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

March 2024

Working Paper no/yr

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

18 March 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 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

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

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 iid 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 2 or 3 variables. There are also usually many 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 behind the use of this proposed semi-parametric model is to allow nonparametric components in a regression-based modelling framework with serially correlated errors. The model fitted for each half-hourly period (q) can be written as

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

where the response variable is the logarithm of electricity demand at time t during period q. The term $h_q(t)$ models several calendar effects as 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. It is important to notice here 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 an interaction of two predictor variables are captured by the sum of $\sum_{k=0}^{K} f_2(p_{t-k}, k)$ and $\sum_{k=0}^{K} f_3(o_{t-k}, w_{t-k}, k)$ respectively. The smooth functions f_2 and f_3 are proposed to be estimated using tensor product smooths.

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. (2022) for predicting census survey response rates. Ravindra et al. (2019) provides 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 such models have been used to address problems from diverse application areas, there are several unresolved issues in using them. 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 to include 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. This semiparametric 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.

This model subsumes the models discussed above, and includes fully additive models (Wood 2011, 2017), where each predictor is in its own index, and single index models (Stoker 1986; Härdle, Hall & Ichimura 1993; Radchenko 2015), where all predictors are 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 limit the order and form of interactions allowed. In our model formulation, the number of indices p is unknown, and the predictor grouping among indices is unknown. We propose algorithmic selection of the predictors to include in each indices, 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 was to develop a methodology for optimal predictor selection in the context of high-dimensional nonparametric additive index models. Moreover, due to computational advancements in the field, the use of mathematical optimisation concepts 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 in this paper 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 variable selection algorithm and estimation procedure. Some benchmark comparison methods are briefly introduced in Section 3. In Section 4, we demonstrate the functionality and the characteristics of the proposed algorithm through a simulation experiment. Section 5 illustrates two empirical applications of the proposed estimation and variable selection methodology, related to forecasting heat exposure related daily mortality and daily solar intensity. Concluding remarks are given in Section 6.

2 Sparse Multiple Index Model

2.1 Optimisation Problem Formulation

We implement variable selection for the proposed semi-parametric additive 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:

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

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 condition ensures that a 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 & Dedieuc 2022; 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 $\boldsymbol{z} = (\boldsymbol{z}_{11}, \ldots, \boldsymbol{z}_{pm})^T$, $j = 1, \ldots, p$, and $m = 1, \ldots, q$. In other words, we have introduced 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 $\max\left(|\alpha^*_{jm}|\right) \leq M$. The big-$M$ constraints ensure 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 $\lambda_0 \sum_{j=1}^{p} \sum_{m=1}^{q} z_{jm}$ influences some of the binary variables z_{jm} to be zero, while the ℓ_2-penalty term $\lambda_2 \sum_{j=1}^{p} \sum_{m=1}^{q} \alpha_{jm}^2$ enforces additional shrinkage on the estimated coefficients. Together, these components perform 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

In order to start solving the MIQP given in Equation 5, we first need to provide an 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 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 $\tau = 0.1 \times \max(\text{PPR coefficients})$.
   d. Set to zero all coefficients that fall below the calculated threshold.

e. For predictors appearing in multiple indices, assign them to the index with the maximum 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 index coefficients of the predictors.

The final optimised SMI Model may change depending on the initialisation provided to the algorithm. Hence, we also considered 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. In our simulations and applications, this is denoted as **Multiple**.

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 ($\hat{\boldsymbol{\alpha}}_j = \boldsymbol{\alpha}_{j,init}$) 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 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 $\boldsymbol{\alpha}^{new}$ through a MIQP:

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

where $\boldsymbol{\alpha}^{old}$ is the current value of $\boldsymbol{\alpha}$, and z_{jm}^{new} are the updated set of binary variables to be estimated. $\boldsymbol{V}$ 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 $r_i = y_i - \beta_0 - \sum_{j=1}^{p} g_j(\boldsymbol{\alpha}_j^{(old)T}\boldsymbol{x}_i)$, is denoted by $\boldsymbol{r}$. It is important to note that the additional covariates w_{ik} and $\boldsymbol{u}_i$ do not step in to the process of updating $\boldsymbol{\alpha}_j$, 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. (2022), the MIQP objective function in above Equation 6 ignores the Hessian (or 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. The quasi-Newton Method is an alternative to the Newton's Method, avoiding the calculation of the Hessian to circumvent its computational burden (Peng 2022). Therefore, the $\boldsymbol{\alpha}$ updating step given in above Equation 6 is assured to be in a descent direction.

