{\sigma_l}\right)\right]\right\}.
\end{aligned} \tag{3.4}$$

Notice that the decomposition of the term $n\log(2\pi)/2$ to the L factors is inessential. One can use any other weights, e.g., $\sum_{i=1}^{n}\eta_{il}^{t}$ instead of the constant weights $1/L$ for factor l.

The uniqueness of the maximizer is ensured by the fact that the function

$$f(y,\sigma) = p\log\left(\frac{1}{\sigma}\phi\left(\frac{y}{\sigma}\right)\right)+(1-p)\log\left(\Phi\left(\frac{y}{\sigma}\right)\right),$$

with p a known constant is a concave function of (y, σ^2) and that affine transformations do not change the concavity or convexity of a function. The optimization of the M-step given η_{il} can be separated into L optimization problems by the groups since the parameters $\theta_l = (\tilde{\beta}_l^T, \sigma_l)^T$ only appear in $Q_l(\theta|\theta^t)$. This property facilitates group-wise optimization at M-step. The EM algorithm iterates between E-step that updates the winning probability of each factor, and M-step that maximizes the group-wise conditional expectation of the complete log-likelihood function. The details of the algorithm are deferred to Section 3.3 after regularization is introduced.

3.2. Theoretical results

3.2.1. Asymptotics of MLE

In this section, we provide the theory for consistency and asymptotic normality of maximum likelihood estimator of the max-linear model.

The following assumptions are required for the consistency and asymptotic normality.

A1. The parameter space Θ is compact.
A2. The distribution that generates X_i has finite first and second moments.
A3. $P(f(Y_i|X_i;\theta) \neq f(Y_i|X_i;\theta_0)) > 0, \forall \theta \in \Theta$ with $\theta \neq \theta_0$, where θ_0 is the true parameter.

$$f(Y_i|X_i;\theta) = \sum_{l=1}^{L} \frac{1}{\sigma_l}\phi\left(\frac{Y_i - \log\alpha_l - X_i^l\beta_l}{\sigma_l}\right)\prod_{k\neq l}\Phi\left(\frac{Y_i - \log\alpha_k - X_i^k\beta_k}{\sigma_k}\right).$$

A4. θ_0 is in the interior of Θ.
A5. The Fisher information $I(\theta) = E\left[\frac{\partial}{\partial\theta}\log f(Y_i|X_i;\theta)\frac{\partial}{\partial\theta^T}\log f(Y_i|X_i;\theta)\right]$ is well defined and positive definite at θ_0.
A6. For all $1 \leq i \leq n$, $\{X_i^{(l)}\}_{l=1}^{L}$ has finite third to fifth order moments.

We obtain the following Theorem 3.1, and the proof is given in the Appendix (online supplementary file).

Theorem 3.1. *Assume $Y_i, i = 1,\ldots,n$, follow the max-linear model in (2.4) with independent error terms ϵ_{ij} for all i, j. Denote θ_0 as the true value of the parameter, and $\widehat{\theta}_n$ as the MLE. Under Assumptions A1–A3, as $n \to \infty$,*

$$\widehat{\theta}_n \to_p \theta_0.$$

If Assumptions A4–A6 are satisfied in addition, we have

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \to N(0, [I(\theta_0)]^{-1}),$$

where $I(\theta_0)$ is the Fisher information matrix.

3.2.2. Convergence of EM algorithm

We now provide the convergence of observed-data likelihood and EM sequence by theorems in Wu et al. (1983) and McLachlan and Krishnan (2007).

Theorem 3.2. *Define the observed-data likelihood function:*

$$L_O(\theta) = p(y;\mathbf{X},\theta) = \sum_{l=1}^{K} \frac{1}{\sigma_l}\phi\left(\frac{y-\mu_l}{\sigma_l}\right)\prod_{k\neq l}\Phi\left(\frac{y-\mu_k}{\sigma_k}\right), \tag{3.5}$$

and the following assumptions:

A7. $\sigma_l \geq \epsilon > 0, \forall l = 1,\ldots,K$, i.e., variances are bounded away from 0.
A8. $\|\theta^{t+1} - \theta^t\| \to 0$, as $t \to \infty$.

Then under Assumptions A1 and A7, all the limit points of θ^t are stationary points and $L_O(\theta^t)$ converges monotonically to $L_O^* = L_O(\theta^*)$ for some stationary point $\theta^* \in \Omega$.

If we further assume that A8 holds, then θ^t converges to some θ^* in $\Omega(L_O^*) = \{\theta \in \Omega : L_O(\theta) = L_O^*\}$.

The proof of this theorem straightforwardly follows the proof in Wu et al. (1983).

The assumption A7 rules out the cases where the noise variances approach infinity. This assumption is to guarantee the compactness regularity condition (condition (3.19) in McLachlan and Krishnan (2007)) for max linear regression structure. The assumption A8 gives the stop rule for our algorithm: empirically, we adopt the stop criterion: $\|\theta^{t+1} - \theta^t\| \leq 0.001$ for simulation studies and real data applications.

3.3. Penalization

In this section, we introduce sparsity into the max-linear model by incorporating penalization. Consider the following penalization on $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$,

$$\min_{\theta} \quad -L(\theta;\mathbf{y}) + \sum_{l=1}^{L}\left[\lambda_1 e^{\beta_{l0}} + \lambda_2\|\beta_l\|_1\right], \tag{3.6}$$

where $\beta_{l0} = \log\alpha_l$ and $\alpha_l > 0$, and $\|\cdot\|_p$ is the L_p norm of a vector. The penalization on the covariates within each factor is of standard lasso type, hence introduces within-group variable sparsity, while the penalization on the intercept is of exponential type e^x, which is equivalent to penalizing the multiplicative factor α_l in model (2.5) to the direction towards 0, with the rate of a L_1 penalty. Due to the parameterization, the penalization on the group intercept would not penalize any intercept to exactly negative infinity and hence would not introduce group sparsity directly. However, group sparsity

can still be observed based on winning probability of each competing factor. Lower intercept results in smaller winning probability and consequently lesser effectiveness for the corresponding competing factor.

Next we briefly describe the algorithm. With the penalization terms, $Q_l(\theta|\theta^t)$ becomes

$$Q_l(\theta|\theta^t) = \sum_{i=1}^{n}\left[\eta_{il}^t \log\sigma_l + \frac{\eta_{il}^t\left(Y_i - \mathbf{X}_{i\cdot}^{(l)}\widetilde{\beta}_l\right)^2}{2\sigma_l^2} - (1-\eta_{il}^t)\log\Phi\left(\frac{Y_i - \mathbf{X}_{i\cdot}^{(l)}\widetilde{\beta}_l}{\sigma_l}\right)\right] + \lambda_1 e^{\beta_{l0}} + \lambda_2\|\beta_l\|_1 + \frac{n}{2L}\log(2\pi).$$

Taking derivative with respect to $\widetilde{\beta}$, we have

$$\frac{\partial Q_l(\theta|\theta^t)}{\partial\widetilde{\beta}_l} = \frac{1}{\sigma_l}\sum_{i=1}^{n}\left[(1-\eta_{il}^t)\frac{\phi\left(\frac{Y_i - \mathbf{X}_{i\cdot}^{(l)}\widetilde{\beta}_l}{\sigma_l}\right)}{\Phi\left(\frac{Y_i - \mathbf{X}_{i\cdot}^{(l)}\widetilde{\beta}_l}{\sigma_l}\right)} - \frac{\eta_{il}^t\left(Y_i - \mathbf{X}_{i\cdot}^{(l)}\widetilde{\beta}_l\right)}{\sigma_l}\right](\mathbf{X}_{i\cdot}^{(l)})^T + \lambda_1 e^{\widetilde{\beta}_{l0}} e_0 + \lambda_2 t_l,$$

where e_0 is a vector with all components 0 except the first one, t_l is a vector with $t_{l0} = 0$, and

$$t_{lj} = \begin{cases} sign(\beta_{lj}), & \beta_{lj} \neq 0, \\ [-1, 1], & \beta_{lj} = 0. \end{cases}$$

The optimization of $Q_l(\theta|\theta^t)$ can be accomplished via gradient descent, iterating over each parameter using the following partial derivatives.

$$\frac{\partial Q_l(\theta|\theta^t)}{\partial\beta_{l0}} = \frac{1}{\sigma_l}\sum_{i=1}^{n}\left[(1-\eta_{il}^t)\frac{\phi\left((\widetilde{e}_i^{(0)} - \beta_{l0})/\sigma_l\right)}{\Phi\left((\widetilde{e}_i^{(0)} - \beta_{l0})/\sigma_l\right)} - \eta_{il}^t\left((\widetilde{e}_i^{(0)} - \beta_{l0})/\sigma_l\right)\right] + \lambda_1 e^{\beta_{l0}},$$

$$\begin{aligned}\frac{\partial Q_l(\theta|\theta^t)}{\partial\beta_{lj}} &= \frac{1}{\sigma_l}\sum_{i=1}^{n}\left[(1-\eta_{il}^t)\frac{\phi\left((\widetilde{e}_i^{(j)} - \beta_{lj}X_{ij}^{(l)})/\sigma_l\right)}{\Phi\left((\widetilde{e}_i^{(j)} - \beta_{lj}X_{ij}^{(l)})/\sigma_l\right)} - \eta_{il}^t\left((\widetilde{e}_i^{(j)} - \beta_{lj}X_{ij}^{(l)})/\sigma_l\right)\right]X_{ij}^{(l)} \\ &\quad + \lambda_2 t_{lj}, j \neq 0 \\ &= \frac{1}{\sigma_l}\sum_{i=1}^{n}X_{ij}^{(l)}\left[(1-\eta_{il}^t)\frac{\phi\left((\widetilde{e}_i^{(j)} - \beta_{lj}X_{ij}^{(l)})/\sigma_l\right)}{\Phi\left((\widetilde{e}_i^{(j)} - \beta_{lj}X_{ij}^{(l)})/\sigma_l\right)} - \eta_{il}^t\widetilde{e}_i^{(j)}/\sigma_l\right] \\ &\quad + \sum_{i=1}^{n}\eta_{il}^t(X_{ij}^{(l)})^2/\sigma_l^2\cdot\beta_{lj} + \lambda_2 t_{lj},\end{aligned}$$

where $\widetilde{e}_i^{(j)} = Y_i - \sum_{m\neq j}X_{im}^{(l)}\beta_{lm}$. To update σ_l for the lth factor, it is convenient to rewrite $Q_l(\theta|\theta^t)$ with the updated residual,

$$Q_l(\theta|\theta^t) = \sum_{i=1}^{n}\left[\frac{\eta_{il}^t}{2}\left(\widetilde{e}_i^{(j)}/\sigma_l - \left(X_{ij}^{(l)}/\sigma_l\right)\beta_{lj}\right)^2 - (1-\eta_{il}^t)\log\Phi\left((\widetilde{e}_i^{(j)} - \beta_{lj}X_{ij}^{(l)})/\sigma_l\right)\right] + \lambda_1 e^{\beta_{l0}} + \lambda_2\|\beta_l\|_1 + \frac{1}{2}\log(2\pi\sigma_l^2)\sum_{i=1}^{n}\eta_{il}^t.$$

σ_l^2 can be updated using any one-dimensional optimization method. Then the EM algorithm for solving (3.6) is listed in Algorithm 1.

