MONASH University

MONASH BUSINESS SCHOOL

ISSN 1440-771X

Department of Econometrics and Business Statistics

http://monash.edu/business/ebs/research/publications

Sparse Multiple Index Models for High-dimensional Nonparametric Forecasting

Nuwani Palihawadana, Rob J Hyndman, Xiaoqian Wang

May 2024

Working Paper no/yr

AACSB ACCREDITED EQUIS ACCREDITED ASSOCIATION OF MBAS ACCREDITED

Sparse Multiple Index Models for High-dimensional Nonparametric Forecasting

Nuwani Palihawadana
Department of Econometrics & Business Statistics
Monash University
Clayton VIC 3800
Australia
Email: nuwani.kodikarapalihawadana@monash.edu
Corresponding author

Rob J Hyndman
Department of Econometrics & Business Statistics
Monash University
Clayton VIC 3800
Australia
Email: rob.hyndman@monash.edu

Xiaoqian Wang
Department of Econometrics & Business Statistics
Monash University
Clayton VIC 3800
Australia
Email: xiaoqian.wang@monash.edu

14 May 2024

JEL classification: C10,C14,C22

Sparse Multiple Index Models for High-dimensional Nonparametric Forecasting

Abstract

Forecasting often involves high-dimensional predictors, which have nonlinear relationships with the outcome of interest. Nonparametric additive index models can capture these relationships, while addressing the curse of dimensionality. This paper introduces a new algorithm, ***S***parse ***M***ultiple ***I***ndex (***SMI***) ***Modelling***, tailored for estimating high-dimensional nonparametric/semi-parametric additive index models, while limiting the number of parameters to estimate, by optimising predictor selection and predictor grouping. The SMI Modelling algorithm uses an iterative approach based on mixed integer programming to solve an ℓ_0-regularised nonlinear least squares optimisation problem with linear constraints. We demonstrate the performance of the proposed algorithm through a simulation study, along with two empirical applications to forecast heat-related daily mortality and daily solar intensity.

Keywords: Additive index models; Variable selection; Dimension reduction; Predictor grouping; Mixed integer programming.

1 Introduction

Forecasts are often contingent on a very long history of predictors, which are nonlinearly related to the variable of interest. For example, when forecasting half-hourly electricity demand, it is common to use at least a week of historical half-hourly temperatures and other weather observations (Hyndman & Fan 2010). The relationships between the lagged temperatures and electricity demand are nonlinear (due to both heating and cooling effects), and involve complex interactions due to thermal inertia in buildings (Fan & Hyndman 2012). Similarly, when forecasting bore levels, rainfall data from up to thousand days earlier can impact the result (Peterson & Western 2014; Bakker & Schaars 2019; Rajaee, Ebrahimi & Nourani 2019) due to the complex nonlinear flow dynamics of rainfall into aquifers.

These examples suggest a possible nonlinear "*transfer function*" model of the form

$$y_t = f(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \ldots, \boldsymbol{x}_{t-p}, y_{t-1}, \ldots, y_{t-k}) + \varepsilon_t, \tag{1}$$

where y_t is the observation of the response variable at time t, $\boldsymbol{x}_t$ is a vector of predictors at time t, and ε_t is an i.i.d. random error. By including lagged values of y_t along with the lagged predictors, we allow for any serial correlation in the data.

The form of f is typically nonlinear, and involves complicated interactions with a high value of p (and possibly also large k). It is infeasible to estimate f in high-dimensional settings (where p is large) due to the curse of dimensionality (Bellman 1957; Stone 1982). Instead, we normally impose some form of additivity constraint, and ignore interactions of more than two or three variables. There are also numerous ad hoc model choices in selecting the appropriate predictors to include.

For example, Fan & Hyndman (2012) proposed a ***semi-parametric additive model*** to obtain short-term forecasts of the half-hourly electricity demand for power systems in the Australian National Electricity Market. In this model, f is assumed to be fully additive, and is used to capture the effects of recent predictor values on the demand. The main objective is to allow nonparametric components in a regression-based modelling framework, particularly to address serially correlated errors. The model fitted for each half-hourly period (q) can be written as

$$\log(y_{t,q}) = h_q(t) + f_q(\boldsymbol{w}_{1,t}, \boldsymbol{w}_{2,t}) + \sum_{j=1}^{k} a_{q,j}(y_{t-j}) + \varepsilon_t. \tag{2}$$

The term $h_q(t)$ models several calendar effects as either linear or smooth terms. Temperature effects are modelled using the nonparametric component $f_q(\boldsymbol{w}_{1,t}, \boldsymbol{w}_{2,t})$, where $\boldsymbol{w}_{i,t} = [w_{i,t}, \ldots, w_{i,t-p}]'$ is a vector of lagged temperatures at site i. The terms $a_{q,j}(y_{t-j})$ capture the lagged effects of the response. Note that the error term ε_t is serially uncorrelated within each half-hourly model, because the serial correlation is eliminated by the inclusion of lagged responses in the model. However, some correlation may still exist between residuals from various half-hourly models (Fan & Hyndman 2012).

Similarly, a ***distributed lag model*** was proposed by Wood (2017) to forecast daily death rates in Chicago using measurements of several air pollutants. The response variable is modelled via a sum of smooth functions of lagged predictor variables, which is quite similar to the semi-parametric additive model used by Fan & Hyndman (2012). However, unlike in Fan & Hyndman (2012), Wood (2017) suggested allowing the smooth functions for lags of the same covariate to vary smoothly over lags, preventing large differences in estimated effects between adjacent lags. Thus, the model is of the form

$$\log(y_t) = f_1(t) + \sum_{k=0}^{K} f_2(p_{t-k}, k) + \sum_{k=0}^{K} f_3(o_{t-k}, w_{t-k}, k),$$

where y_t is the death rate at day t, f_1 is a nonparametric term to capture the *time* effect, and p_t, o_t, and w_t are various predictor variables. The model incorporates the current value ($k = 0$) and several

lagged values ($k = 1, \ldots, K$) of the predictors, where the distributed lag effect of a single predictor variable, and of the interaction of two predictor variables are captured by the sums $\sum_{k=0}^{K} f_2(p_{t-k}, k)$ and $\sum_{k=0}^{K} f_3(o_{t-k}, w_{t-k}, k)$ respectively.

Further examples include Ho, Chen & Hwang (2020), who used semi-parametric additive models to estimate ground-level $PM_{2.5}$ concentrations in Taiwan, while nonparametric additive models were utilised by Ibrahim et al. (2023) for predicting census survey response rates. Ravindra et al. (2019) provided a comprehensive review of the applications of additive models in environmental data, with a special focus on air pollution, climate change, and human health related studies.

In this paper, we are interested in high-dimensional applications that exhibit complicated interactions among predictors, particularly in the presence of a large number of lagged variables, and correlated errors. In such situations, *index models* prove beneficial in improving the flexibility of the broader class of nonparametric additive models (Radchenko 2015), while mitigating the difficulty of estimating a nonparametric component for each individual predictor.

While index models have been used to address problems from diverse application areas, there are several unresolved issues in their implementation. In this paper, we attempt to address two of those issues. First, the estimation of the model is challenging in high-dimensional settings due to the large number of nonparametric components to be estimated. Second, there is a noticeable subjectivity in selecting predictor variables (from the available predictors) for the model, and in identifying which terms should be grouped together to model interactions. In most of the applications discussed above, the choices were based on empirical explorations or domain expertise.

We propose to address these issues using a Sparse Multiple Index (SMI) model with automatic variable selection and grouping. This semi-parametric model can be written as

$$y_i = \beta_0 + \sum_{j=1}^{p} g_j(\boldsymbol{\alpha}_j^T \boldsymbol{x}_{ij}) + \sum_{k=1}^{d} f_k(w_{ik}) + \boldsymbol{\theta}^T \boldsymbol{u}_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{3}$$

where y_i is the univariate response, β_0 is the model intercept, $\boldsymbol{x}_{ij} \in \mathbb{R}^{l_j}$, $j = 1, \ldots, p$ are p subsets of all the predictors entering indices, $\boldsymbol{\alpha}_j$ is a vector of index coefficients corresponding to the index $h_{ij} = \boldsymbol{\alpha}_j^T \boldsymbol{x}_{ij}$, g_j is a smooth nonlinear function (possibly estimated by a spline). Note that we also allow for the inclusion of predictors that do not enter any of the indices, including covariates w_{ik} that relate to the response through the nonlinear functions f_k, $k = 1, \ldots, d$, and linear covariates denoted by $\boldsymbol{u}_i$. Although our interest is in forecasting time series data, the model can be used more widely, and so we have not included any notation specific to time series in the model formulation.

The SMI model subsumes the models discussed above, incorporating fully additive models (Wood 2011, 2017), where each predictor has its own index, and single index models (Stoker 1986; Härdle, Hall & Ichimura 1993; Radchenko 2015), where all predictors are included in a single index. The greater generality allows us to address the two issues mentioned earlier. First, the number of parameters to estimate is reduced by combining variables using linear indices, and by grouping predictors into indices to constrain the order and form of interactions. In our model formulation, both the number of indices p and the predictor grouping among indices are unknown. We propose algorithmic selection of the predictors to include in each index, thereby reducing the subjectivity in model formulation. We assume that no predictor enters more than one index (i.e. overlapping of predictors among indices is not allowed).

To our knowledge, no previous research has been done to explore how predictor choices can be made more objective and principled in nonparametric additive index models. Hence, our goal is to develop a methodology for optimal predictor selection and grouping in the context of high-dimensional nonparametric additive index models. Moreover, due to computational advancements in the field, the use of mathematical optimisation in solving statistical problems has gained a lot of recent interest (Theußl, Schwendinger & Hornik 2020). This motivated us to develop a variable selection algorithm based on mathematical optimisation techniques.

It is crucial to point out that any variable selection methodology naturally renders inferential statistics invalid, since we do not assume that the resulting model obtained through the variable selection procedure represents the true data generating process. Hence, our focus is only on improving forecasts, but not on making inferences on the resulting parameter estimates.

The rest of this paper is organised as follows. Section 2 presents our proposed *Sparse Multiple Index Model* and describes the algorithm for variable selection and grouping, and the estimation procedure. In Section 3, we demonstrate the functionality and the characteristics of the proposed algorithm through a simulation experiment. Section 4 illustrates two empirical applications of the proposed estimation and variable selection methodology, related to forecasting heat exposure related daily mortality and daily solar intensity. Some benchmark comparison methods are briefly introduced in Section 4.1. Concluding remarks are given in Section 5.