When $\boldsymbol{\alpha}^{new}$ is obtained, if any of the estimated individual index coefficient vectors $\boldsymbol{\alpha}_j^{new}$ contains all $\boldsymbol{z}$eros (i.e. zero index), such indices will be dropped out from the model. Furthermore, similar to Section 2.2, once the new estimate $\boldsymbol{\alpha}^{new}$ is obtained, we scale each estimated index coefficient vector $\hat{\boldsymbol{\alpha}}_j = \boldsymbol{\alpha}_j^{new}$ to have unit norm.

The algorithm alternates updating the index coefficients $\boldsymbol{\alpha}$ and estimating nonlinear functions $\boldsymbol{g}$ with the updated $\boldsymbol{\alpha}$ until meeting 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. In the empirical applications in Section 5, we used a

convergence tolerance of 0.001 and a maximum of 50 iterations, stopping at the first criterion reached.

Next, we consider changing the index structure of the model to exploit any 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 these dropped predictors, and the alternating updating process in the previous step is repeated. This increment 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.

Note that, 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 lower bound (Bertsimas, King & Mazumder 2016). According to Hazimeh, Mazumder & Radchenko (2023), several methods have been used to select M in practice. For a description on estimating M in a linear regression setting, refer to Bertsimas, King & Mazumder (2016).

The following ***Algorithm 1*** summarises the key steps of the SMI Modelling algorithm.

Algorithm 1: SMI Modelling Algorithm

1. Initialise index structure and index coefficients $\boldsymbol{\alpha}$:
   Initialise p, predictor grouping among indices, and obtain $\boldsymbol{\alpha}^{init}$ using one of the five options in Section 2.2. Then scale each $\hat{\boldsymbol{\alpha}}_j = \boldsymbol{\alpha}_j^{init}$ to have unit norm.
2. Estimate nonlinear functions g_js:
   Estimate g_js using a GAM taking y_i as the response, $\hat{h}_{ij} = \hat{\boldsymbol{\alpha}}_j^T \boldsymbol{x}_i$s as predictors.
3. Update index coefficients $\boldsymbol{\alpha}$:
   Estimate updated value $\boldsymbol{\alpha}^{new}$ through the MIQP in Equation 6, and scale each $\hat{\boldsymbol{\alpha}}_j = \boldsymbol{\alpha}_j^{new}$ to have unit norm.

4. Iterate steps 2 and 3 until convergence, loss increase for three consecutive iterations, or reaching the maximum iterations.
5. Update index structure:

   Include a new index consisting of dropped predictors if applicable, and proceed to step 4. Otherwise, terminate the algorithm.
6. Iterate step 5 with increased number of indices p:

   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, in estimating all the SMI Models.

3 Benchmark Methods

Before demonstrating the SMI model on simulated and real data, we will briefly introduce three popular benchmark methods that can be used in comparisons.

Backward Elimination

Fan & Hyndman (2012) used a stepwise procedure for variable selection through cross-validation to find a model of the form of Equation 2. In each half-hourly model, the data is split into training and validation sets, and the predictors are selected based on the Mean Absolute Percentage Error (MAPE) calculated for the validation set. Starting from the full model with all predictors, the predictive power of each variable is evaluated by dropping one at a time. When the validation MAPE increases with the exclusion of a predictor, the predictor is omitted from the model in subsequent steps.

3.1 Projection Pursuit Regression

Friedman & Stuetzle (1981) introduced ***Projection Pursuit Regression (PPR)*** given by

$$y_i = \sum_{j=1}^{q} g_j(\boldsymbol{\alpha}_j^T \boldsymbol{x}_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

where y_i is the response, $\boldsymbol{x}_i$ is a p-dimensional predictor vector, $\boldsymbol{\alpha}_j = (\alpha_{j1}, \ldots, \alpha_{jp})^T, j = 1, \ldots, q$ are p-dimensional projection vectors (or vectors of *"index coefficients"*), g_j's are unknown univariate functions, and ε_i is the random error.