3.4. Prediction

Predictions under linear models are expected values of the response variable given the covariates and the estimated coefficients. Similarly we can make unbiased predictions for the response variable under the max-linear model. The conditional expectations are not the maximum of the expected means across factors due to the random error terms

Algorithm 1 EM algorithm for solving max-linear regression.

```
initialization: OLS
while not converged and maximum step not exceeded do
    E step: calculate η_il^t.
    M step:
    for l = 1 : L
        Update β_l0 with penalization λ_1 e^{β_l0}.
        for j = 1 : p_l
            Check if 0 is solution. If not, update β_lj via coordinate descent.
        end for
        update σ_l^2
    end for
end while
```

in each factor. The expectation of Y has the following analytical form due to Lemma 11,

$$\begin{aligned} E[Y] &= \sum_{l=1}^{L} \int_y \frac{y}{\sigma_l} \phi\left(\frac{y-\mu_l}{\sigma_l}\right) \prod_{k\neq l} \Phi\left(\frac{y-\mu_k}{\sigma_k}\right) dy \\ &= \sum_{l=1}^{L} \mu_l \Phi_{L-1}(\mu_l - \boldsymbol{\mu}_{-l}; 0, diag(\boldsymbol{\sigma}^2_{-l}) + \sigma_l^2 J_{L-1}) + \sum_{l=1}^{L}\sum_{k\neq l} \sigma_l^2 \phi\left(\mu_l - \boldsymbol{\mu}_k; 0, \sqrt{\sigma_l^2+\sigma_k^2}\right) \cdot \\ &\quad \Phi_{L-2}\left(\mu_l - \boldsymbol{\mu}_{-(l,k)}; \frac{\mu_l-\mu_k}{1+(\sigma_k/\sigma_l)^2}\mathbf{1}_{L-2}, diag(\boldsymbol{\sigma}^2_{-(l,k)}) + \frac{1}{\sigma_k^{-2}+\sigma_l^{-2}} J_{L-2}\right), \end{aligned} \tag{3.7}$$

where $x_{-(ind)}$ is the vector obtained by removing the elements with indices in the set ind from the vector x, and $\Phi_k(x; \mu, \Sigma)$ is the c.d.f. of the k-dimensional multivariate normal distribution with mean μ and covariance matrix Σ. $\mathbf{1}_k$ is k-dimensional vector of ones, and J_k is the $k \times k$ matrix with all elements being one.

On the other hand we have the probability of lth factor dominating,

$$\eta_l = P(y = \mu_l) = P\left(\epsilon_k \leq \epsilon_l + \mu_l - \mu_k, k \neq l\right) = E\left(\prod_{k\neq l} \Phi\left(\frac{\mu_l - \mu_k + \epsilon_l}{\sigma_k}\right)\right). \tag{3.8}$$

Lemma 9 indicates that $\eta_l = \Phi_{L-1}(\mu_l - \boldsymbol{\mu}_{-l}; \mathbf{0}, \text{diag}(\boldsymbol{\sigma}^2_{-l}) + \sigma_l^2 J_{L-1})$, thus

$$\begin{aligned} E(Y) &= \sum_{l=1}^{L} \mu_l \eta_l + \sum_{l=1}^{L}\sum_{k\neq l} \sigma_l^2 \phi\left(\mu_l - \boldsymbol{\mu}_k; 0, \sqrt{\sigma_l^2+\sigma_k^2}\right) \cdot \\ &\quad \Phi_{L-2}\left(\mu_l - \boldsymbol{\mu}_{-(l,k)}; \frac{\mu_l-\mu_k}{1+(\sigma_k/\sigma_l)^2}\mathbf{1}_{L-2}, diag(\boldsymbol{\sigma}^2_{-(l,k)}) + \frac{1}{\sigma_k^{-2}+\sigma_l^{-2}} J_{L-2}\right). \end{aligned} \tag{3.9}$$

Although it is tempting to go along with the path $E(Y) = E(E(Y|K))$, where $E(Y|K = l)$ seems to be simply μ_l, which eventually gives rise to $\sum_{l=1}^{L} \mu_l \eta_l$ as the expectation of Y, we note that $(Y|K = l)$ is no longer distributed as $N(\mu_l, \sigma_l^2)$ since the condition $K = l$ has imposed certain constraint on Y that alters the conditional distribution. Eq. (3.9) has the weighted average of group means as one component, but in addition incorporates the variances.

4. Simulation

In this section, we assess the performance of the estimation procedure derived in Section 3 for both the max-linear regression base (fixed dimension) model by providing the mean values and standard errors for estimated parameters and for high dimensional settings by comparing its prediction accuracy, model sparsity and estimation precision with other benchmark methodologies using simulated data. The first benchmark method is the random forest (Breiman, 2001). We mostly use the default setting from `RandomForestRegressor` in `sklearn` library in Python, except that we select five times of trees than the default value. We choose random forest as a benchmark for prediction accuracy because its power to handle non-linearity and interaction made it one of the most powerful methodologies for prediction. Sparse group lasso, proposed by Yuan and Lin (2006) has structural similarity to our methodology in that it has variable groups, and uses regularization for sparsity. All the examples in this section are simulated based on model (2.4) with varying parameter setting. The independent variable $X_i^{(l)}$ is generated independently and identically distributed from $N(0, I)$ where I is the

Table 4.1
True parameters of max-linear models for both based model settings (β_is with zero values are not included in model fittings) and high dimensional settings (independent features are generated for model fittings for both zero and nonzero coefficients).

		α	σ	β_1	β_2	β_3	β_4	β_5	...	β_{10}
Example 1	Competing factor 1	1.2	1	−3	2	0	0	0	...	0
	Competing factor 2	0.5	1	2	2	0	0	0	...	0
	Competing factor 3	1.5	1	−2	3	0	0	0	...	0
Example 2	Competing factor 1	1.2	1	−3	2	0	0	0	...	0
	Competing factor 2	0.5	1.5	2	2	0	0	0	...	0
	Competing factor 3	1.5	2	−2	3	0	0	0	...	0
Example 3	Competing factor 1	1.2	2	−2	3	0	0	0	...	0
	Competing factor 2	0.5	1.5	−2	3	1	0	0	...	0
	Competing factor 3	1.5	1	−2	3	1	1.5	0	...	0

identity matrix. Five-fold cross-validation is used for selecting the best tuning parameter in each method. Candidates of the two tuning parameters are formed in a grid. Ranges of candidates are decided empirically. For each parameter set, we conducted 500 repetitions.

Four metrics are used to evaluate model performance. First, mean squared error, i.e., $\sum_{i=1}^{n}(y_i-\widehat{y_i})^2/n$, is used to measure prediction performance. Secondly to assess parameter estimation, we calculate the distance of the normalized coefficients between the estimated and true values. This metric is not interpretable for other models which are mis-specified, but is useful to reveal the effectiveness of our algorithm. Specifically, for each of the parameters $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and σ, we take the L2-norm between the normalized true value and normalized estimated, e.g. $\left\| \frac{\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|} - \frac{\widehat{\boldsymbol{\beta}}}{\|\widehat{\boldsymbol{\beta}}\|} \right\|_2$. Lastly, to evaluate the estimated model sparsity, we compare the sensitivity and specificity of variable selection, defined as

$$\text{sensitivity} = \frac{\text{No. of coefficients correctly estimated as non-zero}}{\text{No. of non-zero coefficients in true } \boldsymbol{\beta}},$$

$$\text{specificity} = \frac{\text{No. of coefficients correctly estimated as zero}}{\text{No. of zero coefficients in true } \boldsymbol{\beta}},$$

respectively. Sensitivity measures the ability to accurately identify truly important variables, while specificity measures the precision of the selected variables on hitting the truly important variables.

4.1. Max-linear regression models

We conduct simulation for three-factor max-linear regression models. For base model cases, three factors are used to generate the true model and all coefficients for features are non-zeros. For high dimensional settings, each factor contains 10 features, where only a part of coefficients in each factor are non-zeros. We progressively made the simulation setup harder with tweaks to the model as follows.

1. Example 1 has constant variance in each competing factor.
2. In Example 2, competing factors have varying variances.
3. In Example 3, we create partially overlapping variables between competing factors in addition to varying variances. Specifically, $\mathbf{X}_{\cdot 1}^{(l)}$, the first variable of each factor, is shared across three competing factors.

Table 4.1 lists the parameter values of this simulation for base model and high dimensional settings.

$n = 500$ observations are generated in each repetition. Table 4.2 shows the simulation results for base model settings. Estimated values and standard deviations for $\boldsymbol{\sigma}$, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ are provided. Table 4.3 shows the key simulation results for high dimensional settings, such as prediction error, sensitivity, and specificity, etc. Certain metrics are missing for some methods because of unavailability. For instance, random forest model, as a tree model, does not have an explicit parametric form, thus estimation error for $\boldsymbol{\beta}$, sensitivity and specificity are not available.