2 Sparse Multiple Index Model

2.1 Optimisation Problem Formulation

We implement variable selection for the proposed ***Sparse Multiple Index*** (SMI) model (Equation 3) by allowing for zero index coefficients for predictors. Suppose we observe $y_1, \ldots, y_n$, along with a set of potential predictors, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, with each vector $\boldsymbol{x}_i$ containing q predictors. The optimisation problem we seek to address is of the form below, where the sum of the squared error of the model (Equation 3) is minimised together with an ℓ_0 penalty term and an ℓ_2 (ridge) penalty term:

$$\begin{aligned} \min_{\beta_0, p, \boldsymbol{\alpha}, \boldsymbol{g}, \boldsymbol{f}, \boldsymbol{\theta}} \quad & \sum_{i=1}^{n} \left[y_i - \beta_0 - \sum_{j=1}^{p} g_j(\boldsymbol{\alpha}_j^T \boldsymbol{x}_i) - \sum_{k=1}^{d} f_k(w_{ik}) - \boldsymbol{\theta}^T \boldsymbol{u}_i \right]^2 \\ & + \lambda_0 \sum_{j=1}^{p} \sum_{m=1}^{q} \mathbb{1}(\alpha_{jm} \neq 0) + \lambda_2 \sum_{j=1}^{p} \|\boldsymbol{\alpha}_j\|_2^2 \\ \text{such that} \quad & \sum_{j=1}^{p} \mathbb{1}(\alpha_{jm} \neq 0) \in \{0, 1\} \quad \forall m, \end{aligned} \tag{4}$$

where $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_p^T]^T$, $\boldsymbol{g} = \{g_1, g_2, \ldots, g_p\}$, $\boldsymbol{f} = \{f_1, f_2, \ldots, f_d\}$, $\mathbb{1}(\cdot)$ is the indicator function, $\lambda_0 > 0$ is a tuning parameter that controls the number of selected predictors entering indices, and $\lambda_2 \geq 0$ is another tuning parameter that controls the strength of the additional shrinkage imposed on the estimated index coefficients. The constraint ensures that every predictor can only have non-zero coefficient in at most one index.

Applying an ℓ_2-penalty in addition to the ℓ_0-penalty is motivated by related literature (Hazimeh & Mazumder 2020; Mazumder, Radchenko & Dedieu 2023; Hazimeh, Mazumder & Radchenko 2023), where it is suggested that the prediction performance of best-subset selection is enhanced by the inclusion of an additional ridge penalty, especially when there is a low signal-to-noise ratio (SNR).

To solve the optimisation problem in Equation 4, we present a big-M based ***Mixed Integer Quadratic Programming*** (MIQP) formulation:

$$\begin{aligned} \min_{\beta_0, p, \boldsymbol{\alpha}, \boldsymbol{g}, \boldsymbol{f}, \boldsymbol{\theta}, \boldsymbol{z}} \quad & \sum_{i=1}^{n} \left[y_i - \beta_0 - \sum_{j=1}^{p} g_j(\boldsymbol{\alpha}_j^T \boldsymbol{x}_i) - \sum_{k=1}^{d} f_k(w_{ik}) - \boldsymbol{\theta}^T \boldsymbol{u}_i \right]^2 + \lambda_0 \sum_{j=1}^{p} \sum_{m=1}^{q} z_{jm} + \lambda_2 \sum_{j=1}^{p} \sum_{m=1}^{q} \alpha_{jm}^2 \\ \text{s.t.} \quad & |\alpha_{jm}| \leq M z_{jm} \quad \forall j, \forall m, \\ & \sum_{j=1}^{p} z_{jm} \leq 1 \quad \forall m, \\ & z_{jm} \in \{0, 1\}, \end{aligned} \tag{5}$$

where $j = 1, \ldots, p$, and $m = 1, \ldots, q$. We have introduced here binary variables $z_{jm} = \mathbb{1}(\alpha_{jm} \neq 0)$ to indicate in which index (if any) each predictor enters. The pre-specified *big-M parameter* is denoted by $M < \infty$, and it should be sufficiently large. If $\boldsymbol{\alpha}^*$ is the optimal solution to the problem given in Equation 5, then the big-M parameter should satisfy $\max\left(\{|\alpha^*_{jm}|\}_{j\in[p],m\in[q]}\right) \leq M$. The big-$M$ constraint ensures that α_{jm} is zero if and only if z_{jm} is zero, and if $z_{jm} = 1$, then $|\alpha_{jm}| \leq M$. At the same time, the ℓ_0-penalty term influences some of the binary variables z_{jm} to be zero, while the ℓ_2-penalty term enforces additional shrinkage on the estimated coefficients. Together, these components achieve variable selection.

2.2 Estimation Algorithm

We now show how to efficiently find a minimiser for the problem given in Equation 5. Since the number of indices p, the vector of index coefficients $\boldsymbol{\alpha}$, and the set of nonparametric functions $\boldsymbol{g}$ are all unknown, it is impossible to solve the above MIQP given in Equation 5 directly. Hence, we propose an iterative algorithm to solve the problem.

Initialising the Index Structure and Index Coefficients

We first need to provide a feasible initialisation for the index structure (i.e. the number of indices p and the grouping of predictors among indices) as well as for the index coefficients ($\boldsymbol{\alpha}$) of the model.

Based on several experiments, we propose three alternative methods for initialising the SMI model as follows.

1. **PPR: Projection Pursuit Regression Based Initialisation**

   A Projection Pursuit Regression model (Friedman & Stuetzle 1981) is a multiple index model, where each index includes all the available predictors. Since the proposed SMI model requires that there are no overlapping indices, it is impossible to use an estimated PPR model directly as a starting model for the algorithm. Thus, we follow the steps presented below to come up with a feasible initialisation for the index structure and the index coefficients.

   a. Scale all the variables of the data set by dividing each variable by its standard deviation (so that it is possible to compare the estimated coefficients among predictors).
   b. Fit a PPR model and obtain estimated index coefficients. (The user can decide the number of initial indices p^* to be estimated; we use $p^* = 5$ in our simulations and applications.)
   c. Calculate a threshold $\tau = 0.1 \times \max(\text{abs}(\text{PPR index coefficients}))$.
   Should it be: max(abs(PPR index coefficients))? - **Corrected**

d. Set to zero all coefficients whose absolute values fall below the calculated threshold.
Should it be: ...whose absolute values...? - **Corrected**

e. For predictors appearing in multiple indices, assign them to the index with the maximum absolute coefficient and zero out their coefficients in other indices.

f. After performing the above steps a-e, if any originally estimated index has all zero coefficients, it will be excluded from the model.

Now, the index structure and the index coefficients obtained through the above steps are considered to be a feasible initialisation for the proposed algorithm. Once the optimal SMI model is obtained through the algorithm, each index coefficient will be back-transformed to the original scale of the respective predictor variable, reversing the scaling effect applied at the beginning.

2. **Additive: Nonparametric Additive Model Based Initialisation**

   As a fully additive model is a special case of the SMI model, we can set $p = q$ and assign each predictor to its own index.

3. **Linear: Linear Regression Based Initialisation**

   We first regress the response variable on the predictors using a multiple linear regression. Then, we construct a single index (i.e. $p = 1$) using the estimated regression coefficients as the initial index coefficients of the predictors.

4. **Multiple: Picking One From Multiple Initialisations**

   The final optimised SMI model may change depending on the initialisation provided to the algorithm. Hence, we also consider using several different models as initialisations, optimise the SMI model for each of them, and pick the initial model that results in the lowest loss for the MIQP problem.

Of course, it is also possible for a user to specify an initialisation, based on their own domain expertise or prior knowledge in initialising the algorithm.

In each of the above initialisation options, once the estimate for $\boldsymbol{\alpha}$ is obtained, the estimated initial index coefficients for each index are scaled to have unit norm to ensure identifiability.

Estimating Nonlinear Functions

Once we have an estimate for $\boldsymbol{\alpha}$, estimating the SMI model is equivalent to estimating a Generalized Additive Model (GAM) as

$$y_i = \beta_0 + \sum_{j=1}^{p} g_j(\hat{h}_{ij}) + \sum_{k=1}^{d} f_k(w_{ik}) + \boldsymbol{\theta}^T \boldsymbol{u}_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

where y_i is the response, and $\hat{h}_{ij} = \hat{\boldsymbol{\alpha}}_j^T \boldsymbol{x}_i$ is the estimated index. The R packages ***mgcv*** (Wood 2011) and ***gam*** (Hastie 2023), for example, can be used to fit GAMs.

Updating the Index Structure and Index Coefficients

We obtain the updated index coefficients $\boldsymbol{\alpha}^{\text{new}}$ through a MIQP:

$$\begin{aligned}
&\min_{\boldsymbol{\alpha}^{\text{new}}, \boldsymbol{z}^{\text{new}}} (\boldsymbol{\alpha}^{\text{new}} - \boldsymbol{\alpha}^{\text{old}})^T \boldsymbol{V}^T \boldsymbol{V} (\boldsymbol{\alpha}^{\text{new}} - \boldsymbol{\alpha}^{\text{old}}) - 2(\boldsymbol{\alpha}^{\text{new}} - \boldsymbol{\alpha}^{\text{old}})^T \boldsymbol{V}^T \boldsymbol{r} + \lambda_0 \sum_{j=1}^{p} \sum_{m=1}^{q} z_{jm}^{\text{new}} + \lambda_2 \sum_{j=1}^{p} \sum_{m=1}^{q} (\alpha_{jm}^{new})^2 \\
&\text{s.t. } |\alpha_{jm}^{\text{new}}| \leq M z_{jm}^{\text{new}} \quad \forall j, \forall m, \\
&\quad z_{jm}^{\text{new}} \in \{0, 1\}, \\
&\quad \sum_{j=1}^{p} z_{jm}^{\text{new}} \leq 1 \quad \forall m,
\end{aligned} \tag{6}$$

where $j = 1, \ldots, p$, $m = 1, \ldots, q$, $\boldsymbol{\alpha}^{\text{old}}$ is the current value of $\boldsymbol{\alpha}$, z_{jm}^{new} are the updated set of binary variables to be estimated, and $\boldsymbol{V}$ is the matrix of partial derivatives of the right hand side of Equation 3 with respect to $\boldsymbol{\alpha}_j$. The i^{th} line of $\boldsymbol{V}$ contains $[\boldsymbol{v}_{i1}, \ldots, \boldsymbol{v}_{ip}]$, where $\boldsymbol{v}_{ij} = \boldsymbol{x}_i g_j'(h_{ij})$. The current residual vector, which contains $r_i = y_i - \beta_0 - \sum_{j=1}^{p} g_j\big((\boldsymbol{\alpha}_j^{\text{old}})^T \boldsymbol{x}_i\big)$, is denoted by $\boldsymbol{r}$. It is important to note that the additional covariates w_{ik} and $\boldsymbol{u}_i$ are not required to update $\boldsymbol{\alpha}_j$, because they are constants with respect to $\boldsymbol{\alpha}_j$, and thus they disappear from $\boldsymbol{V}$.