Instead of estimating a single index, PPR estimates multiple indices and connects them to the response through a sum of univariate nonlinear functions. These indices are constructed through a *Projection Pursuit (PP)* (Kruskal 1969; Friedman & Tukey 1974) algorithm, which is considered to be "interesting" low-dimensional projections of a high-dimensional feature space, obtained through the maximisation of an appropriate objective function or a "projection index" (Huber 1985).

According to Zhang et al. (2008), PPR increases the power of additive models in high-dimensional settings, but it has two major drawbacks. Firstly, since PP increases the freedom of the additive model, it tends to overfit in a situation, where there are a lot of unimportant predictors. Secondly, the interpretation of the model estimated by PPR will be troublesome as many non-zero elements will be present in each projection vector $\boldsymbol{\alpha}_j$. To overcome these issues, Zhang et al. (2008) introduced an ℓ_1 regularised projection pursuit algorithm, where the resultant regression model is named as ***Sparse Projection Pursuit Regression*** (SpPPR). In SpPPR, an ℓ_1 penalty (i.e. a LASSO penalty) on index coefficients is added to the cost function (the squared error) at each iteration of the PP, thereby performing variable selection and model estimation simultaneously. See Zhang et al. (2008) for more details.

Although Zhang et al. (2008) claimed that the SpPPR algorithm can detect important predictors even in a noisy data set, our experiments show that it is not particularly scalable for large data sets with both higher number of predictors and observations.

3.2 Group-wise Additive Index Model

Even though PPR introduces flexibility and the ability to model interactions among predictors into additive models, the indices obtained through PPR contain all the predictors at hand. Hence, even with a variable selection mechanism like SpPPR (Zhang et al. 2008), PPR creates indices possibly by mixing heterogeneous variables in a single linear combination, making model interpretation challenging (Masselot et al. 2022).

Typically, in many real-world problems, natural groupings can be identified in predictor variables. For example, naturally interacting variables can be grouped together, such as several lags of a predictor, weather related variables, and genes or proteins that are grouped by biological pathways in a biological study (Masselot et al. 2022; Wang, Xu & Zhu 2015).

This suggests the use of a ***Group-wise Additive Index Model (GAIM)***, which can be written as

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

where y_i is the univariate response, $\boldsymbol{x}_{ij} \in \mathbb{R}^{lj}$, $j = 1, \ldots, p$ are naturally occurring p groups of predictors, which are p non-overlapping subsets of $\boldsymbol{x}_i$ - the vector of all predictors, $\boldsymbol{\alpha}_j$ is a l_j-dimensional vector of index coefficients corresponding to the index $h_{ij} = \boldsymbol{\alpha}_j^T \boldsymbol{x}_{ij}$, 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. 2022).

Since GAIM uses groups of predictors that are naturally or logically belonging together to construct indices, such derived indices will be more expressive and interpretable. However, at the same time, this introduces a certain level of subjectivity into the model formulation as different users can group the available predictors in different ways based on different logical reasoning.

In this paper, our aim is to reduce that subjectivity induced by personal judgement or domain expertise. Hence, we propose a methodology that injects more objectivity into the estimation of multiple index models by algorithmically grouping predictors into indices, resulting in a model with a higher predictive accuracy.

4 Simulation Experiment

This section presents the results of a simple simulation experiment designed to demonstrate the performance and characteristics of the proposed SMI Modelling algorithm. Particularly, we try to investigate how the estimated SMI Model varies depending on the initialisation (as discussed in Section 2.2) used.