For base model estimation, the parameters are closely estimated with relatively small standard errors. Results for those examples validate the theoretical results derived in Section 3.2.1. For high-dimensional settings, without surprise random forest outperforms group lasso in prediction accuracy. Max-linear model outperforms the flexible random forest by a big margin due to its correctly specified model. The max() function splits the hyperspace into sub-spaces with linear surfaces, which random forest might have difficulty identifying due to the complexity of the true model. Further, the random effect ϵ_l in each competing factor also increases difficulty for random forest to discover the model structure. In terms of variable selection, max-linear model outperforms group lasso in sensitivity and specificity most of the times. The MSE in Example 2 is larger compared to Example 1, due to lower signal to noise ratio.

In order to further stress how the values of λ_1 and λ_2 affect type I and type II errors, we also provide ROC curves corresponding to simulation Examples 1, 2 and 3. Fig. 4.1 demonstrates the ROC curves for these examples.

To create the ROC curves, we fixed levels of the group penalty constant λ_1 at different values, and observe true positive rates and false positive rates for different candidates of the variable selection penalty constant λ_2.

Table 4.2
Estimated parameters of max-linear models for base model settings. ("–" means no corresponding feature).

		α (SE)	σ (SE)	β_1 (SE)	β_2 (SE)	β_3 (SE)	β_4 (SE)
Example 1	CF 1	1.218 (0.246)	0.988 (0.055)	−2.999 (0.141)	2.001 (0.110)	–	–
	CF 2	0.510 (0.105)	0.988 (0.074)	2.003 (0.218)	2.001 (0.136)	–	–
	CF 3	1.516 (0.308)	0.985 (0.053)	−1.996 (0.113)	3.003 (0.138)	–	–
Example 2	CF 1	1.232 (0.253)	0.986 (0.060)	−2.991 (0.137)	1.998 (0.111)	–	–
	CF 2	0.529 (0.155)	1.480 (0.118)	2.003 (0.189)	1.997 (0.186)	–	–
	CF 3	1.493 (0.444)	1.982 (0.112)	−2.020 (0.188)	3.029 (0.213)	–	–
Example 3	CF 1	1.192 (0.340)	1.983 (0.126)	−1.989 (0.178)	3.016 (0.225)	–	–
	CF 2	0.528 (0.173)	1.459 (0.111)	−1.997 (0.168)	2.988 (0.225)	0.998 (0.163)	–
	CF 3	1.522 (0.353)	0.990 (0.063)	−1.996 (0.100)	2.999 (0.146)	1.002 (0.088)	1.503 (0.098)

Table 4.3
Results for max-linear models corresponding to Table 4.1.

	Max-linear	(SE)	Sparse group lasso	(SE)	Random forest	(SE)
Prediction MSE	0.928	(0.001)	3.499	(0.005)	2.126	(0.005)
Sensitivity	1.000	(0.000)	1.000	(0.000)		
Specificity	0.847	(0.004)	0.853	(0.005)		
β	0.058	(0.001)	0.263	(0.002)		
$\log\alpha$	0.378	(0.007)				
σ	0.223	(0.005)				
Prediction MSE	2.303	(0.023)	4.452	(0.007)	3.357	(0.006)
Sensitivity	0.999	(0.001)	1.000	(0.000)		
Specificity	0.908	(0.004)	0.858	(0.005)		
β	0.108	(0.002)	0.272	(0.003)		
$\log\alpha$	0.605	(0.049)				
σ	2.725	(0.300)				
Prediction MSE	2.273	(0.005)	5.123	(0.009)	4.285	(0.009)
Sensitivity	1.000	(0.000)	0.988	(0.002)		
Specificity	0.847	(0.005)	0.792	(0.006)		
β	0.095	(0.001)	0.252	(0.002)		
$\log\alpha$	0.524	(0.008)				
σ	2.788	(0.026)				

Table 4.4
True parameters of a simple linear regression in Section 4.2.

	α	σ	β_1	β_2	β_3	β_4	...	β_{10}
Competing factor 1	1.2	1	−2	3	1	0	...	0

As the complexity of the true model increases, the ROC curve plots for Examples 1–3 reveal that the variable selection accuracy of max linear model drops slightly, but in general, performs plausibly. This observation is consistent with the sensitivity and specificity rows in Table 4.3. We also note that the choice of group penalization parameter does not affect ROC curves significantly. This phenomenon is because in the simulation settings of those examples, only three competing factors are used to generate the true model and every competing factor has non-ignorable chance of taking the lead in max linear regressions. To avoid penalizing the number of groups too much, the given range for candidates of λ_1 are not extremely large.

4.2. Simple linear models

In this example, we generate a simple linear model and test how effective our penalization is in reducing to the one-factor max-linear model. The parameters are listed in Table 4.4. We still supply three competing factors of variables, each with 10 independent and identically distributed features, but only the first three features in the first competing factor have truly non-zero coefficients in the underlying linear model.

Under this setting, max-linear model behaves very similarly to sparse group lasso, as both identify linear structure and penalize non-important coefficients. Random forest predicts less accurately than the other two methods which both have the correctly specified model structure (see Table 4.5).

Although the penalization on the intercept does not introduce sparsity, i.e., we cannot estimate the true values of intercepts $\log\alpha_2$ and $\log\alpha_3$ as $-\infty$, their estimated values are penalized enough so that they are extremely small, which makes it very unlikely for the second and third competing factors to dominate. This observation means that simple linear models can be algorithmically identified as a special case of max-linear model with one factor of true simple linear structure, and intercepts of all other competing factors approaching $-\infty$.

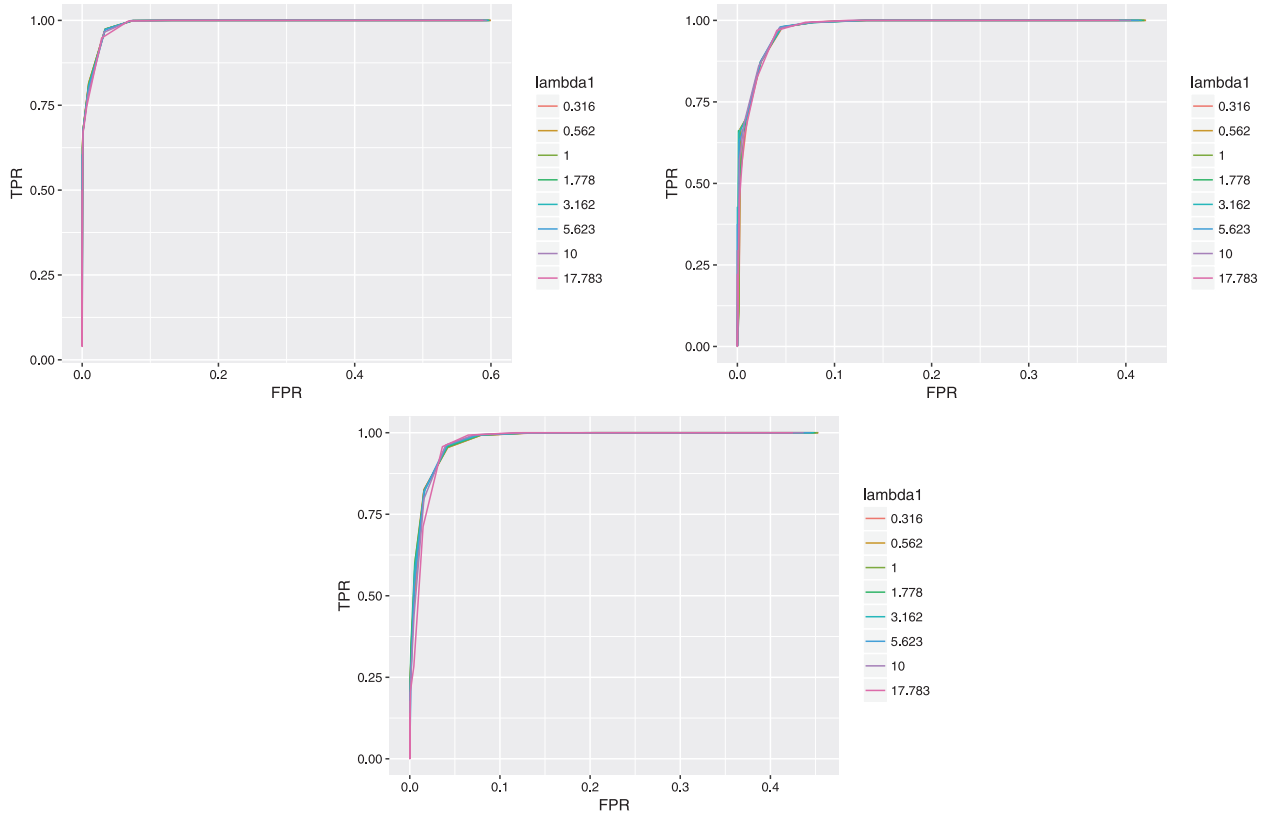

Fig. 4.1. ROC curves for Examples 1, 2 and 3 with different choices of λ_1.

Table 4.5
Results of the simple linear regression in Section 4.2.

	Max-linear	(SE)	Sparse group lasso	(SE)	Random forest	(SE)
Prediction MSE	1.020	(0.001)	1.284	(0.003)	2.046	(0.005)
Sensitivity	1.000	(0.000)	1.000	(0.000)		
Specificity	0.880	(0.004)	0.998	(0.000)		
$\boldsymbol{\beta}$	0.030	(0.001)	0.036	(0.000)		

In summary, the above two simulation results show that our method outperforms other methods when the underlying model has a true max-linear structure. The results imply that the non-linearity introduced by max-linear structure cannot even be handled by powerful prediction models like random forest. Sparse group lasso and the max-linear model has a comparable capability in variable selection. However, our estimations of regression coefficient and predictions are more accurate. When the underlying model is a linear model, performances of max-linear model and group lasso are very similar, and both outperform random forest.

4.3. Robustness to dependent errors

In this section, we introduce violations to our model assumptions by having dependent error terms and show the robustness of the model.