Similar to the explanation given by Masselot et al. (2023), the MIQP objective function in Equation 6 ignores the Hessian (the matrix of second derivatives of Equation 3, with respect to $\boldsymbol{\alpha}_j$), and considers only the matrix of first derivatives, which is a quasi-Newton step (Peng 2022). Therefore, the $\boldsymbol{\alpha}$ updating step given in Equation 6 is assured to be in a descent direction.

After obtaining $\boldsymbol{\alpha}^{\text{new}}$, if any of the estimated individual index coefficient vectors $\boldsymbol{\alpha}_j^{\text{new}}$ contains all zeros, they will be dropped from the model. Then we scale each estimated index coefficient vector $\hat{\boldsymbol{\alpha}}_j = \boldsymbol{\alpha}_j^{\text{new}}$ to have a unit norm.

Next, the algorithm alternates between updating the index coefficients $\boldsymbol{\alpha}$ and estimating the nonlinear functions $\boldsymbol{g}$ until it meets one of the three criteria: (i) the reduction ratio of the objective (loss)

function value in Equation 5, calculated between consecutive iterations, reaches a pre-specified convergence tolerance; (ii) the loss increases consecutively for three iterations; or (iii) the maximum number of iterations is reached. The selection of convergence tolerance and maximum iterations depends on the specific problem or data.

Finally, we consider adjusting the index structure of the model to exploit any potential benefits in terms of further minimising the loss function in Equation 5. As indices can be automatically reduced by dropping zero indices in each optimisation iteration, this step focuses on potential index additions to the current model. Specifically, we consider adding a new index to the current model by identifying dropped predictors. If applicable, a new index is constructed with all these dropped predictors, and the alternating updating process from the previous step is repeated. This step continues until one of these termination criteria is met: (i) the number of indices reaches q, selecting the final model as output; (ii) loss increases after the increment, selecting the previous iteration model as the final SMI model; or (iii) the solution maintains the same number of indices as the previous iteration, and the absolute difference of index coefficients between two successive iterations is not larger than a pre-specified tolerance, choosing the model with the smaller loss as the final SMI model in this case.

To obtain an estimated model with the best possible forecasting accuracy, it is important to select appropriate values for the non-negative penalty parameters λ_0 and λ_2. One possible way to do this is to estimate the model over a grid of possible values for λ_0 and λ_2, and then select the combination that yields the lowest loss function value. Moreover, it is also crucial to choose a suitable value for the big-M parameter, as the strength of the MIP formulation depends on the choice of a good upper bound (Bertsimas, King & Mazumder 2016). According to Hazimeh, Mazumder & Radchenko (2023), several methods have been used to select M in practice. For more details on estimating M in a linear regression setting, refer to Bertsimas, King & Mazumder (2016).

The following algorithm summarises the key steps of the SMI Modelling algorithm.

Algorithm 1: SMI Modelling Algorithm

1. Initialise p, the predictor grouping among indices, and obtain the initial estimate of $\boldsymbol{\alpha}$ using one of the options in Section 2.2. Then scale each $\hat{\boldsymbol{\alpha}}_j$, $j = 1, \ldots, p$, to have a unit norm.
2. Estimate g_j, $j = 1, \ldots, p$, using a GAM taking y_i as the response and $\hat{\boldsymbol{\alpha}}_j^T \boldsymbol{x}_i$ as predictors.
3. Update $\boldsymbol{\alpha}$ through the MIQP in Equation 6, and scale each $\hat{\boldsymbol{\alpha}}_j$ to have a unit norm.
4. Iterate steps 2 and 3 until convergence, loss increase for three consecutive iterations, or reaching the maximum iterations.

5. If there are no dropped predictors, stop. Otherwise include a new index consisting of dropped predictors, and repeat step 4.
6. Increase p by one in each iteration of step 5 until meeting one of the termination criteria below.
    - The number of indices in the iteration reaches q; select the final fitted model as output.
    - Loss increases after the increment; select previous iteration model as the final SMI model.
    - The solution maintains the same number of indices as the previous iteration, and the absolute difference of index coefficients between two successive iterations is not larger than a pre-specified tolerance; select the model with smaller loss as the final SMI model.

Throughout the experiments in the paper, we use $M = 10$, a convergence tolerance of 0.001, and a maximum of 50 iterations in step 4 of Algorithm 1, and a convergence tolerance of 0.001 for coefficients in step 6 of Algorithm 1.

3 Simulation Experiment

This section presents the results of a simulation experiment designed to demonstrate the performance and characteristics of the proposed SMI Modelling algorithm. In particular, we investigate how the estimated SMI model varies depending on the initialisation used.

3.1 Data Generation

Generating predictor variables:

First, we generate two time series, each of length 1205: x_0 from a uniform distribution on the interval $[0, 1]$, and z_0 from a normal distribution $N(5, 4)$. Next, we construct lagged series of both x_0 and z_0, where x_i denotes the i^{th} lag of x_0, and z_i denotes the i^{th} lag of z_0. Then $\boldsymbol{x} = \{x_0, x_1, \ldots, x_5\}$, and $\boldsymbol{z} = \{z_0, z_1, \ldots, z_5\}$) are taken as predictors in the simulation experiment.

Generating response variables:

We generate two response variables y_1 and y_2, using two different index structures and a normally distributed white noise component with two different variances as follows:

- Low noise level - $N(\mu = 0, \sigma^2 = 0.01)$:
  $y_1 = (0.9 * x_0 + 0.6 * x_1 + 0.45 * x_3)^3 + \epsilon, \quad \epsilon \sim N(0, 0.01)$
  $y_2 = (0.9 * x_0 + 0.6 * x_1 + 0.45 * x_3)^3 + (0.35 * x_2 + 0.7 * x_5)^2 + \epsilon, \quad \epsilon \sim N(0, 0.01)$

- High noise level - $N(\mu = 0, \sigma^2 = 0.25)$:
  $y_1 = (0.9 * x_0 + 0.6 * x_1 + 0.45 * x_3)^3 + \epsilon, \quad \epsilon \sim N(0, 0.25)$
  $y_2 = (0.9 * x_0 + 0.6 * x_1 + 0.45 * x_3)^3 + (0.35 * x_2 + 0.7 * x_5)^2 + \epsilon, \quad \epsilon \sim N(0, 0.25)$

Hence, the response y_1 is constructed using a single index consisting of the predictor variables x_0, x_1, and x_3, whereas the other response y_2 is constructed using two indices, where the first index consists of the predictors x_0, x_1, and x_3, and the second index consists of x_2 and x_5. Neither the variable x_4 nor any of the z variables are used in generating y_1 and y_2.

Once the data set is generated, the first five observations are discarded due to the missing values introduced by lagged variables, leaving a data set with 1200 observations. We use the first 1000 observations as the training set, while the remaining 200 observations are kept aside as the test set for evaluating the estimated models.

3.2 Experiment Setup

We estimate SMI models through the proposed algorithm for each of the two response variables, using three different sets of predictors as inputs. Our aim is to assess the algorithm's capability to correctly pick the relevant predictor variables (and drop the irrelevant predictors), and to estimate the index structure of the true model.

The three different sets of predictors considered are as follows:

1. All $\boldsymbol{x}$ variables (denoted as "all $\boldsymbol{x}$");
2. All $\boldsymbol{x}$ variables and all $\boldsymbol{z}$ variables (denoted as "all $\boldsymbol{x}$ + all $\boldsymbol{z}$");
3. The first three $\boldsymbol{x}$ variables (i.e. x_0, x_1 and x_2) and all $\boldsymbol{z}$ variables (denoted as "some $\boldsymbol{x}$ + all $\boldsymbol{z}$").

We apply the proposed SMI Modelling algorithm with each of the above predictor combinations, for both variations of the responses concerning the noise level. Moreover, we consider every initialisation options discussed in Section 2.2, for each of the two responses.

3.3 Results

We summarise the results of the simulation experiment in Table 1. In the columns, we indicate the index structure (i.e. the predictor grouping among indices) estimated by the proposed algorithm under each of the initialisation options, i.e., "PPR", "Additive", "Linear", and "Multiple". This is detailed for each combination explored, considering response, input predictors, and noise levels.

In the simulation experiment, we did not perform any tuning for the penalty parameters λ_0 and λ_2. Our experiments indicated that, for this simple example, different values of penalty parameters have

a negligible impact on the estimated models. The default values $\lambda_0 = 1$ and $\lambda_2 = 1$ were used in estimating all the models presented in Table 1.

Table 1: *Simulation experiment results.*

True Model	Predictors	PPR	Additive	Linear	Multiple
Low noise level					
y_1	all $\boldsymbol{x}$	(x_0, x_1, x_3)	(x_0, x_1, x_3)	(x_0, x_1, x_3)	(x_0, x_1, x_3)
y_1	all $\boldsymbol{x}$ + all $\boldsymbol{z}$	(x_0, x_1, x_3)	(x_0, x_1, x_3)	(x_0, x_1, x_3)	(x_0, x_1, x_3)
y_1	some $\boldsymbol{x}$ + all $\boldsymbol{z}$	(x_0, x_1, z_2, z_4)	(x_0, x_1) (z_4) (z_1)	(x_0, x_1, z_2, z_4)	(x_0, x_1, z_2, z_4)
y_2	all $\boldsymbol{x}$	(x_0, x_1, x_3) (x_2, x_5)	(x_0, x_1, x_3) (x_2, x_5)	$(x_0, x_1, x_2, x_3, x_5)$	(x_0, x_1, x_3) (x_2, x_5)
y_2	all $\boldsymbol{x}$ + all $\boldsymbol{z}$	(x_0, x_1, x_3) (x_2, x_5)	(x_0, x_1, x_3) (x_2, x_5)	$(x_0, x_1, x_2, x_3, x_5)$	(x_0, x_1, x_3) (x_2, x_5)
y_2	some $\boldsymbol{x}$ + all $\boldsymbol{z}$	(x_0, x_1) (x_2) (z_4)	(x_0, x_1, z_4) (x_2)	(x_0, x_1, x_2, z_2)	(x_0, x_1) (x_2, z_2, z_3)
High noise level					
y_1	all $\boldsymbol{x}$	(x_0, x_1, x_3) (x_2, x_4, x_5)	(x_0, x_1) (x_3)	(x_0, x_1, x_3)	(x_0, x_1, x_3)
y_1	all $\boldsymbol{x}$ + all $\boldsymbol{z}$	(x_0, x_1, x_3)	(x_0, x_1, x_3) (z_0)	(x_0, x_1, x_3)	(x_0, x_1, x_3) (z_0)
y_1	some $\boldsymbol{x}$ + all $\boldsymbol{z}$	(x_0, x_1) (z_1) (z_4)	(x_0, x_1) (z_1) (z_4)	(x_0, x_1, z_2, z_4)	(x_0, x_1) (z_0, z_4) (z_1)
y_2	all $\boldsymbol{x}$	(x_0, x_1, x_3) (x_2, x_5) (x_4)	(x_0, x_1, x_3) (x_2, x_5) (x_4)	$(x_0, x_1, x_2, x_3, x_5)$ (x_4)	(x_0, x_1, x_3) (x_2, x_5) (x_4)
y_2	all $\boldsymbol{x}$ + all $\boldsymbol{z}$	(x_0, x_1, x_3) (x_5, z_1) (x_2, z_0)	(x_0, x_1, x_3) (x_2, x_5, z_1)	$(x_0, x_1, x_2, x_3, x_5, z_0)$	(x_0, x_1, x_3) (x_2, x_5)
y_2	some $\boldsymbol{x}$ + all $\boldsymbol{z}$	$(x_0, x_1, z_0, z_3, z_4)$ (x_2)	$(x_0, x_1, z_0, z_1, z_3, z_4)$ (x_2)	$(x_0, x_1, x_2, z_0, z_3, z_4)$	$(x_0, x_1, z_0, z_1, z_3, z_4)$ (x_2)