4.1 Data Generation

Generating predictor variables:

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

Generating response variables:

We generated two response variables y_1 and y_2, with two different index structures: single-index and 2-index, and added a random normal noise component with two different strengths as follows:

- Low noise level - $N(\mu = 0, \sigma = 0.1)$:

  $y_1 = (0.9 * x_0 + 0.6 * x_1 + 0.45 * x_3)^3 + \epsilon, \quad \epsilon \sim N(0, 0.01)$

  $y_2 = (0.9 * x_0 + 0.6 * x_1 + 0.45 * x_3)^3 + (0.35 * x_2 + 0.7 * x_5)^2 + \epsilon, \quad \epsilon \sim N(0, 0.01)$

- High noise level - $N(\mu = 0, \sigma = 0.5)$:

  $y_1 = (0.9 * x_0 + 0.6 * x_1 + 0.45 * x_3)^3 + \epsilon, \quad \epsilon \sim N(0, 0.25)$

  $y_2 = (0.9 * x_0 + 0.6 * x_1 + 0.45 * x_3)^3 + (0.35 * x_2 + 0.7 * x_5)^2 + \epsilon, \quad \epsilon \sim N(0, 0.25)$

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 were 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 of 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.

4.2 Experiment Setup

We estimated SMI Models through the proposed algorithm for each of the two response variables (the two "true models"), using three different sets of predictors as inputs. Our aim was to assess the algorithm's capability to correctly pick the relevant predictor variables (and drop the irrelevant predictors), and to estimate the correct 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. A part of $\boldsymbol{x}$ variables (i.e. x_0, x_1 and x_2) and all $\boldsymbol{z}$ variables (denoted as some $\boldsymbol{x}$ + all $\boldsymbol{z}$).

We applied the proposed SMI Modelling algorithm with each of the above predictor combinations, for both variations of the responses concerning the noise level. Moreover, we considered each of the first four initialisation options that we discussed in Section 2.2, for each of the two responses.

4.3 Results

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

Table 1: *Simulation experiment results.*

True Model	Predictors	PPR	Additive	Linear	Multiple
Low noise level					
y_1	all $\boldsymbol{x}$	1 index	1 index	1 index	1 index
		(x_0, x_1, x_3)	(x_0, x_1, x_3)	(x_0, x_1, x_3)	(x_0, x_1, x_3)
	all $\boldsymbol{x}$ + all $\boldsymbol{z}$	1 index	1 index	1 index	1 index
		(x_0, x_1, x_3)	(x_0, x_1, x_3)	(x_0, x_1, x_3)	(x_0, x_1, x_3)
	some $\boldsymbol{x}$ + all $\boldsymbol{z}$	1 index	3 indices	1 index	1 index
		(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 $\boldsymbol{x}$	2 indices	2 indices	1 index	2 indices
		(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)
	all $\boldsymbol{x}$ + all $\boldsymbol{z}$	2 indices	2 indices	1 index	2 indices
		(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)
	some $\boldsymbol{x}$ + all $\boldsymbol{z}$	3 indices	2 indices	1 index	2 indices
		(x_0, x_1, z_2) (x_2) (z_3, 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 $\boldsymbol{x}$	1 index	2 indices	1 index	1 index
		(x_0, x_1, x_3)	(x_0, x_1) (x_3)	(x_0, x_1, x_3)	(x_0, x_1, x_3)
	all $\boldsymbol{x}$ + all $\boldsymbol{z}$	1 index	2 indices	1 index	2 indices
		(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)
	some $\boldsymbol{x}$ + all $\boldsymbol{z}$	3 indices	3 indices	1 index	3 indices
		(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 $\boldsymbol{x}$	3 indices	3 indices	2 indices	3 indices
		(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)
	all $\boldsymbol{x}$ + all $\boldsymbol{z}$	2 indices	2 indices	1 index	2 indices
		(x_0, x_1, x_3) (x_2, x_5, z_1)	(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)
	some $\boldsymbol{x}$ + all $\boldsymbol{z}$	3 indices	2 indices	1 index	2 indices
		(x_0, x_1, z_0, z_3) (x_2) (z_1, z_4, z_5)	$(x_0, x_1, z_0, z_1, z_3, z_4)$ (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)

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.