To fully test out our algorithm, we consider two extreme cases where the random error terms for the factors are perfectly dependent, either positively or negatively. More specifically, we consider the following two models.

$$Y_i = \max\left\{\log\alpha_1 + \mathbf{X}_{i\cdot}^{(1)}\beta_1, \ldots, \log\alpha_L + \mathbf{X}_{i\cdot}^{(L)}\beta_L\right\} + \epsilon_i, \tag{4.1}$$

$$Y_i = \max\left\{\log\alpha_1 + \mathbf{X}_{i\cdot}^{(1)}\beta_1 + \epsilon_i, \log\alpha_L + \mathbf{X}_{i\cdot}^{(2)}\beta_2 + \epsilon_i, \log\alpha_L + \mathbf{X}_{i\cdot}^{(3)}\beta_3 - \epsilon_i.\right\}. \tag{4.2}$$

Table 4.6 lists the parameter values of this simulation, which are the same as simulation Example 1.

Table 4.7 shows the results of the two simulation examples. The performance of prediction and estimation of coefficients are both comparable with setting under models with independent errors. The algorithm we developed is robust for the regression problems in Eqs. (4.1) and (4.2), with modeling assumptions in (2.4) being violated.

Table 4.6
True parameters for (4.1)–(4.2).

	α	σ	β_1	β_2	β_3	...	β_{10}
Competing factor 1	1.2	1	3	−2	0	...	0
Competing factor 2	0.5	1	2	2	0	...	0
Competing factor 3	1.5	1	−2	3	0	...	0

Table 4.7
Results of models given by Eqs. (4.1)–(4.2).

	Max-linear (SE)	Sparse group lasso (SE)	Random forest (SE)
Prediction MSE	1.086(0.002)	3.970(0.006)	2.389(0.006)
Sensitivity	1(0.000)	1(0.000)	
Specificity	0.863(0.005)	0.855(0.006)	
$\boldsymbol{\beta}$	0.064 (0.001)	0.277(0.002)	
$\log\alpha$	0.381(0.007)		
σ	0.461(0.008)		
Prediction MSE	0.853(0.011)	3.290(0.005)	1.996(0.005)
Sensitivity	0.997 (0.001)	1 (0.000)	
Specificity	0.876(0.005)	0.855(0.005)	
$\boldsymbol{\beta}$	0.062 (0.002)	0.277(0.002)	
$\log\alpha$	0.469(0.007)		
σ	0.383(0.041)		

5. Econometric and business applications

Econometric and business models are statistical models that specify the relationship between various informative quantities and economic phenomena or core aspects of a business. Most econometric and business models only utilize linear relationship due to its simplicity and interpretability, whereas linearity barely provides satisfactory explanation and prediction of economic phenomena towards making the right business decisions. For instance, one widely used linear econometric model is the Fama–French factor model, an asset pricing model using size of firms, book-to-market values and excess return on the market as predictors. In addition, those linear models often assume normal random errors and fail to capture potential asymmetry in the distribution of the response variable.

In this section, we present two econometric applications predicting bond risk premia and volatility index respectively with market indicators, and one business application in modeling soccer players' market values. These examples illustrate the extraordinary prediction power of our max-linear model in the absence of Gaussianity and linearity, even compared with other state-of-the-art machine learning methods known for their outstanding prediction performances. Meanwhile, the fitted models preserve an explicit and interpretable structure compared with those black-box machine learning models where structures are elusive.

5.1. Max-linear modeling Fama–Bliss bond risk premia

Bond risk premium is defined as the excess return compared with one-year bond, if we buy an n year bond and sell it as an $n-1$ year bond the next year. Studies on bond risk premia are numerous in the literature. Ludvigson and Ng (2009) and Fan et al. (2021) discussed the prediction of bond risk premia using macroeconomic market indicators with factor analysis. The former used dynamic factor model with block structure, while the latter proposed an estimation for the factor models that is robust to possibly heavy-tailed distribution. Following Ludvigson and Ng (2009), the macroeconomic market indicators (total 126) can be categorized into eight blocks: Output and Income, Labor Market, Housing, Consumption, Orders and Inventories, Money and Credit, Interest and Exchange Rates, Prices and Stock Market. In this section, we develop the max-linear competing factor model to predict risk premia of Fama–Bliss discount bonds using market indicators as predictor variables. A detailed description of this panel data can be found in the Appendix of Ludvigson and Ng (2009).

In our case, the bond risk premia are computed based on prices of the one — through five-year Fama–Bliss discount bonds from January 1990 to December 2017, which is available from the Center for Research in Securities Prices (CRSP). We model the bond risk premia to a max-linear relationship with the macroeconomic market indicators considering different competing factors may determine the market risk under changing economic environment. In Sections 5.1 and 5.2, we use the block structure mentioned above to form eight competing factors. We find that many variables within the same competing factor have high correlations, and hence variable selection within each competing factor is essential for the model to provide more insights.

To evaluate the prediction accuracy of bond risk premia using different models, we hold out the most recent one-year data (year 2017) as testing set consisting of 12 data points. We then use 100 seeds to randomly split the rest of data from year 1990 to 2016 into training and validation sets with the ratio 7:3, and construct an ensemble of 100 models using the

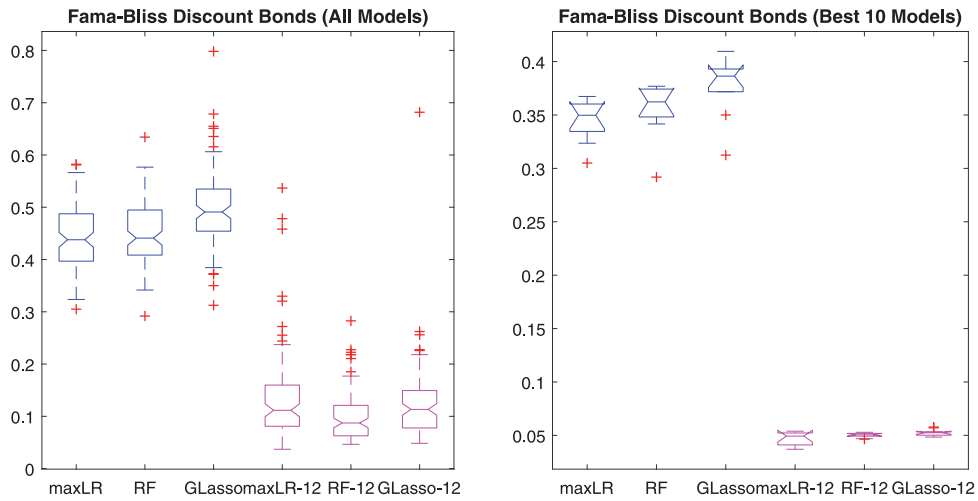

Fig. 5.1. (Fama–Bliss Bond Risk Premia). Prediction MSEs of validation set and testing set (ended with "−12") for max-linear competing regression, random forest, and sparse group lasso in 100 random cases (left panel) and 10 best performing models respectively on validation set and testing set (right panel).

Table 5.1
Average coefficients of Fama–Bliss bond risk premia data.

Important variable	UMCSENTx	EXUSUKx	EXSZUSx	EXJPUSx	EXCAUSx	AAA
Average coef.	0.039	0.055	0.022	−0.012	0.0009	0.442

Table 5.2
(Fama–Bliss bond risk premia). Average winning probability for the best performing model.

Competing Factor(CF)	Winning probability (Validation)	Winning probability (Test)
1. Output and income	0.006	0.003
2. Labor market	0.005	$<10^{-3}$
3. Consumption and orders	0.030	$<10^{-3}$
4. Orders and inventories	0.006	0.025
5. Money and credit	0.070	0.001
6. Interest and exchange rates	0.873	0.967
7. Price	0.007	0.002
8. Stock market	0.004	0.002

training set. The best model among the 100 models, evaluated using the associated validation dataset, can be chosen as the final predictive model for the testing set. Performance of the chosen model is used for comparing different methods in predicting future bond risk premia. One can also select the top K models from the ensemble, and build a meta-regressor to enhance prediction performance.

We compare the max-linear model with random forest, and sparse group lasso. Prediction MSEs on the validation and testing sets for the three methods are shown in the left panel of Fig. 5.1. If we restrict to the 10 best models for each method evaluated on validation sets as spearhead models and apply them on both validation set and testing set, we have the boxplots in the right panel of Fig. 5.1. The top max-linear models within its class show better performance than the other methods, both on validation and testing set.

To further uncover market indicators that are more important in predicting bond risk premia, we compute frequencies of nonzero estimates for each predictor among the ensemble models. Since the penalization on groups in max-linear model does not introduce group sparsity, we enforce group sparsity by discarding groups with average leading probability less than 5%, and coefficients corresponding to variables within those groups will be excluded from the frequency counting. Important market indicators are displayed in Fig. 5.2. UMCSENTx (Consumer Sentiment Index), EXUSUKx (U.S. / U.K. Foreign Exchange Rate) are shown to be important as their coefficients are estimated to be nonzero in approximately 50% of cases in 100 experiments. Besides, EXSZUSx (Switzerland/U.S. Foreign Exchange Rate), EXJPUSx (Japan/U.S. Foreign Exchange Rate), EXCAUSx (Canada/U.S. Foreign Exchange Rate) and AAA (Moody's Seasoned Aaa Corporate Bond Yield) also have relatively high selection frequencies. Corresponding average coefficients are given in Table 5.1.

Table 5.2 reports the average winning probability of each competing factor for the best performing model in terms of prediction accuracy in validation set and test set, respectively. The corresponding coefficients are reported in Table 5.3. Note that coefficients in competing factors with extremely low winning probabilities ($< 10^{-3}$) are omitted.