At a low noise level, in both cases of "all $\boldsymbol{x}$" and "all $\boldsymbol{x}$ + all $\boldsymbol{z}$", all initialisations enable the algorithm to estimate the correct index structure for both y_1 and y_2, with an exception in the "Linear" option for y_2. The "Linear" option for y_2 selects the correct variables, but fails to identify the two-index structure. This suggests that initialising the algorithm with a higher number of indices might be more effective than with a lower number. In the case of "some $\boldsymbol{x}$ + all $\boldsymbol{z}$", for both y_1 and y_2, the models estimated under all initialisations include some noise variables. This indicates that when the available predictors are insufficient to capture the data signal, the algorithm might select irrelevant variables to make up for the missing signal.

When the fitted models are evaluated for y_1, in both cases of "all $\boldsymbol{x}$" and "all $\boldsymbol{x}$ + all $\boldsymbol{z}$", all initialisations result in a test set Mean Squared Error (MSE) of ≈ 0.01, which corresponds to the variance of the true model error. This confirms the accuracy with which the SMI Modelling algorithm estimates the index structure for y_1. For y_2, all the estimated models result in a test set MSE of ≈ 0.16. This is an interesting result as the test set MSE of an estimated model with incorrect index structure, but with correct predictors ("Linear") is similar to the models with correct index structure ("PPR", "Additive", and "Multiple"). This suggests that the selection of the predictor variables is more important than determining the index structure of the model in this case. For both y_1 and y_2, in the case of "some $\boldsymbol{x}$ + all $\boldsymbol{z}$", MSE on the test set increases in comparison to the above cases, probably due to the inclusion of the noise variables.

Moreover, in contrast to y_1, the test set MSE values for y_2 are higher than the variance of the true model error. Intuitively, the complexity of the model y_2 is higher than y_1, where the total estimation error of two nonlinear functions (corresponding to the two indices) for y_2, might be higher than the error of estimating a single nonlinear function for y_1.

As expected, the accuracy with which the SMI Modelling algorithm estimates the index structure is in general lower with the high noise level than with the low noise level. For both y_1 and y_2, most of the estimated models have selected irrelevant variables. In both the cases of "all $\boldsymbol{x}$" and "all $\boldsymbol{x}$ + all $\boldsymbol{z}$", all the models estimated for y_1 (except for the "Additive" option in the "all $\boldsymbol{x}$" case) result in a test set MSE of ≈ 0.23 (which is slightly lower than the variance of the true model error) irrespective of the fact that in the "all $\boldsymbol{x}$" case, "PPR" option, and in the "all $\boldsymbol{x}$ + all $\boldsymbol{z}$" case, "Additive" and "Multiple" options include irrelevant variables. This suggests over-fitting when there is low signal-to-noise ratio in the data. The result is the same for y_2, where irrespective of the different index structures and predictor choices, the estimated models in the above two predictor combinations produce similar MSE values on the test set.

Similar to the previous case of low noise level, when only a part of $\boldsymbol{x}$ variables are provided, the test set MSE values increase for both y_1 and y_2, where the estimated models for y_2 produce higher test MSE values in comparison to the models for y_1.

It is worth mentioning here that in real-world forecasting problems, the true data generating process (DGP) is unknown, and we do not expect an estimated model to precisely capture the true DGP. Therefore, as long as the estimated model demonstrates good forecasting accuracy, the index structure of the estimated model is less important.

Finally, the simulation study indicates that the choice of the initialisation depends on the data and application. Thus, users are encouraged to follow a trial-and-error procedure to determine the most suitable initial model for a given application.

4 Empirical Applications

4.1 Forecasting Daily Mortality

We apply the SMI Modelling algorithm to a data set used in Masselot et al. (2023), to forecast daily mortality based on heat exposure. Studying the effects of various environmental exposures such as weather related variables, pollutants and man-made environmental conditions on human health, is of significant importance in environmental epidemiology.

Description of the Data

For this analysis, we consider daily mortality and heat exposure data for the Metropolitan Area of Montreal, Québec, Canada, from 1990 to 2014, for the months June, July, and August (i.e. summer). The daily all-cause mortality data were obtained from the National Institute of Public Health, Québec,

while *DayMet*, a $1km \times 1km$ grid data set (Thornton et al. 2021), was used to extract daily temperature and humidity data.

Figure 1 shows the time plots of daily deaths during the summer for the years from 1990 to 1993. The series for each of the four years are presented separately in a faceted grid for visual clarity.

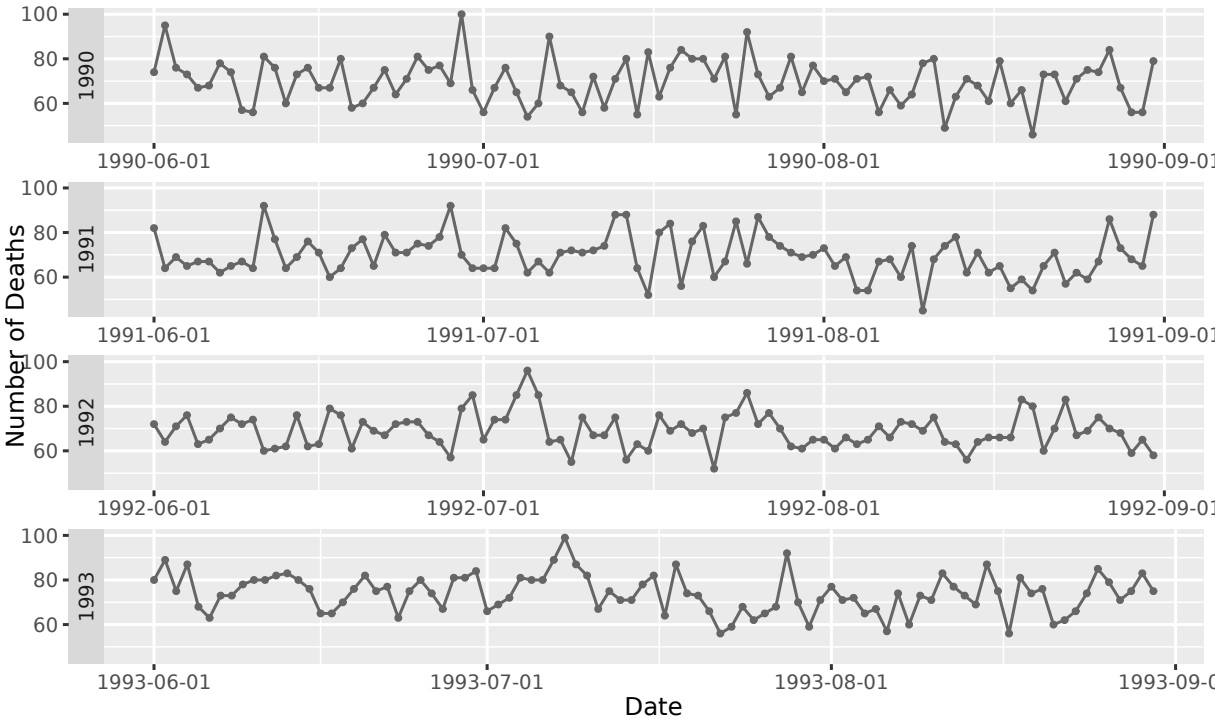

Figure 1: *Daily mortality in summer in Montreal, Canada from 1990 to 1993.*

Maximum temperature, minimum temperature, and vapour pressure (to represent the level of humidity) are considered as predictors in this empirical study. The number of daily deaths is plotted against each of these predictors in Figures 2–4, respectively, where we can observe that the relationships between these predictors and the response are slightly nonlinear.

Predictors Considered

1) Daily deaths lags: Fourteen lags of the daily deaths itself are used as predictors to incorporate the serial correlations presented in the data into the modelling process.

2) Current maximum/minimum temperatures and lags: In addition to current maximum and minimum temperatures, the temperature measurements up to 14 days prior are considered as predictors in the forecasting model. This accounts for the cumulative impact of both current and recent past temperatures on a person's heat exposure.

3) Current vapour pressure and lags: Similarly, the current value and 14 lags of vapour pressure are considered as predictors, as a proxy for the level of humidity.