At low noise level, in both the cases "all $\boldsymbol{x}$" and "all $\boldsymbol{x}$ + all $\boldsymbol{z}$", all four initialisations enable the algorithm to estimate the correct index structure for both y_1 and y_2, with an exception in "Linear" option for y_2. The "Linear" option for y_2 selects the correct variables, but fails to identify the 2-index structure. This suggests that initialising the algorithm with a higher number of indices might be more effective than 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 four 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 the cases "all $\boldsymbol{x}$" and "all $\boldsymbol{x}$ + all $\boldsymbol{z}$", all four initialisations resulted in a test MSE of ≈ 0.01, which is the random squared error of the true model. This confirms the accuracy with which the SMI Modelling algorithm estimated the index structure for

y_1. For y_2, all the models estimated resulted in a test MSE of ≈ 0.16. This is an interesting result as the test 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. For both y_1 and y_2, in the case of "some $\boldsymbol{x}$ + all $\boldsymbol{z}$", the test MSEs increased in comparison to the above cases, probably due to the inclusion of noise variables.

Moreover, in contrast to y_1, the test MSE values for y_2 are higher than the random squared error of the corresponding true model. Intuitively, the complexity of the model y_2 is higher than y_1, where the total estimation error of two nonlinear link 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, in comparison to the low noise level. For both y_1 and y_2, most of the estimated models have selected irrelevant variables. In both the cases "all $\boldsymbol{x}$" and "all $\boldsymbol{x}$ + all $\boldsymbol{z}$", all the models estimated for y_1 (except for "Additive" option in "all $\boldsymbol{x}$" case) resulted in a test MSE of ≈ 0.23 (which is slightly lower than the random squared error of the true model; this probably indicates a slight level of over-fitting), irrespective of the fact that in "all $\boldsymbol{x}$ + all $\boldsymbol{z}$" case, "Additive" and "Multiple" options included a noise variable. This is an indication of the effect of the low signal-to-noise ratio in the data. The observation 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 produced similar test MSE values.

Similar to the previous case of low noise level, when only a part of $\boldsymbol{x}$ variables are provided, the test MSE values increased for both y_1 and y_2, where the estimated models for y_2 produced 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, the users are encouraged to follow a trial-and-error procedure to determine the most suitable initial model for a given application.

5 Empirical Applications

5.1 Forecasting Daily Mortality

We apply the SMI Modelling algorithm to a data set from Masselot et al. (2022), 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 etc. on human health, is of significant importance in environmental epidemiology. Therefore, forecasting daily deaths taking heat related variables as predictors is an interesting application.

Description of the Data

For this analysis, we consider daily mortality and heat exposure data for the Metropolitan Area of Montreal, Province of 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, Province of Québec, while *DayMet* — a 1 km × 1 km grid data set (Thornton et al. 2021) was used to extract daily temperature and humidity data (Masselot et al. 2022).

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.