In Fig. 5.3, we provide pairwise scatter plots of response variable versus extracted competing factors. The extracted competing factors are based on the best performing model from the validation set in Table 5.3. If a linear relationship

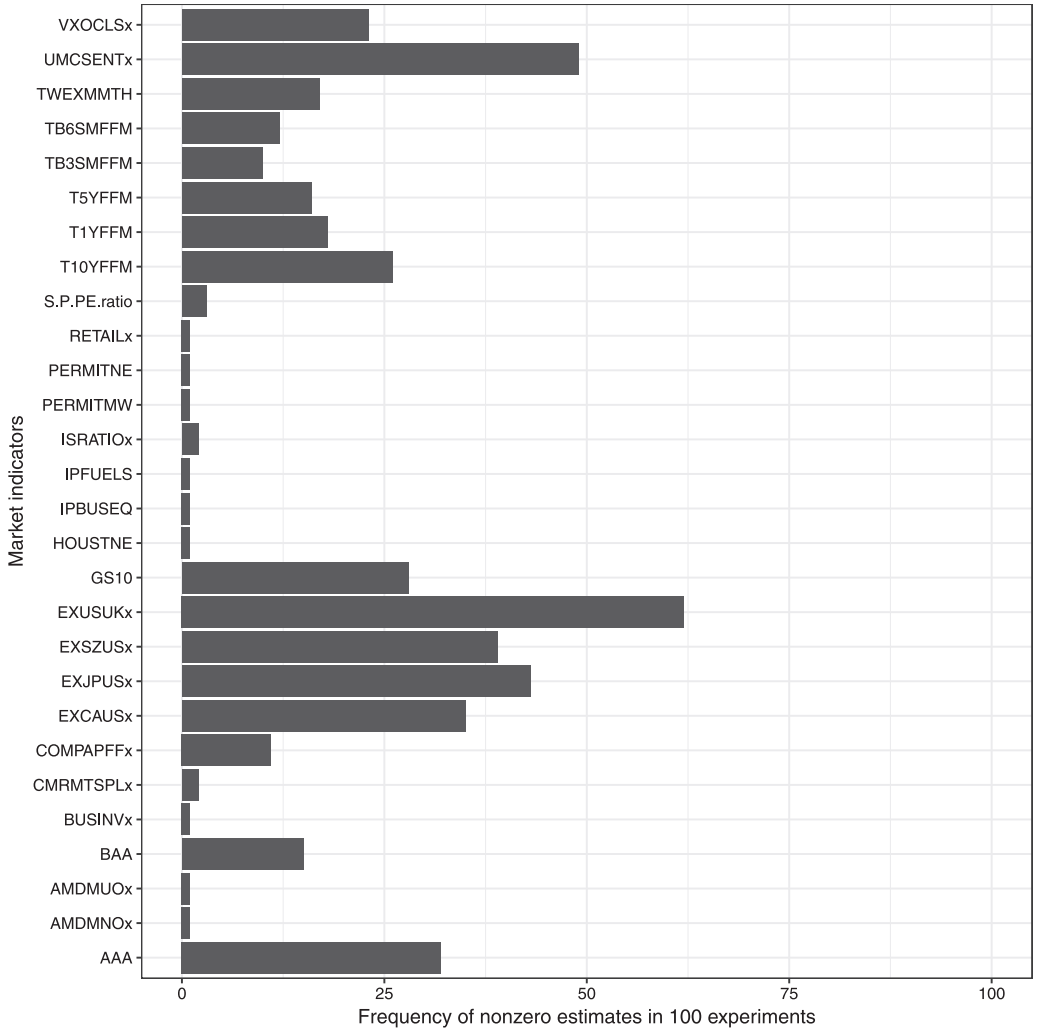

Fig. 5.2. (Fama–Bliss Bond Risk Premia). Frequency of nonzero estimates in 100 models. Predictors with 0 frequency are omitted.

Table 5.3
(Fama–Bliss bond risk premia). Coefficients for the best performing model.

Variable	HOUSTMW	VXOCLSx	EXJPUSx	EXCAUSx	UMCSENTx	EXSZUSx
CF	3	5	6	6	6	6
Coef. (Validation)	−0.003	0.625	−0.065	−0.0002	0.090	0
Coef. (Test)	0	0	0	0.013	0	0.033

between the response (denoted as Y) and extracted competing factors (denoted as CF) is plausible, it is expected to see that the scatter plot of Y vs. each CF shows relatively strong linear trend. However, what we observe from the scatter plots in the first columns of Figure 4 and Figure 6 is that there are several clouds of the points, where within each cloud, there is roughly a linear trend, while when clouds are combined, the linear trend is not quite clear. This phenomenon exactly tells that max linear structure is a better fit here comparing with linear model: max linear regression allows a part of samples to be dominated by a CF while the other parts being dominated by other CFs. The clouds in scatter plots show the existence of partition of samples and competing over competing factors.

We demonstrated that max-linear competing factor model can perform as well as random forest in terms of predicting bond risk premia by capturing the non-linearity and avoiding overfitting. Table 5.2 shows that the factor derived from Interest and Exchange Rates is in dominance. In 2017, its competing winning probability is as high as 0.967. From Table 5.3, one can see that the exchange rate of Canadian dollar to US dollar (EXCAUSx) and the exchange rate of Swiss franc to US dollar (EXSZUSx) have high impacts on Fama–Bliss bond risk premia prediction, which is consistently observed in Table 5.1. These empirical evidences derived from our model provide insights into risk premia evaluations. As a result, policy makers and market strategists can use our results as references in their decision making and portfolio management.

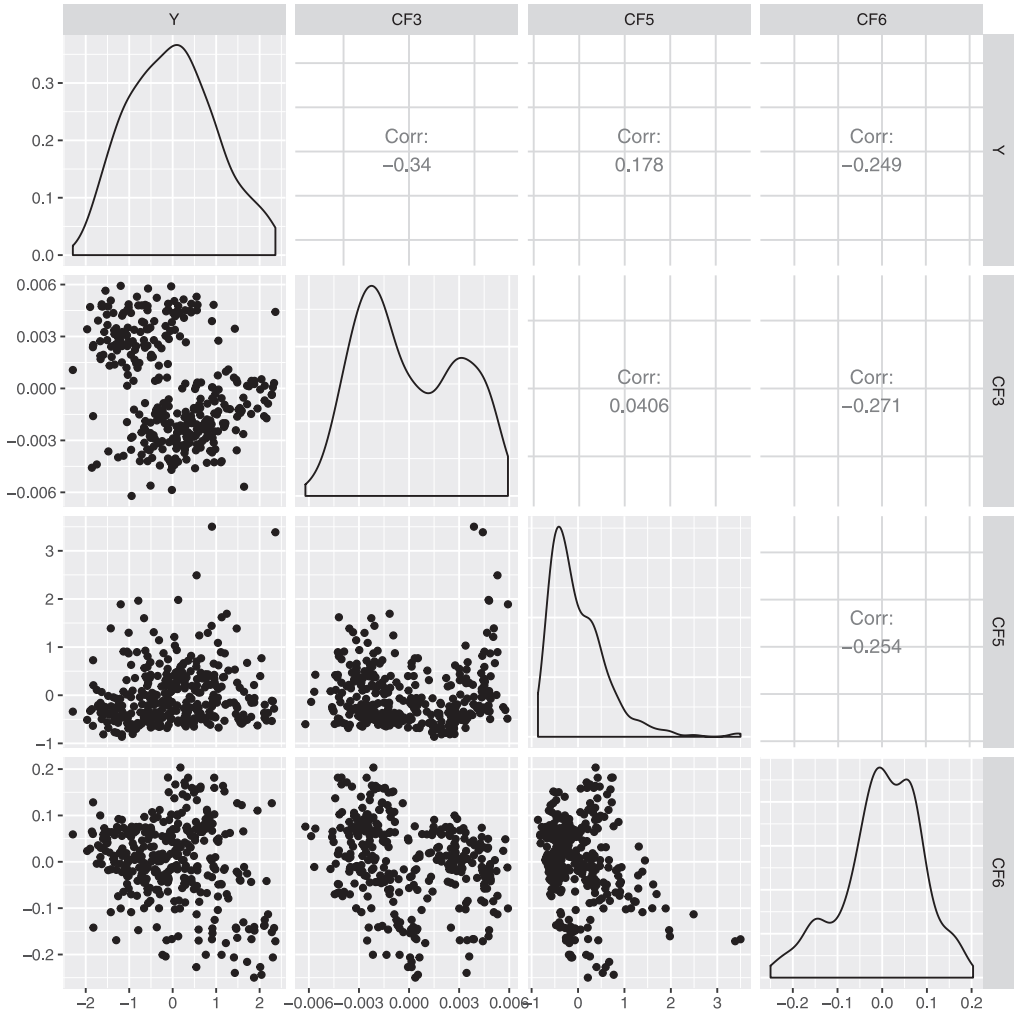

Fig. 5.3. (Bond risk premia) pairwise plots for response and extracted factors.

5.2. Max-linear modeling monthly maximum of VIX with market indicators

The CBOE Volatility Index (VIX), computed by the Chicago Board Options Exchange (CBOE), is a measure of the market expectation of volatility in the near future. Volatility is widely used as a measure of market risks in financial investments. For instance, volatility is one of the key factors determining the price of put and call options, as higher volatility implies higher probability that these options will expire in the money. Therefore, investors are willing to pay greater premiums for an option if the expected volatility is higher. Prediction of VIX or volatility often relies on call and put option prices, or the recent past volatility with the assumption that similar pattern of stock market will continue in the near future (Fleming et al., 1995). Different from the other existing research studies, Kapoor et al. (2017) predicted VIX using market indicators with linear regression.

In this study, we apply the max-linear model to the monthly maximum VIX, which reflects the largest fluctuation hence risk climax of stock market within each month. The monthly maximum VIX is computed from daily VIX from January 1990 to December 2017 available from the Federal Reserved Economic Data (FRED). The same market indicators in Section 5.1 are used as predictors. We exclude VXOCLSx (CBOE S&P 100 Volatility Index: VXO) because it was the original-formula of VIX based on S&P 100 (OEX) Index Options before 2003/09/22 (www.cboe.com/VXO), hence is highly correlated with the monthly maximum VIX.

To evaluate the prediction accuracy of monthly maximum VIX using different models, we perform the same data partitions as in the analysis procedure in Section 5.1, i.e., 100 models are generated using the training set. Best 10 models among the 100 are again detected using validation set and applied to forecast of the monthly maximum VIX in 2017.