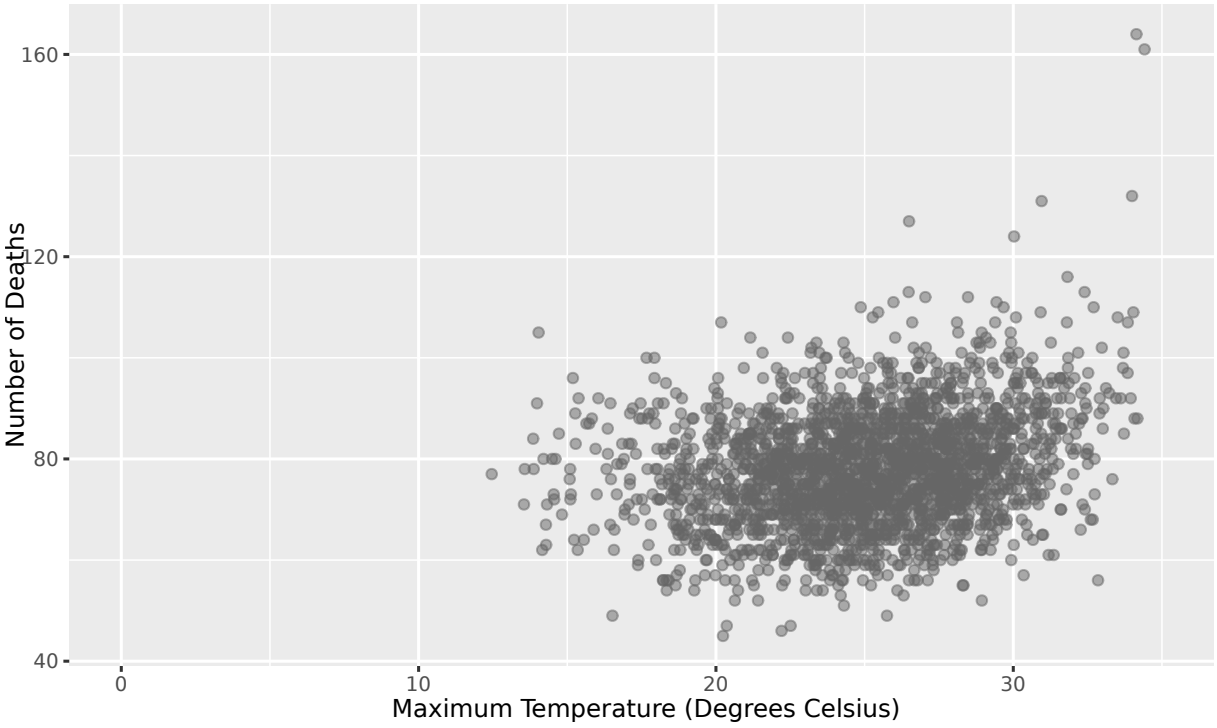

Figure 2: *Daily mortality in summer (from 1990 to 2014) plotted against maximum temperature.*

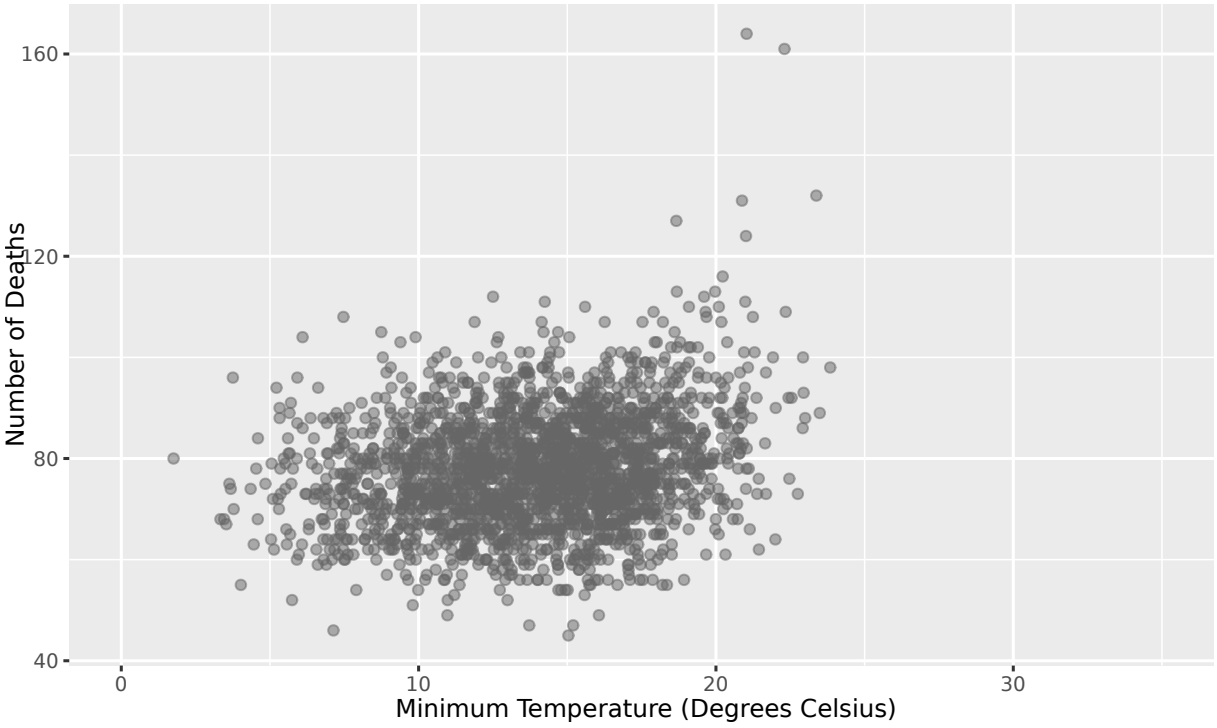

Figure 3: *Daily mortality in summer (from 1990 to 2014) plotted against minimum temperature.*

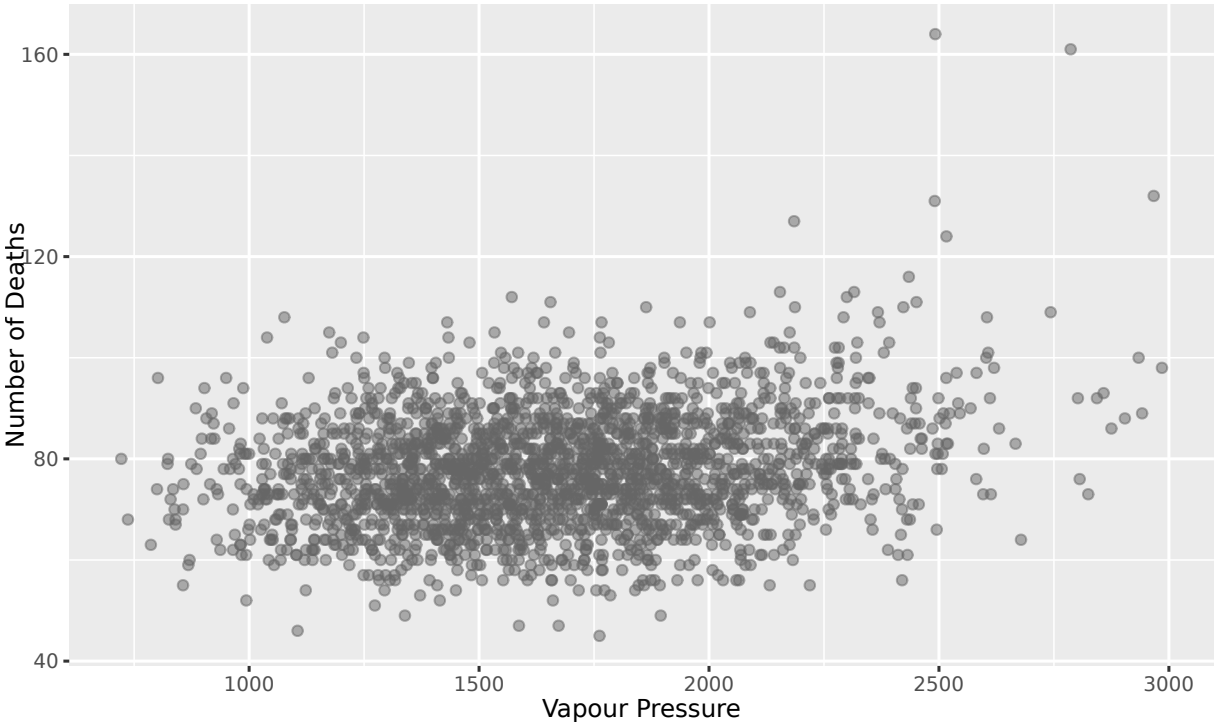

Figure 4: *Daily mortality in summer (from 1990 to 2014) plotted against vapour pressure.*

4) Calendar effects: Finally, a couple of calendar variables (*day of the season (DOS)* and *Year*) are incorporated into the model to capture annual trend and seasonality, which is a common practice in environmental epidemiology.

Modelling Framework

Death lags, maximum temperature lags, minimum temperature lags, and vapour pressure lags are predictors that may enter indices. The two calendar variables, *DOS* and *Year*, are included in the model as separate nonparametric components that do not enter any of the indices. Hence, the relevant SMI model can be written as

$$\textbf{Deaths} = \beta_0 + \sum_{j=1}^{p} g_j(\boldsymbol{X}\boldsymbol{\alpha}_j) + f_1(\textbf{DOS}) + f_2(\textbf{Year}) + \boldsymbol{\varepsilon}, \tag{7}$$

where

- **Deaths** is the vector containing daily deaths observations;
- β_0 is the model intercept;
- p is the unknown number of indices to be estimated via the proposed algorithm;
- $\boldsymbol{X}$ is a matrix containing the $q = 59$ predictor variables that may enter indices (i.e. death lags, maximum temperature lags, minimum temperature lags, and vapour pressure lags);
- $\boldsymbol{\alpha}_j$, $j = 1, \ldots, p$ are the index coefficient vectors, each of length q;

- $g_1, \ldots, g_p$, f_1, and f_2 are unknown nonparametric functions; and
- $\boldsymbol{\varepsilon}$ is the error term.

The data from 1990 to 2012 are used as the training set to estimate the model, while the data from the three summer months of year 2013 are kept aside as a validation set. This validation set is used to perform the penalty parameter tuning of the SMI model. Here, once the tuning for the two penalty parameters is completed, the model is re-fitted for the entire data set combining training and validation sets. The data of year 2014 are used as the test set for evaluating forecasting performance.

Have the model refitted using the data from 1990 to 2013?

Comment addressed

The forecasting accuracy on the test set is evaluated using MSE and Mean Absolute Error (MAE). We assume that the future values of the maximum/minimum temperatures and vapour pressure are known to use in the forecasting model; thus it is a post hoc analysis.

Benchmark Methods

In addition to our proposed SMI model, we also evaluate three benchmark models for comparison purposes.

The first benchmark is a nonparametric additive model formulated through ***Backward Elimination***, as proposed by Fan & Hyndman (2012). The process starts with all variables included in an additive model, and variables are progressively omitted until the optimal model is obtained based on the validation set. Once the optimal model is obtained, the final model is re-fitted for the entire data set combining training and validation sets.

Is this the only method that has used the validation set?