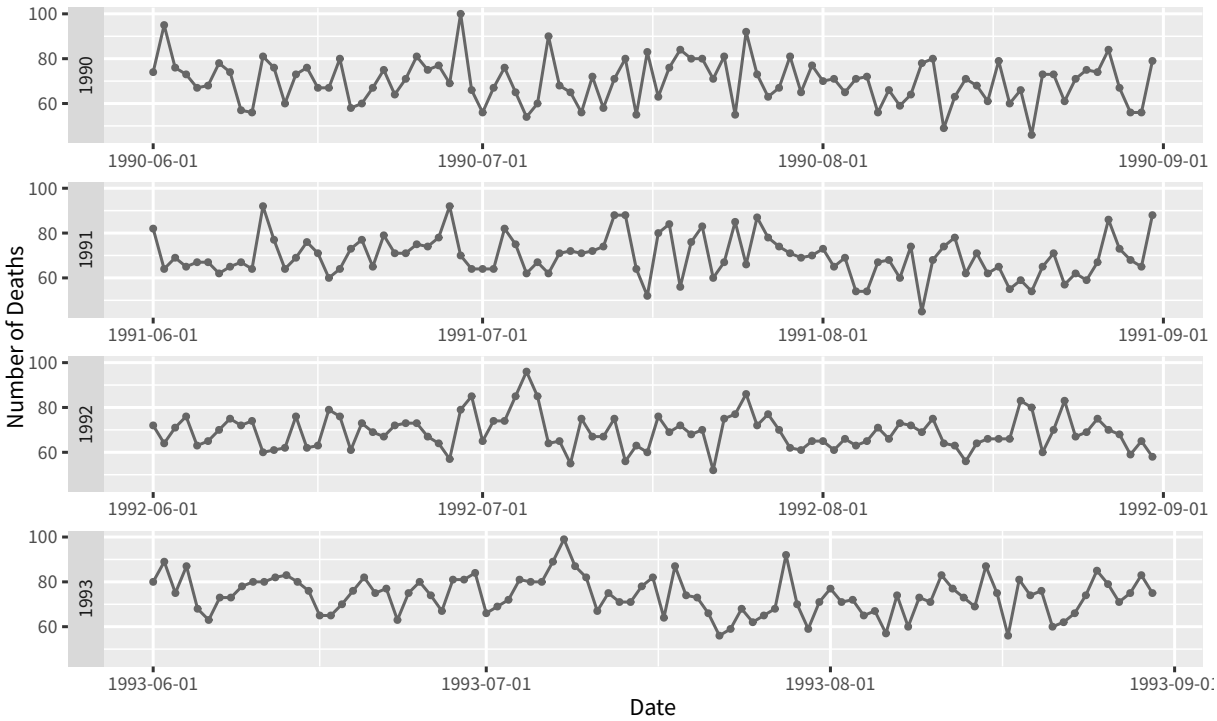

Figure 1: *Daily mortality in summer in Montreal, Canada from 1990 to 1993.*

The three main predictors considered in this empirical study are maximum temperature, minimum temperature, and vapour pressure (to represent the level of humidity). The number of daily deaths

are plotted against each of these predictors in Figure 2, Figure 3, and Figure 4, respectively, where we can observe that the relationships between these predictors and the response are slightly non-linear.