The boxplots in Fig. 5.4 show that the sparse group lasso fails to capture the relationship between monthly maximum VIX and market indicators. In fact, the estimated parameters in sparse group lasso are all zeros. In the left panel, max-linear model performs the best on testing set while random forest performs better on validation set, signaling overfitting

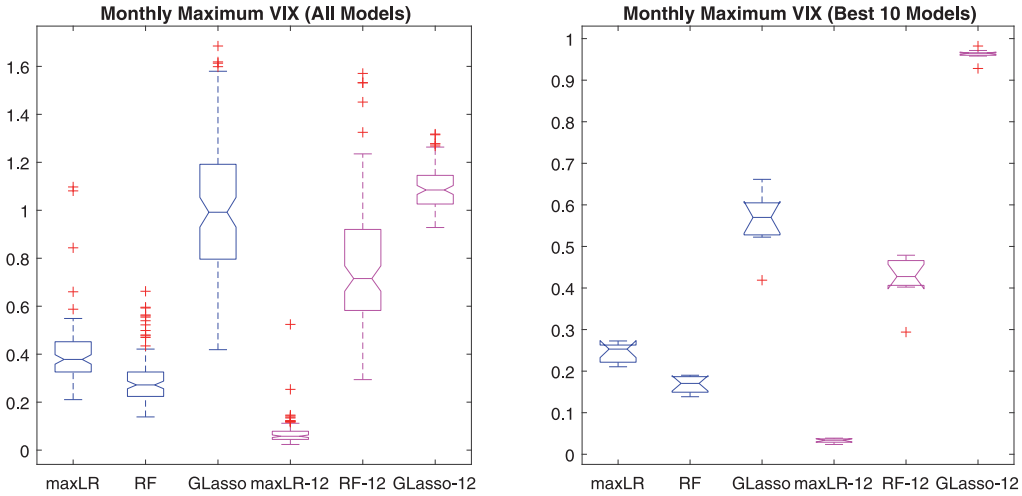

Fig. 5.4. (VIX). Prediction MSEs of validation and testing set (ended with "−12") for max-linear competing regression, random forest, and sparse group lasso in 100 random cases (left panel) and 10 best performing models respectively on validation set and testing set (right panel).

Table 5.4
(VIX). Average coefficients of monthly VIX maximum data.

Important variable	S.P.PE.ratio	S.P.div.yield	S.P.500	EXUSUKx	EXJPUSx	EXCAUSx
Average coef.	0.215	−0.817	−0.662	0.082	0.197	0.322

Table 5.5
(VIX). Average winning probability for best performing model.

Competing factor (CF)	Winning probability (Validation)	Winning probability (Test)
1. Output and income	0.026	0.038
2. Labor market	$<10^{-3}$	$<10^{-3}$
3. Consumption and orders	0.197	0.150
4. Orders and inventories	0.008	0.061
5. Money and credit	0.027	0.016
6. Interest and exchange rates	0.495	0.468
7. Price	0.019	0.020
8. Stock market	0.227	0.246

in random forest. Shown in the right panel of Fig. 5.4, the max-linear model outperforms all tested models on the testing set, winning over random forest which is renowned for its prediction performance.

Fig. 5.5 displays the proportion of non-zero coefficients out of the ensemble model for each predictor using similar method as Section 5.1, and we can see that a few variables are selected with probability greater than 75%, including S.P.PE.ratio (S&P's Composite Common Stock: Price-Earnings Ratio), S.P.div.yield (S&P's Composite Common Stock: Dividend Yield), S.P.500 (S&P's Common Stock Price Index: Composite), EXUSUKx (U.S. / U.K. Foreign Exchange Rate), EXJPUSx (Japan/U.S. Foreign Exchange Rate) and EXCAUSx (Canada/U.S. Foreign Exchange Rate). These results suggest that stock market and foreign currency exchange rates are the most important in predicting monthly maximum VIX. The average estimated coefficients for important variables are listed in Table 5.4.

Similar to Section 5.1, we now focus on the best-performing model in terms of prediction accuracy in the validation set and testing set, respectively. In Table 5.5 the average winning probability of each competing factor is reported. The corresponding coefficients are reported in Table 5.6. The best performing model in validation sets mainly uses variables in competing factors including Consumption and Orders, Interest and Exchange Rates, and Stock Market, while variables in Orders and Inventories have escalated impact in the best performing model for testing set.

Similar to Section 5.1, we provide pairwise scatter plots of response variable versus extracted competing factors. There are also several clouds of points in response variable versus extracted competing factors plots, showing that linear relationships solely are not adequate in modeling the relations between response and features. Clouds of scatter points show partition of samples exists, and max linear relations which allow different parts of samples to be dominated by different competing factors are more plausible in this scenario (see Fig. 5.6).

Viewing Table 5.5, we can see that among all eight competing factors, Interest and Exchange Rates has the highest winning probability, followed by Stock Market and Consumption and Orders. These empirical findings match what market analysis had believed in but yet to find supporting quantitative evidences technically, which is now unprecedentedly provided by the max-linear model. The impact of exchange rates on monthly maximum VIX can also be seen in Table 5.4,

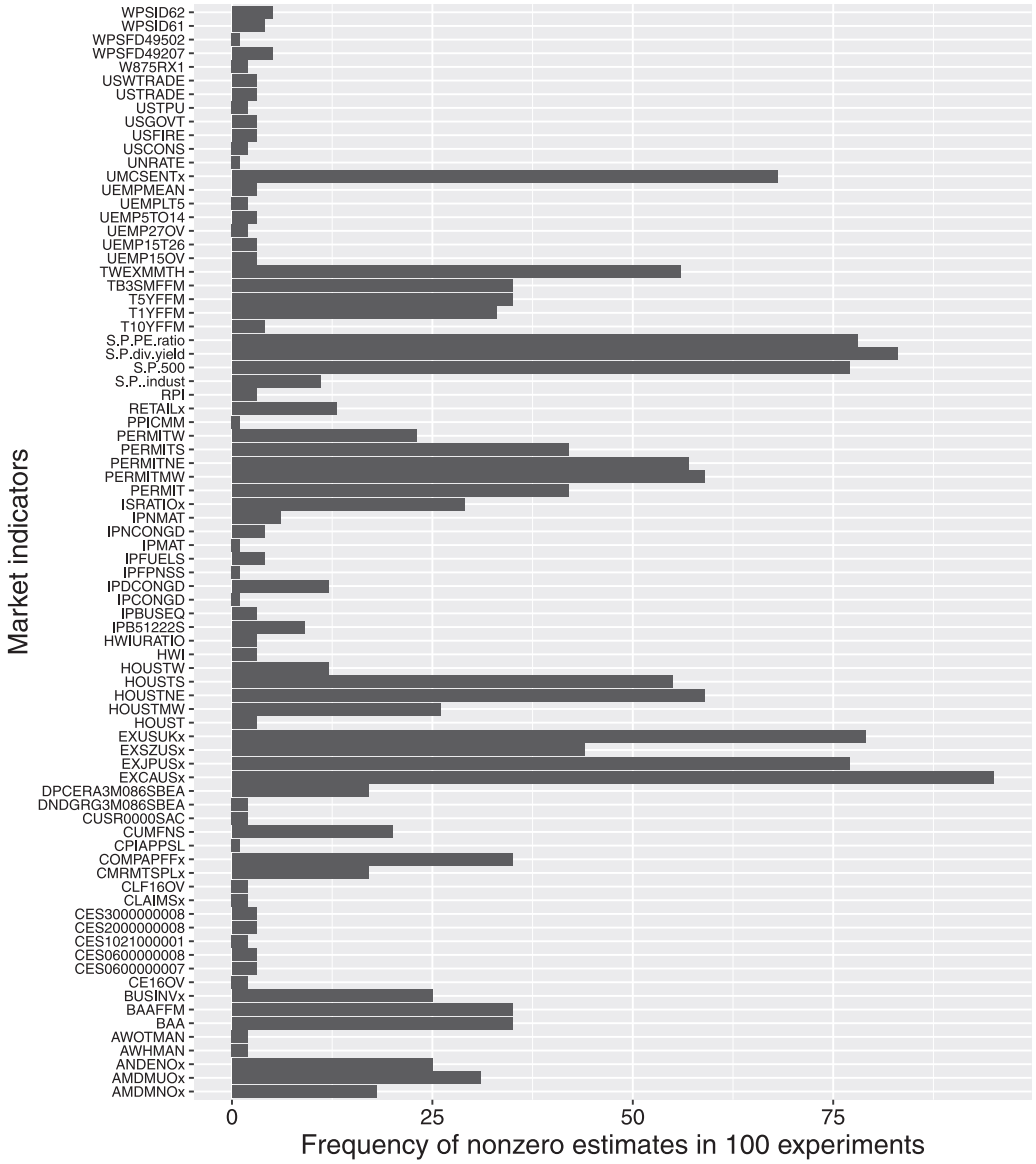

Fig. 5.5. (VIX). Frequency of nonzero estimates in 100 models. Predictors with 0 frequency are omitted.

Table 5.6
(VIX). Coefficients for best performing model.

Variable	HOUSTMW	HOUSTS	PERMIT	PERMITNE	PERMITMW	PERMITS	PERMITW
CF	3	3	3	3	3	3	3
Coef. (Validation)	−0.046	−0.072	0	−0.003	−0.059	−0.231	0
Coef. (Test)	0	−0.255	−0.267	0	−0.030	−0.063	−0.126
Variable	TWEXMMTH	EXJPUSx	EXUSUKx	EXCAUSx	UMCSENTx	S.P.500	S.P.div.yield
CF	6	6	6	6	6	8	8
Coef. (Validation)	0.380	0	0.066	0.235	−0.267	−0.630	−0.838
Coef. (Test)	0.020	0.210	0	0.301	−0.306	−1.040	−1.053
Variable	S.P.PE.ratio	AMDMNOx	ANDENOx	AMDMUOx	BUSINVx	ISRATIOx	
CF	8	4	4	4	4	4	
Coef. (Validation)	0.269	0	0	0	0	0	
Coef. (Test)	0.278	0.008	−0.042	−0.081	0.130	0.033	

namely, the exchange rates of EXUSUKx (US dollars to British pound), EXJPUSx (Japanese yen to US dollar), and EXCAUSx (Canadian dollar to US dollar). In the same table, we can see that a rise in S.P.500 (S&P500 index) or S.P.div.yield (S&P's Composite Common Stock: Dividend Yield) will lower the monthly maximum VIX, while S.P.ratio has the opposite effect