Comment addressed

The second benchmark is a ***Group-wise Additive Index Model (GAIM)***, which can be written in the form

$$y_i = \sum_{j=1}^{p} g_j(\boldsymbol{\alpha}_j^T \boldsymbol{x}_{ij}) + \varepsilon_i,$$

where y_i is the univariate response, $\boldsymbol{x}_{ij} \in \mathbb{R}^{lj}$, $j = 1, \ldots, p$ are pre-specified non-overlapping subsets of $\boldsymbol{x}_i$, and $\boldsymbol{\alpha}_j$ are the corresponding index coefficients, g_j is an unknown (possibly nonlinear) component function, and ε_i is the random error, which is independent of $\boldsymbol{x}_i$ (Wang et al. 2015; Masselot et al. 2023). For this model, death lags, maximum temperature lags, minimum temperature lags, and

vapour pressure lags are intuitively categorised into four separate groups, with an index estimated for each group.

The third benchmark is a ***Projection Pursuit Regression (PPR)*** model (Friedman & Stuetzle 1981) given by

$$y_i = \sum_{j=1}^{p} g_j(\boldsymbol{\alpha}_j^T \boldsymbol{x}_i) + \varepsilon_i,$$

where y_i is the response, $\boldsymbol{x}_i$ is the q-dimensional predictor vector, $\boldsymbol{\alpha}_j = (\alpha_{j1}, \ldots, \alpha_{jp})^T$, $j = 1, \ldots, p$ are q-dimensional projection vectors (or vectors of "index coefficients"), g_j's are unknown nonlinear functions, and ε_i is the random error. The number of indices in the PPR model was taken as four, matching the number of indices estimated by the GAIM.

Results

We estimated SMI models for the mortality data using three different initialisation options: "PPR", "Additive" and "Linear", for comparison purposes. We tuned the penalty parameters λ_0 and λ_2 over ranges of integers from 1 to 12, and 0 to 12 respectively, based on validation set MSE. Here, a greedy search is used instead of a grid search to reduce computational time.

The penalty parameter combination ($\lambda_0 = 5, \lambda_2 = 12$) was selected for the model fitted with "PPR" initialisation. The estimated model, ***SMI Model (5, 12) - PPR***, resulted in seven indices without dropping any of the index variables. The optimal penalty parameter combination for the model initiated with "Additive" was ($\lambda_0 = 1, \lambda_2 = 0$), resulting in the ***SMI Model (1, 0) - Additive***, equivalent to a nonparametric additive model (no index variables or indices were dropped). The model estimated with "Linear" initialisation selected ($\lambda_0 = 6, \lambda_2 = 11$) (***SMI Model (6, 11) - Linear***), and resulted in two indices, without dropping any of the index variables.

We evaluated forecasting errors of the estimated models using two subsets of the original test set:

1. ***Test Set 1:*** original test set spanning three months (June, July and August 2014); and
2. ***Test Set 2:*** a test set covering one month (June 2014).

The MSE and MAE values on the two different test sets for the estimated SMI models along with all benchmark methods considered are presented in Table 2. We observe that the SMI model estimated with "PPR" initialisation, ***SMI Model (5, 12) - PPR***, shows the best forecasting performance on both test sets, compared to the other two estimated SMI models.

Table 2 also shows that ***SMI Model (5, 12) - PPR*** outperforms all three benchmark models in terms of forecasting accuracy, for both *Test Set 1* and *Test Set 2*. However, the SMI Models estimated using

"Additive" or "Linear" initialisations have inferior forecasting performance compared to GAIM and PPR models. The actual number of deaths and the predicted values from the ***SMI Model (5, 12) - PPR*** and benchmark models on *Test Set 2* are plotted in Figure 5 for further comparison.

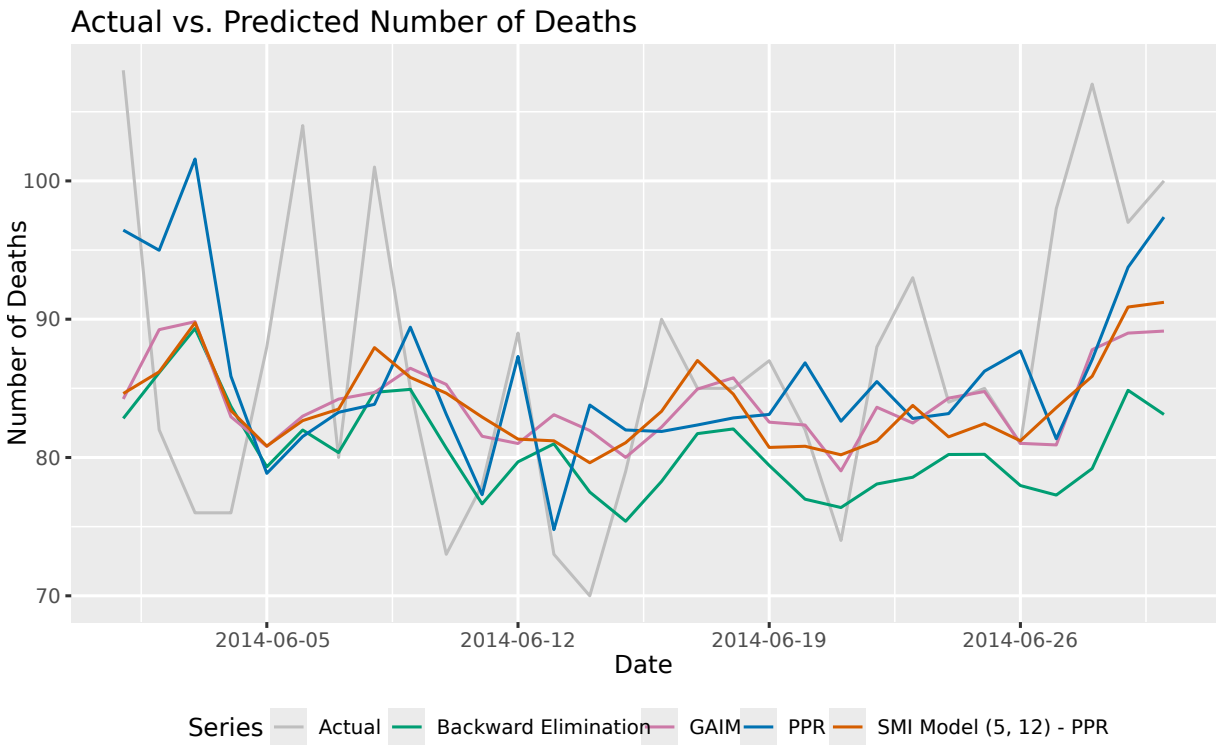

Figure 5: *Actual number of deaths vs. predicted number of deaths from "SMI Model (12, 0) - PPR" and benchmark models for Test Set 2.*

4.2 Forecasting Daily Solar Intensity

Next, we utilise the SMI Modelling algorithm to forecast daily solar intensity, using other weather conditions.

Table 2: *Daily mortality forecasting - Out-of-sample point forecast results.*

			Test Set 1		Test Set 2	
Model	Predictors	Indices	MSE	MAE	MSE	MAE
SMI Model (5, 12) - PPR	61	7	**85.233**	**7.140**	**97.353**	**7.772**
SMI Model (1, 0) - Additive	61	59	96.398	7.481	112.199	8.156
SMI Model (6, 11) - Linear	61	2	100.231	7.719	120.542	8.598
Backward Elimination	40		136.204	9.319	140.867	9.385
GAIM	61	4	90.763	7.247	106.251	7.928
PPR	61	4	90.698	7.343	110.497	8.057

Description of the Data

We use solar intensity and other weather variables measured at the Davis weather station in Amherst, Massachusetts, obtained from the *UMass Trace Repository* (University of Massachusetts 2023). The data was recorded at every five minutes, from 24 February 2006 to 27 February 2013, using sensors for measuring temperature, wind chill, humidity, dew point, wind speed, wind direction, rain, pressure, solar intensity, and UV. We converted the five minutes data to daily data by averaging each variable.

Figure 6 shows the time plot of daily solar intensity for the entire period, which clearly depicts the annual seasonality in the data. The gaps show days for which the observations were missing; these were excluded from the analysis.

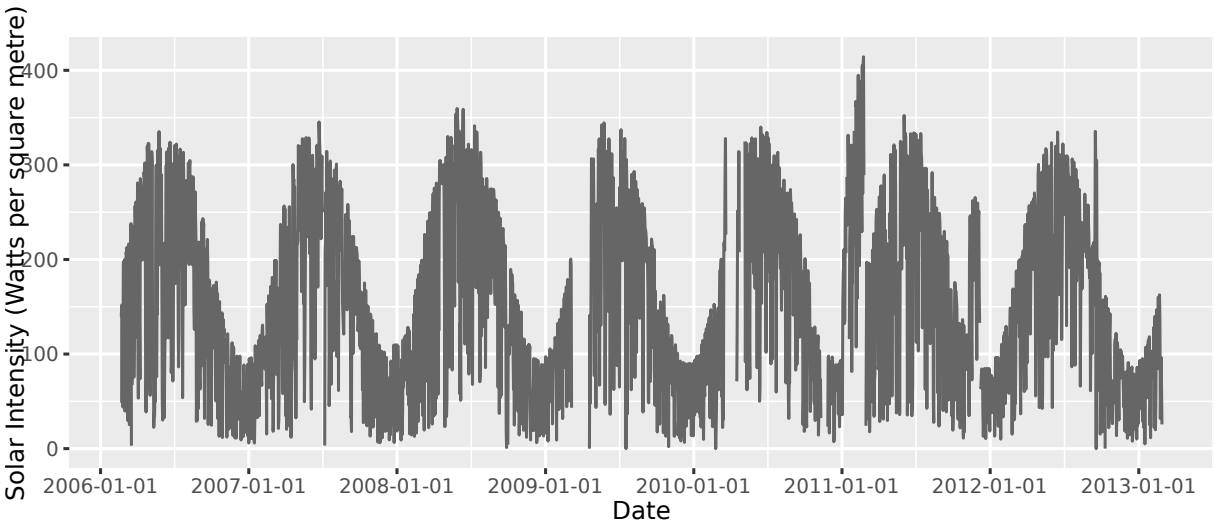

Figure 6: *Daily solar intensity in Amherst, Massachusetts, from February 2006 to February 2013.*

In Figure 7, the daily solar intensity is plotted against each of the predictors: temperature, dew point, wind, rain and humidity. We can observe that the relationships between these predictors and the response are nonlinear.

Predictors Considered

1) Solar intensity lags: Three lags of the daily solar intensity itself are used as predictors to incorporate the serial correlations presented in the data into the modelling process.

2) Current weather variables and lags: In addition to current temperature, dew point, wind speed, rain and humidity, the measurements of the three previous days for each of the these weather variables are also included as predictors in the forecasting model.