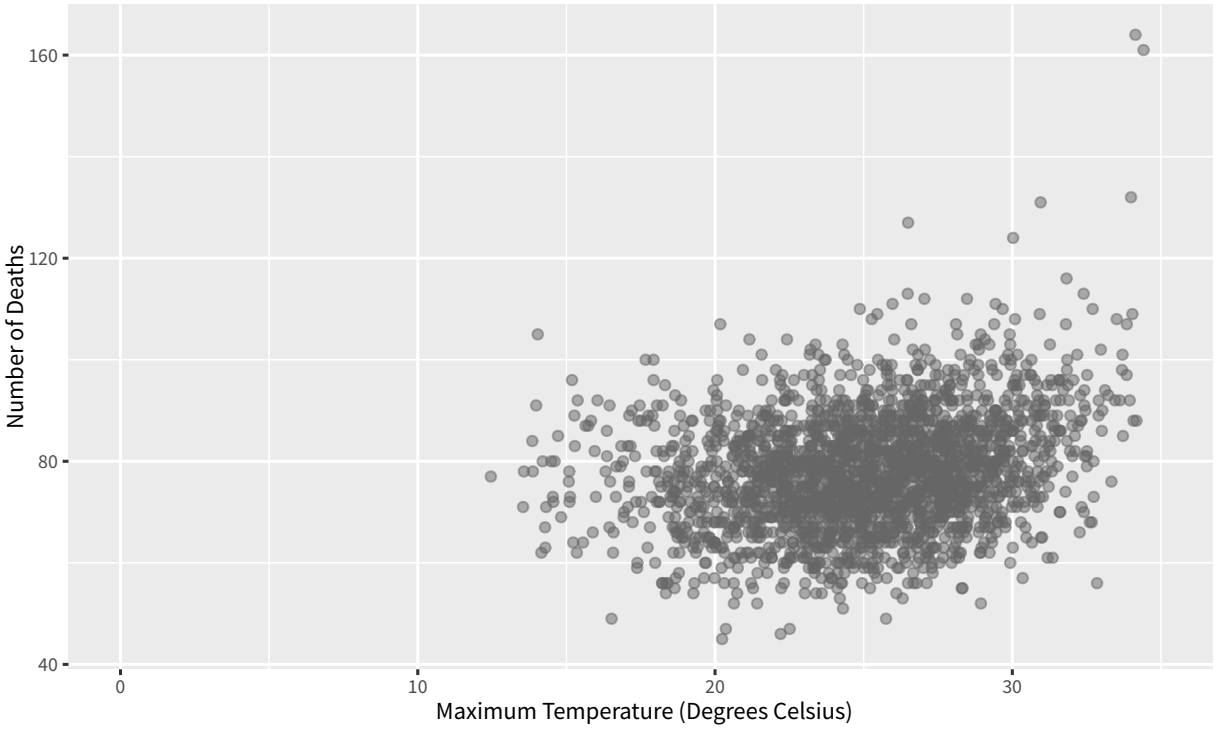

Figure 2: *Daily mortality in summer (from 1990 to 2014) plotted against maximum temperature.*

Predictors Considered

1) Current maximum/minimum temperatures and lags:

In addition to current maximum and minimum temperatures, the temperature measurements up to 14 days prior (i.e. 0^{th} to 14^{th} lag) 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.

2) Current vapour pressure and lags:

Similar to temperature variables, the current value and 14 lags of vapour pressure are considered as predictors, as a proxy to the level of humidity.

3) 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, and also to control the autocorrelation in residuals, which is a common practice in environmental epidemiology (Masselot et al. 2022).

Modelling Framework

Maximum temperature lags, minimum temperature lags, and vapour pressure lags are considered as predictors entering indices. The two calendar variables, *DOS* and *Year*, are included into the model as

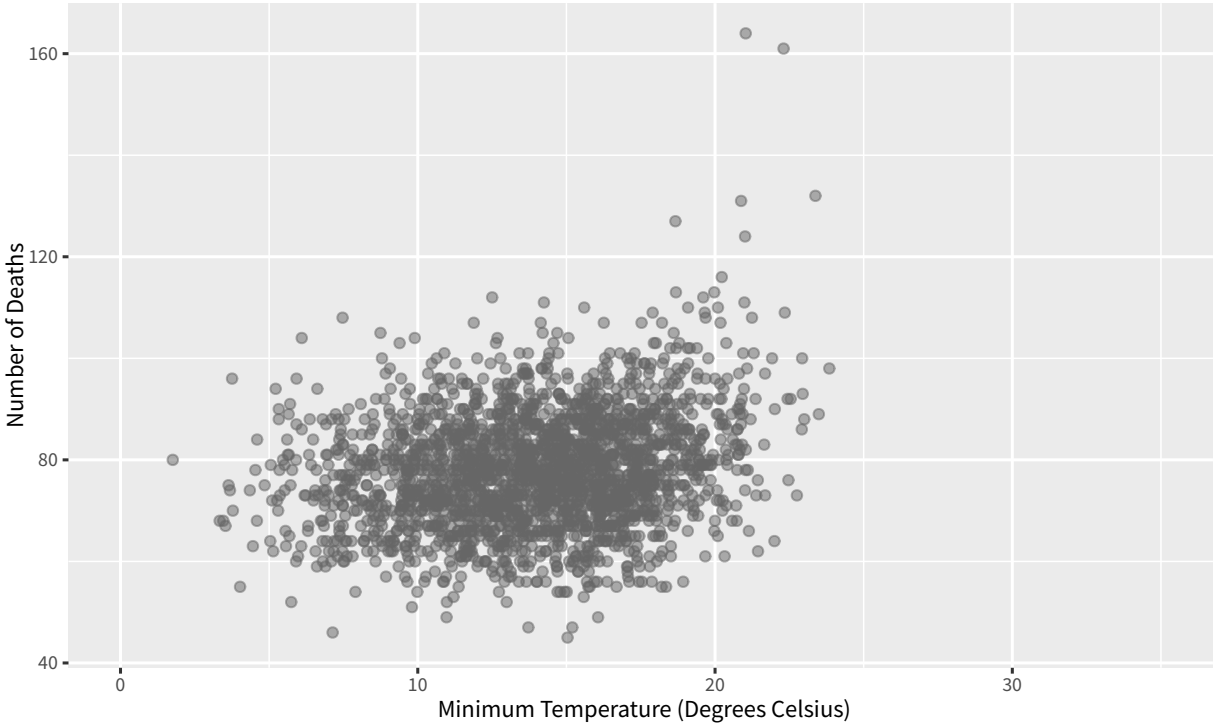

Figure 3: *Daily mortality in summer (from 1990 to 2014) plotted against minimum temperature.*