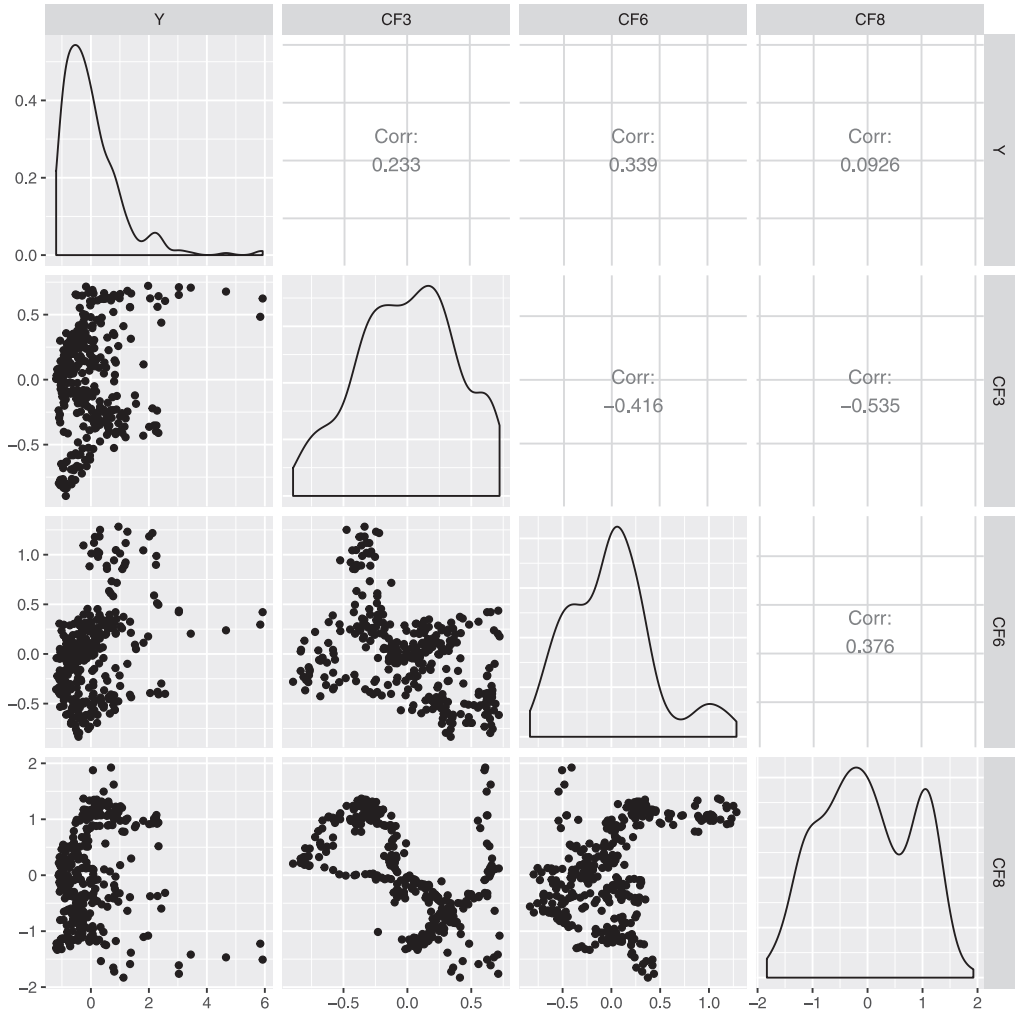

Fig. 5.6. (VIX) pairwise plots for response and extracted factors.

on average. The best performing model, in Table 5.6, reveals similar findings, with the exception of S.P.div.yield and S.P.PE.ratio (S&P's Composite Common Stock: Price-Earnings Ratio). The empirical evidence presented in Table 5.6 on the impact from Interest and Exchange Rates and Stock Market again matches common beliefs from various researches and reports, with our new model being able to deliver an overall market interpretation from the viewpoint of competing risk factors.

It is worth noting that, competing factor Interest and Exchange Rates takes lead in modeling both Fama–Bliss bond risk premia and monthly maximum VIX, with foreign currency exchange rates EXJPUSx and EXCAUSx shown up important in examples. On the other hand, competing factors Consumption and Orders and Stock Market contribute to monthly maximum VIX prediction prominently, while they are not deemed powerful in predicting risk premia.

5.3. Market values of soccer players

Soccer has been the most popular sport in the world for decades. According to the Fédération Internationale de Football Association (FIFA), thousands of registered soccer clubs play in different soccer leagues every season. After the end of a season, the transfer window opens for the soccer clubs to transfer players. Hundreds of millions of dollars change hands, and hundreds of players change teams during summer transfer window. Taking Premier League in England for example, seasons run from August to May. During the transfer window of summer 2018, the transfer expenses amount to £1,280,196,000, with 239 player arrivals and 255 departures.

Transfers involving top players with high market value are never less thrilling than goals or trophies. Player's market values vary greatly for different players, leagues, and periods of time. Although inauthoritative organizations, such as https://www.transfermarkt.com from Germany, publish estimated player values several times a season, an effective, accurate and updated evaluation system of player values has always been aspired. This market value evaluation system

can be a huge support to club's budgeting and decision making on player transfers. To establish such a system, many variables should be taken into the valuation assessment, from personal data such as age, height, and weight, to playing statistics reflecting the player's abilities, etc.

In this section, we model the relationship between the player's market value and playing attributes. Starting from 1995 the FIFA soccer games have a coherent scout of players worldwide. The data we collected from the full version of FIFA 18 (https://sofifa.com) contain statistics of players from the major soccer leagues, including personal data such as age, height and weight, team and contract, etc., and 20+ playing attributes each evaluated with a score between 0 and 100, such as shot accuracy, dribbling, aggression, goalkeeping, etc., together with current estimated market value (Cotta et al., 2016).

We select 5,110 non-goalkeeping soccer players from the database, with their market values (in million), age and 29 non-goalkeeping playing attributes. The attributes are naturally grouped by attacking attributes, skill attributes, physical and mental attributes, and defense attributes. We compare max-linear model with sparse group lasso, random forest and linear regression model. We used half of the observations for model training and the rest held out for testing prediction performance.

Table 5.7 shows the grouping of attributes used in the max-linear model and sparse group lasso. Our preliminary analysis shows a non-linear relationship between value and age, which coincides with our expectation of players' heyday during the mid-term of their career, so we add age into all four groups and model it with second-order polynomials. In addition, players' position is deterministic of a general trend of lower value for defenders as opposed to forwarders; hence we include two position indicators in all groups using dummy variables: frequencies of a player playing as a forwarder/defender, both categorized into three levels, low, medium and high.

We estimated the standard errors of the model coefficients using parametric bootstrap, i.e., simulating 500 sets of response variables under the max-linear model assumption with the group mean estimated by using the coefficients obtained from the EM-algorithm upon convergence and refitting the max-linear model with the simulated responses. We then took the standard errors of the 500 estimated coefficients from the bootstrap samples.

Based on the estimated coefficients and t values in Table 5.7, it is clear that age is one of the key ingredients in deciding players' market values, although the estimated coefficients of the age terms vary a lot for different competing factors, as players with different roles in squad have different "golden age" in their career path. Frequency of playing in a forward or defensive position is submerged by other variables in the group of physical and mental attributes but is crucial in the group of defending attributes. We observed that low frequency as a forwarder and high frequency as a defender lead to highest market value for players whose value is dominated by the group of defensive attributes.

We identify important features in predicting soccer players' market values within each group. Players whose value is determined by attacking attributes are valued higher with excellent skills in heading accuracy and short passing. Dribbling, long passing, and ball control are the more valuable skill set in the skill attributes. Reaction is most crucial for players with market value determined by physical and mental attributes. Sliding tackle ability is positively valued for those with value determined by the group of defensive attributes. We observed negative estimated coefficients in each group, often a result of collinearity between predictor variables.

The prediction errors of the selected models on the testing set are reported in Table 5.8. Max-linear model outperforms sparse group lasso and linear regression, whereas is not as good as random forest as far as prediction accuracy is concerned. Note that we are currently using natural grouping of predictor variables. If an optimal grouping strategy can be identified, we expect further improvement on prediction accuracy for max-linear models. In addition, the max-linear regression model preserves explicit and easily interpretable structure, which is impossible for random forest model.

6. Discussions

The max-linear model is a natural extension of linear regression model, where it keeps linear structure within each competing factor of independent variables, and employs the max relationship to allow non-Gaussian, asymmetry, and competition across the factors. It maintains the attractive feature of interpretability in terms of variable contributions to the response variables.

The max-linear formulation has a natural form of augmentation. By augmenting the latent variable of the dominant factor, we find the maximum-likelihood estimator of parameters via EM approach, whose probabilistic properties including consistency and asymptotic normality are derived under certain regularity conditions. In high dimensional settings, penalizing group parameter and coefficients achieves effective variable and group selection. Under true max-linear relationships, the max-linear model enjoys excellent performances compared with the group lasso and random forest. Meanwhile, when the true model is a linear regression model, the max-linear model has comparable performance with that of the group lasso.

In the literature, Zhou and Zhu (2010) considered the hierarchical lasso, where the coefficients were decomposed into a multiplicative form of a group level parameter that controls the signal embedded in a group of variables, and individual parameters reflecting the importance of each predictor within a group. Penalizations were then applied in order to not only select important predictors, but also important groups. The idea of group lasso is the realization of a similar structure under linearity and symmetry as the max-linear model, where groups are connected with maximum rather than linear relationship, circumventing the identifiable issues as long as the groups of predictors are not completely overlapping.

In this paper, we assume that the independent variables are naturally grouped. When the grouping of features is not available, an efficient grouping strategy will be very helpful to enhance the performance of the max-linear model,

Table 5.7
Coefficients of FIFA player data(with bootstrap t-values).