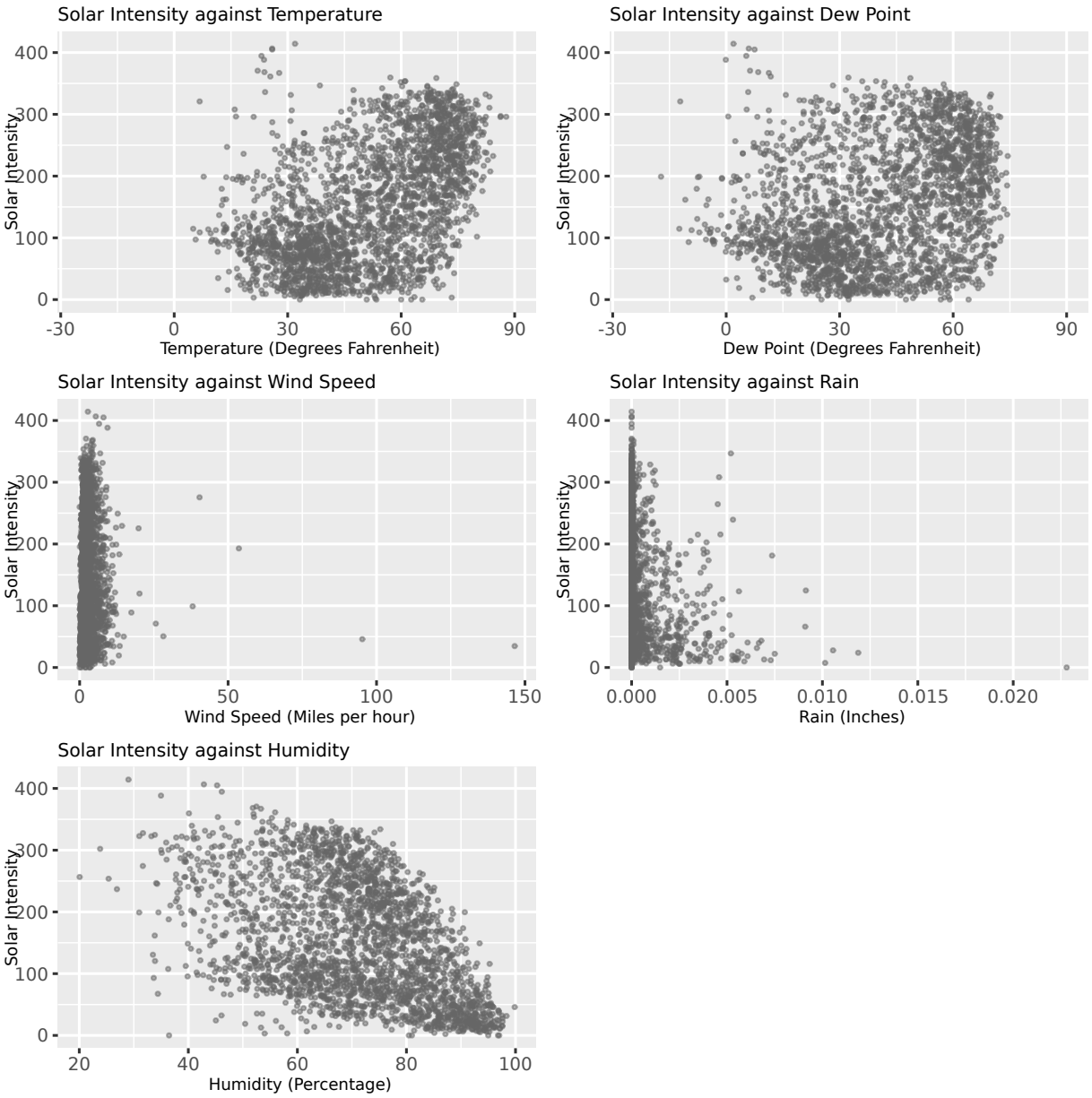

Figure 7: *Daily solar intensity against other weather variables.*

3) Calendar effects: Finally, a calendar variable, *day of the year (DOY)* is incorporated into the model as a smooth function through Fourier terms to capture annual seasonality, and also to control the autocorrelation in residuals.

Modelling Framework

The lags of solar intensity, and the lags of weather variables are considered as predictors that may enter indices. The Fourier terms capturing the day of the year effect are included into the model as linear variables (so that they are not part of the indices).

We experimented with the number of pairs of Fourier terms (K) that best captures the day of the year effect, considering values from $K = 1$ to $K = 10$. The SMI models fitted with one pair of Fourier

terms ($K = 1$) resulted in the lowest test MSE. (i.e. the day of the year seasonal pattern follows a simple sine wave.)

Did you use the test set data when choosing K?

Hence, the relevant SMI model can be written as

$$\textbf{Solar} = \beta_0 + \sum_{j=1}^{p} g_j(\boldsymbol{X}\boldsymbol{\alpha}_j) + \theta_1 \textbf{DOY\_S1} + \theta_2 \textbf{DOY\_C1} + \boldsymbol{\varepsilon}, \tag{8}$$

where

- **Solar** is the vector containing daily observations of solar intensity;
- β_0 is the model intercept;
- p is the unknown number of indices to be estimated via the algorithm;
- $\boldsymbol{X}$ is a matrix containing the $q = 23$ predictor variables that may enter indices (i.e. solar intensity, temperature, dew point, wind speed, rain and humidity lags);
- $\boldsymbol{\alpha}_j$, $j = 1, \ldots, p$ are the index coefficient vectors, each of length q;
- $g_1, \ldots, g_p$ are unknown nonparametric functions;
- **DOY_S1** and **DOY_C1** are the sine and cosine terms of the first pair of Fourier terms corresponding to the variable **DOY** respectively (seasonal period = 365.25);
- θ_1 and θ_2 are the coefficients of **DOY_S1** and **DOY_C1** respectively; and
- $\boldsymbol{\varepsilon}$ is the error term.

The data from February 2006 to October 2012 are used as the training set to estimate the model, while the data of January and February 2013 comprise the test set to evaluate the forecasting performance. The data from November and December 2012 are kept aside as a validation set, which is required to estimate some of the benchmark models for comparison.

Same question as in application 1.

Then we apply the proposed SMI Modelling algorithm to the training set to estimate the model, and the forecasting accuracy on the test set is evaluated using MSE and MAE. We assumed that the future values of the weather variables are known; thus it is a post hoc analysis.

Results

We estimated SMI Models for the solar intensity data using three different initialisation options: "PPR", "Additive" and "Linear". We also tuned the penalty parameters λ_0 and λ_2, over ranges of integers from 1 to 12, and 0 to 12 respectively.

Table 3: *Daily solar intensity forecasting - Out-of-sample point forecast results.*

Model	Predictors	Indices	Test Set	
			MSE	MAE
SMI Model (1, 12) - PPR	25	5	1094.097	26.186
SMI Model (1, 0) - Additive	25	23	784.448	21.698
SMI Model (1, 0) - Linear	2	0	1556.214	30.893
Backward Elimination	17		**564.811**	**19.604**
GAIM	25	6	1664.801	31.787
PPR	23	6	796.779	22.455

The penalty parameter combination ($\lambda_0 = 1, \lambda_2 = 12$) was selected for the model fitted with "PPR" initialisation. The estimated model, ***SMI Model (1, 12) - PPR***, resulted in five indices without dropping any of the index variables. The optimal penalty parameter combination for the model estimated taking "Additive" model as the starting point was ($\lambda_0 = 1, \lambda_2 = 0$). The estimated SMI model did not drop any index variables or indices, and thus the final model, ***SMI Model (1, 0) - Additive***, is equivalent to a semi-parametric additive model. The model estimated with "Linear" initialisation also selected ($\lambda_0 = 1, \lambda_2 = 0$). Unlike the above models, this SMI Model dropped all index variables and resulted in null indices, and hence, the final model, ***SMI Model (1, 0) - Linear***, is just a linear model with the two linear variables DOY_S1 and DOY_C1. Notice that both ***SMI Model (1, 0) - Additive*** and ***SMI Model (1, 0) - Linear*** have $\lambda_2 = 0$, indicating that these two models have omitted the ℓ_2-penalty in the estimation process.

Table 3 presents the MSE and MAE values for the estimated SMI models on the test set. The results indicate that the SMI model estimated with "Additive" initialisation, ***SMI Model (1, 0) - Additive***, shows the best forecasting performance among the three estimated SMI Models.

We also present forecasting errors of the three benchmark models in Table 3, to compare with the estimated SMI models. Here, the GAIM is fitted by grouping the lags of each weather variable into distinct group, resulting in six indices. The number of indices of the PPR model was taken as six, matching the number of indices estimated by the GAIM. Note that here, the calendar variable *DOY* was excluded when estimating the PPR model.

Just to double check, did you also use Fourier terms to allow for seasonal patterns in all benchmark methods? - **Yes, used Fourier terms in GAIM and backward elimination, but not in PPR because it is impossible to include separate linear terms in stats::ppr().**

Table 3 shows that the forecasting errors of ***SMI Model (1, 0) - Additive*** is lower than the GAIM and the PPR model. However, the ***SMI Model (1, 0) - Additive*** is unable to outperform the semi-parametric

additive model with backward elimination, which has resulted in the best forecasting accuracy in this case.

Is it because Backward Elimination method has used the validation set to evaluate out-of-sample performance in the modelling process?

Here, it is worth considering the differences between the SMI model and the backward elimination method that shows superior forecasting performance. The method proposed by Fan & Hyndman (2012) formulates a semi-parametric additive model using a backward elimination of predictors, and does not perform any predictor grouping. In contrast, the SMI model takes a more general and objective approach to estimate an additive index model (which includes the semi-parametric additive model as a special case), where the number of indices as well as predictors within each index are automatically determined through the proposed algorithm. Thus, the SMI model faces a more challenging estimation task due to the limited prior information provided regarding the model structure.

The actual solar intensity and the predicted values from the ***SMI Model (1, 0) - Additive*** and benchmark models are plotted in Figure 8 for further comparison.

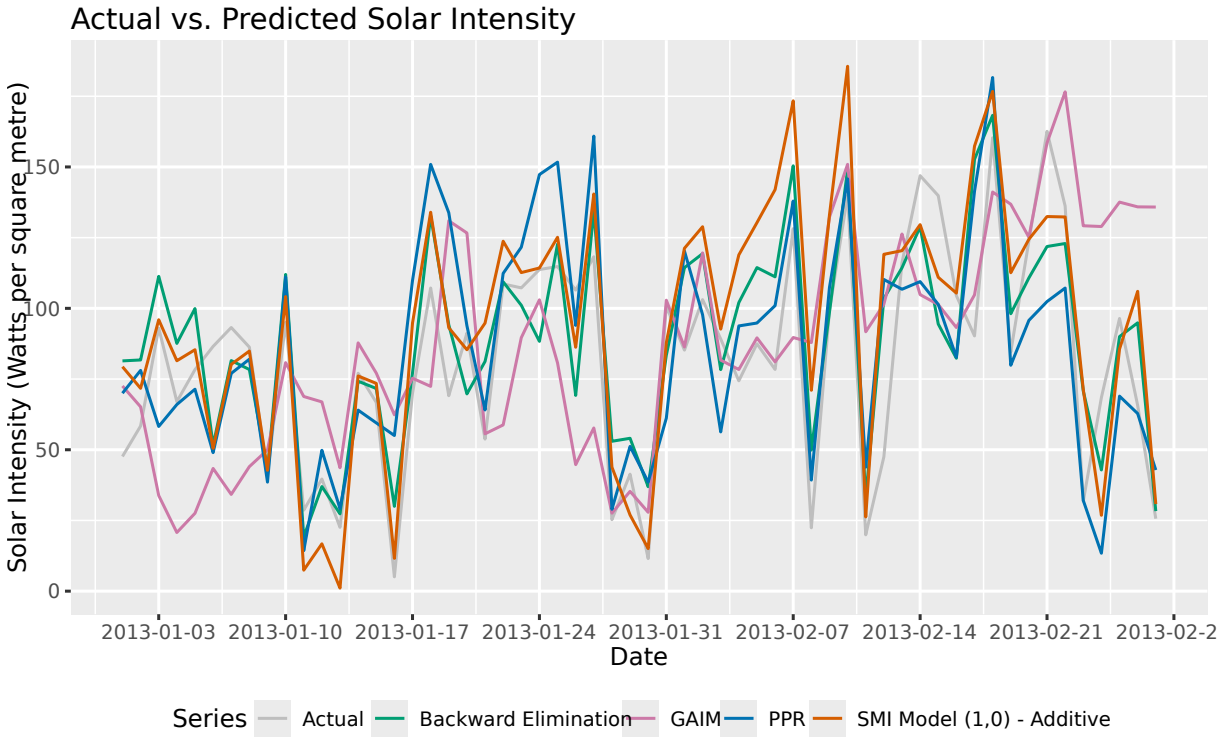

Figure 8: *Actual solar intensity vs. predicted solar intensity from "SMI Model (1, 0) - Additive" and benchmark models.*

4.3 Software

These two empirical applications were performed using the R statistical software (R Core Team 2024). We used the commercial MIP solver ***Gurobi*** (Gurobi Optimization, LLC 2023) to solve the MIQPs related to the proposed SMI Modelling algorithm, through the ***Gurobi plug-in*** (ROI.plugin.gurobi, Schwendinger 2023) available from the ***R Optimization Infrastructure*** (ROI, Hornik et al. 2023; Theußl, Schwendinger & Hornik 2020) package. Furthermore, the GAMs were fitted using the R package ***mgcv*** (Wood 2023; Wood 2011).

5 Conclusions and Further Research

In this paper, we presented a novel algorithm for estimating a nonparametric/semi-parametric additive index model with optimal predictor selection and predictor grouping, which we refer to as Sparse Multiple Index (SMI) Model. The SMI Modelling algorithm is an iterative procedure that is developed based on mixed integer programming to solve an ℓ_0-regularised nonlinear least squares optimisation problem with linear constraints.

The proposed SMI Modelling algorithm has a number of key features: 1) It performs automatic selection of both the number of indices and the predictor grouping when estimating the nonparametric additive index model. Users need to input the set of predictors entering indices and a starting model (index structure and a set of index coefficients) to initiate the algorithm. 2) It performs automatic variable selection, which is particularly beneficial in high-dimensional settings. This feature contributes to an objective and principled estimation, reducing subjectivity across different users. 3) It is capable of estimating a wide spectrum of models, from single index models (one index) to additive models (number of indices equals the number of predictors entering indices). Hence, the SMI Modelling algorithm is a more general estimation tool for nonparametric additive models. 4) It provides the flexibility to include separate nonlinear and linear predictors in the model that are not entering any indices, allowing the estimation of semi-parametric additive models.

We demonstrated the performance of the proposed algorithm through a simulation study and two empirical applications. Due to the limited input information provided to the algorithm, the estimation of a SMI Model is a challenging problem.

The two empirical applications presented above highlight the challenge of finding a universally applicable initialisation option for the SMI Model across various scenarios. As mentioned in Section 3, we encourage users to follow a trial-and-error procedure to identify the most effective initialisation option for their specific application.

Since the difficulty of specifying an initialisation that works in general is a limitation of the proposed algorithm, an interesting future research problem would be to explore the potential for determining a generalised initialisation for the SMI Modelling algorithm that will work well across various applications.

This study should be viewed as a first attempt to develop a more objective methodology for variable selection and model estimation in the broader class of nonparametric additive models for forecasting. An important future research problem is therefore, to assess the performance of the proposed SMI Modelling algorithm across various data sets with diverse properties, identifying scenarios where it outperforms other benchmark methods.

Furthermore, the MIQP in the algorithm is somewhat analogous to the *best subset selection* method frequently used in least squares problems. Thus, another limitation of the proposed algorithm is the increase in computational time as the number of predictors and number of indices increase. Therefore, it would be an interesting research to obtain further insights regarding the algorithm to see what improvements can be made to the algorithm design to reduce the computational cost in a high-dimensional context.

Acknowledgements

We thank Professor Louise Ryan for useful discussions during the initial stage of the project, and for her valuable comments and feedback on this research work.

Furthermore, this research is partially supported by the Monash eResearch Centre through the use of the MonARCH (Monash Advanced Research Computing Hybrid) HPC Cluster, and by the Australian Research Council Industrial Transformation Training Centre in Optimisation Technologies, Integrated Methodologies, and Applications (OPTIMA), Project ID IC200100009.

References

Bakker, M & F Schaars (2019). Solving groundwater flow problems with time series analysis: you may not even need another model. *Groundwater* **57**(6), 826–833.

Bellman, R (1957). *Dynamic Programming*. Princeton, New Jersey: Princeton University Press.

Bertsimas, D, A King & R Mazumder (2016). Best subset selection via a modern optimization lens. *Annals of Statistics* **44**(2), 813–852.

Fan, S & RJ Hyndman (2012). Short-term load forecasting based on a semi-parametric additive model. *IEEE Transactions on Power Systems* **27**(1), 134–141.

Friedman, JH & W Stuetzle (1981). Projection Pursuit Regression. *Journal of American Statistical Association* **76**(376), 817–823.

Gurobi Optimization, LLC (2023). *Gurobi Optimizer Reference Manual*. https://www.gurobi.com.

Härdle, W, P Hall & H Ichimura (1993). Optimal Smoothing in Single-Index Models. *Annals of Statistics* **21**(1), 157–178.

Hastie, T (2023). *gam: Generalized Additive Models*. R package version 1.22-2. https://CRAN.R-project.org/package=gam.

Hazimeh, H & R Mazumder (2020). Fast Best Subset Selection: Coordinate Descent and Local Combinatorial Optimization Algorithms. *Operations Research* **68**(5), 1517–1537.

Hazimeh, H, R Mazumder & P Radchenko (2023). Grouped variable selection with discrete optimization: Computational and statistical perspectives. *Annals of Statistics* **51**(1), 1–32.

Ho, CC, LJ Chen & JS Hwang (2020). Estimating ground-level PM2.5 levels in Taiwan using data from air quality monitoring stations and high coverage of microsensors. *Environmental Pollution* **264**, 114810.

Hornik, K, D Meyer, F Schwendinger & S Theussl (2023). *ROI: R Optimization Infrastructure*. R package version 1.0-1. https://CRAN.R-project.org/package=ROI.

Hyndman, RJ & S Fan (2010). Density forecasting for long-term peak electricity demand. *IEEE Transactions on Power Systems* **25**(2), 1142–1153.

Ibrahim, S, P Radchenko, E Ben-David & R Mazumder (2023). "Predicting Census Survey Response Rates With Parsimonious Additive Models and Structured Interactions". https://arxiv.org/abs/2108.11328.

Masselot, P, F Chebana, C Campagna, É Lavigne, TBMJ Ouarda & P Gosselin (2023). Constrained groupwise additive index models. *Biostatistics* **24**(4), 1066–1084.

Mazumder, R, P Radchenko & A Dedieu (2023). Subset selection with shrinkage: Sparse linear modeling when the SNR is low. *Operations Research* **71**(1), 129–147.

Peng, RD (2022). *Advanced Statistical Computing*. Accessed: 2023-5-19. https://bookdown.org/rdpeng/advstatcomp/.

Peterson, TJ & AW Western (2014). Nonlinear time-series modeling of unconfined groundwater head. *Water Resources Research* **50**(10), 8330–8355.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/.

Radchenko, P (2015). High dimensional single index models. *Journal of Multivariate Analysis* **139**, 266–282.

Rajaee, T, H Ebrahimi & V Nourani (2019). A review of the artificial intelligence methods in groundwater level modeling. *Journal of Hydrology* **572**, 336–351.

Ravindra, K, P Rattan, S Mor & AN Aggarwal (2019). Generalized additive models: Building evidence of air pollution, climate change and human health. *Environment International* **132**, 104987.

Schwendinger, F (2023). *ROI.plugin.gurobi: 'Gurobi' Plug-in for the 'R' Optimization Infrastructure*. R package version 0.4-0. http://r-forge.r-project.org/projects/roi.

Stoker, TM (1986). Consistent Estimation of Scaled Coefficients. *Econometrica* **54**(6), 1461–1481.

Stone, CJ (1982). Optimal global rates of convergence for nonparametric regression. *Annals of Statistics* **10**(4), 1040–1053.

Theußl, S, F Schwendinger & K Hornik (2020). ROI: An Extensible R Optimization Infrastructure. *Journal of Statistical Software* **94**, 1–64.

Thornton, PE, R Shrestha, M Thornton, SC Kao, Y Wei & BE Wilson (2021). Gridded daily weather data for North America with comprehensive uncertainty quantification. *Scientific Data* **8**(1), 190.

University of Massachusetts (2023). *The UMass Trace Repository*. Accessed on 19 March 2024. https://traces.cs.umass.edu/.

Wang, T, J Zhang, H Liang & L Zhu (2015). Estimation of a Groupwise Additive Multiple-Index Model and its Applications. *Statistica Sinica* **25**, 551–566.

Wood, SN (2017). *Generalized Additive Models: An Introduction with R*. 2nd. Chapman & Hall/CRC.

Wood, SN (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (Series B)* **73**(1), 3–36.

Wood, SN (2023). *mgcv: Mixed GAM Computation Vehicle with Automatic Smoothness Estimation*. R package version 1.9-1. https://CRAN.R-project.org/package=mgcv.