Group1	Coef (t-value)	Group2	Coef (t-value)	Group3	Coef (t-value)	Group4	Coef (t-value)
σ_1^2	0.097(209.62)	σ_2^2	0.086(231.46)	σ_3^2	1.709(9.41)	σ_4^2	0.002(10415.78)
β_{01}	−0.606(−102.37)	β_{02}	−0.895(−117.81)	β_{03}	−3.440(−27.72)	β_{04}	−0.560(−7300.31)
age	0.357(49.98)	age	0.778(108.33)	age	0(0)	age	0.184(1382.85)
age^2	−0.544(−84.12)	age^2	−0.936(−131.05)	age^2	−0.942(−4.253)	age^2	−0.230(−2042.13)
attack_ low	−0.013(−4.56)	attack_ low	−0.009(−1.53)	attack_ low	0.230(3.54)	attack_ low	0.001(18.49)
attack_ medium	0.013(3.80)	attack_ medium	−0.043(−14.07)	attack_ medium	−0.021(−0.45)	attack_ medium	−0.001(−19.55)
defend_ low	−0.007(−1.52)	defend_ low	−0.049(−14.92)	defend_ low	0.085(1.50)	defend_ low	−0.002(−19.73)
defend_ medium	−0.022(−7.15)	defend_ medium	−0.039(−9.86)	defend_ medium	0.087(1.61)	defend_ medium	−0.004(−50.10)
Crossing	−0.006(−1.758)	Dribbling	0.414(35.56)	Acceleration	0.186(1.33)	Marking	−0.024(−78.90)
Finishing	−0.017(−3.17)	Curve	0.049(7.34)	Sprint speed	0.332(2.51)	Standing tackle	−0.001(−2.08)
Heading accuracy	0.357(50.27)	Freekick accuracy	−0.008(−1.54)	Agility	0.313(2.53)	Sliding tackle	0.006(19.37)
Short passing	0.161(39.83)	Long passing	0.133(27.81)	Reaction	1.961(21.90)		
Volleys	−0.020(−3.73)	Ball control	0.592(58.35)	Balance	−0.116(−1.46)		
Shot power	0.060(14.17)			Jumping	0.006(0.12)		
Long shot	−0.041(−7.244)			Stamina	0.029(0.51)		
				Strength	0.647(7.52)		
				Aggression	−0.056(−0.92)		
				Interceptions	0(0)		
				Positioning	−0.145(−1.82)		
				Vision	0.062(0.91)		
				Penalties	0.081(1.23)		
				Composure	1.021(12.82)		

Table 5.8
Results of FIFA player data.

Methods	Max-linear	Random forest	Sparse group lasso	Linear regression
Prediction error (MSE)	0.236	0.118	0.508	0.433

especially when there is a strong nonlinear relationship. One simple approach is to form groups using unsupervised clustering on samples (Friedman et al., 2001), while Dettling and Bühlmann (2004) used penalized logistic regression based analysis and combined feature selection, feature grouping, and sample classification in a supervised, simultaneous way. More recently, Bondell and Reich (2008) and Zhong and Kwok (2012) proposed and developed OSCAR (Octagonal Shrinkage and Clustering Algorithm for Regression) respectively, which deals with feature grouping problem when features are not ordered. The listed methods mainly combine similar features into single groups and hence could be very helpful for further study in feature selection methods such as fused lasso (Tibshirani et al., 2005) etc. However in our model, grouping features based on similarity of features may not be a good choice. In fact, features in the same competing factor need not behave in concordant manner. This idea could be seen in the soccer player market value data application in Section 5.3, where two features "balance" and "reactions" in the same competing factor "physical and mental attributes" have as low correlation as 0.14. Separating highly correlated features into different competing factors could be even more plausible for max linear regression, as the max operator only selects one competing factor for each sample. If highly correlated features are separately distributed, they will not dominate the model simultaneously, and problems caused by multi-collinearity can be reduced. Therefore, forming competing factors only based on similarity of features, which is suitable for group lasso or fused lasso, may not be the best choice for forming competing factors in max linear regressions. The grouping strategy for max linear regression is more like an optimization problem, rather than a clustering problem, which requires a lot of work and advanced optimization techniques. In this paper, we decided to use natural groups of features to format competing factors, as in Yuan and Lin (2006) and Fan et al. (2021). This choice is indeed one limitation of the proposed model so far, but can be an interesting research direction which further improves the flexibility and prediction power of the max linear models.

There are many possible extensions of the max-linear model. On one hand, further theoretical developments related to our proposed EM estimation procedure are needed to enhance the understanding of group selections and within groups variable selections. On the other hand, different estimation and inference procedures can be an interesting research direction. The max-linear structure can be applied to generalized linear models including logistic regression or Poisson regression. In survival analysis, one can also embed the max-linear structure in the Cox proportional-hazards model to allow predictors to compete with each other. In real applications, we have showed some empirical evidences of the impacts of economic indicators to Fama–Bliss bond risk premia and monthly maximum VIX values. These empirical findings need further exploration, qualitatively and quantitatively, as they may influence market decisions. The analysis of soccer players' values serves as a reference for sport business. We believe that the max-linear structure will open a new chapter in machine learning literature.

Acknowledgments

The work by Cui was partially supported by NSF-DMS-1505367 and Wisconsin Alumni Research Foundation #MSN130-403. The work by Xu was partially supported by NSF-DMS-1505367 and Wisconsin Alumni Research Foundation #MSN215758. The work by Zhang was partially supported by NSF-DMS-1505367 and NSF-DMS-2012298. The supports from the 2018 International Symposium on Financial Engineering and Risk Management (FERM 2018, Shanghai) and NNSFC 71991471 are also acknowledged.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2020.07.017.

References

Bates, D., Watts, D., 1988. Nonlinear regression analysis and its applications. In: Wiley series in probability and mathematical statistics, Wiley, New York [u.a.].
Bondell, H.D., Reich, B.J., 2008. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. Biometrics 64 (1), 115–123.
Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.
Cotta, L., de Melo, P., Benevenuto, F., Loureiro, A.A., 2016. Using fifa soccer video game data for soccer analytics. In: Workshop on Large Scale Sports Analytics.
Cui, Q., Zhang, Z., 2018. Max-linear competing factor models. J. Bus. Econom. Statist. 36 (1), 62–74.
Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the em algorithm. J. R. Stat. Soc. Ser. B (Methodol.) 1–38.
Dettling, M., Bühlmann, P., 2004. Finding predictive gene groups from microarray data. J. Multivariate Anal. 90 (1), 106–131.
Fan, J., Gijbles, I., 1996. Local Polynomial Smoothing. Chapman and Hall, London.

Fan, J., Ke, Y., Liao, Y., 2021. Augmented factor models with applications to validating market risk factors and forecasting bond risk premia. J. Econometrics 222, 269–294.
Fan, J., Wang, W., Zhong, Y., 2019. Nonlinear and non-gaussian state-space modeling with monte carlo simulations. J. Econometrics 208, 5–22.
Fleming, J., Ostdiek, B., Whaley, R.E., 1995. Predicting stock market volatility: A new measure. J. Futures Mark. 15 (3), 265–302.
Friedman, J., Hastie, T., Tibshirani, R., 2001. The elements of statistical learning, Vol. 1. Springer series in statistics New York.
Heffernan, J.E., Tawn, J.A., Zhang, Z., 2007. Asymptotically (in) dependent multivariate maxima of moving maxima processes. Extremes 10 (1–2), 57–82.
Henderson, D.J., Parmeter, C.F., 2015. Applied nonparametric econometrics. Cambridge University Press.
Kapoor, V., Khedkar, N., O'Keefe, J., Qiao, I., Venkatesan, S., Laghate, S., 2017. Predicting volatility in the s&p 500 through regression of economic indicators.
Linton, O., Xiao, Z., 2019. Efficient estimation of nonparametric regression in the presence of dynamic heteroskedasticity. J. Econometrics http://dx.doi.org/10.1016/j.jeconom.2019.01.016.
Ludvigson, S.C., Ng, S., 2009. A factor analysis of bond risk premia. Technical report, National Bureau of Economic Research.
Malinowski, A., Schlather, M., Zhang, Z., 2016. Intrinsically weighted means and non-ergodic marked point processes. Ann. Inst. Statist. Math. 68 (1), 1–24.
McLachlan, G., Krishnan, T., 2007. The EM Algorithm and Extensions, Vol. 382. John Wiley & Sons.
Naveau, P., Zhang, Z., Zhu, B., 2011. An extension of max autoregressive models. Stat. Interface 4 (2), 253–266.
Pancheva, E., 1989. Max-stability. Theory Probab. Appl. 33 (1), 155–158.
Patton, A.J., Timmermann, A., 2007. Properties of optimal forecasts under asymmetric lossand nonlinearity. J. Econometrics 140, 884–918.
Tanizaki, H., Mariano, R.S., 1998. Nonlinear and non-gaussian state-space modeling with monte carlo simulations. J. Econometrics 83 (1–2), 263–290.
Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., Knight, K., 2005. Sparsity and smoothness via the fused lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 67 (1), 91–108.
Wahba, G., 1990. Spline models for observational data, Vol. 59. Siam.
Wand, M.P., Jones, M., 1994. Kernel smoothing. CRC Press.
Wu, C.J., et al., 1983. On the convergence properties of the em algorithm. Ann. Statist. 11 (1), 95–103.
Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. Ser. B Stat. Methodol. 68 (1), 49–67.
Zhang, Z., 2008. Quotient correlation: a sample based alternative to Pearson's correlation. Ann. Statist. 36 (2), 1007–1030.
Zhang, Z., 2009. On approximating max-stable processes and constructing extremal copula functions. Stat. Inference Stoch. Process. 12, 89–114.
Zhang, Z., Peng, L., Idowu, T., 2016. Max-autoregressive and moving maxima models for extremes. In: Extreme Value Modeling and Risk Analysis: Methods and Applications. CRC Press, pp. 153–178.
Zhang, Z., Qi, Y., Ma, X., et al., 2011. Asymptotic independence of correlation coefficients with application to testing hypothesis of independence. Electron. J. Stat. 5, 342–372.
Zhang, Z., Zhang, C., Cui, Q., 2017. Random threshold driven tail dependence measures with application to precipitation data analysis. Statist. Sinica 27 (2), 685–709.
Zhang, Z., Zhu, B., 2016. Copula structured m4 processes with application to high-frequency financial data. J. Econometrics 194 (2), 231–241.
Zhong, L.W., Kwok, J.T., 2012. Efficient sparse modeling with automatic feature grouping. IEEE Trans. Neural Netw. Learn. Syst. 23 (9), 1436–1447.
Zhou, N., Zhu, J., 2010. Group variable selection via a hierarchical lasso and its oracle property. Stat. Interface 3, 57–574.