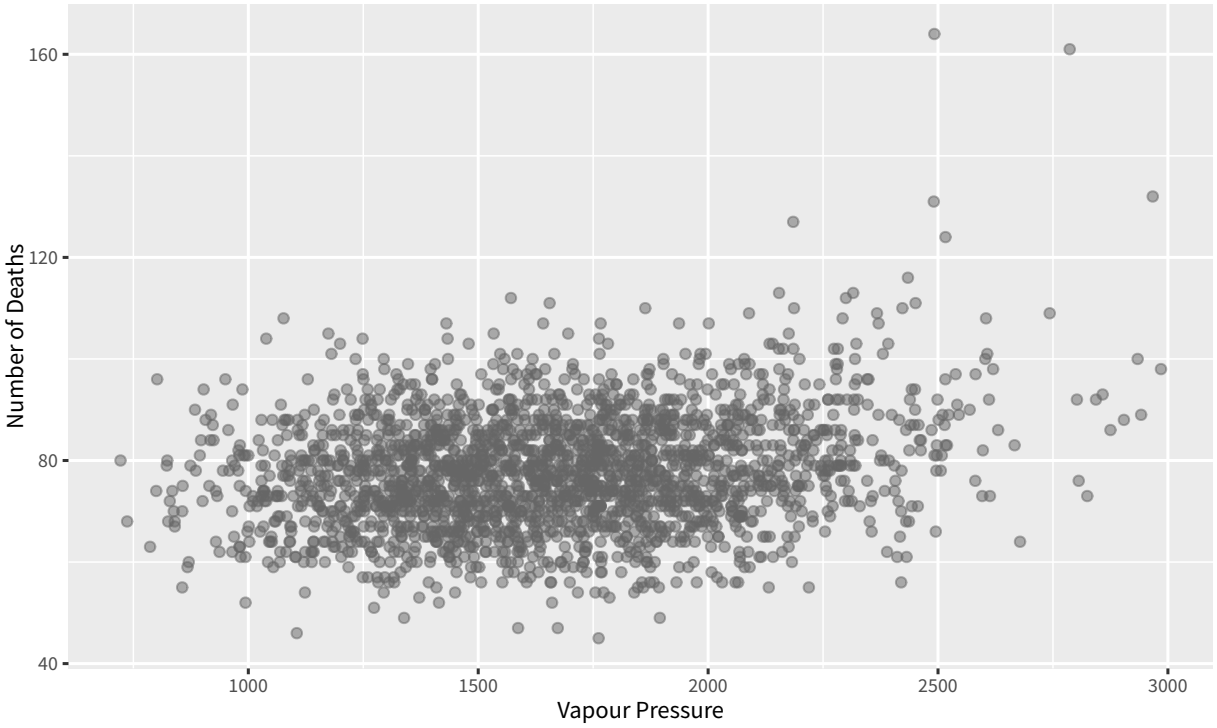

Figure 4: *Daily mortality in summer (from 1990 to 2014) plotted against vapour pressure.*

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 that need to be estimated through the algorithm;
- $\boldsymbol{X}$ is the matrix containing the predictor variables that are entering indices (i.e. maximum temperature lags, minimum temperature lags, and vapour pressure lags);
- $\boldsymbol{\alpha}_j, j = 1, \ldots, p$ are the index coefficient vectors, each with a length equal to the number of predictors entering indices ($q = 45$);
- $g_j, j = 1, \ldots, p$, 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 of year 2014 are separated to be the test set for evaluating forecasting performance. The data from the three summer months of year 2013 are kept aside as a validation set, which is used to estimate benchmark models for comparison purposes.

Then we apply the proposed SMI Modelling algorithm to the training set to estimate the model. Finally, the forecasting accuracy on the test set is evaluated using MSE and Mean Absolute Error (MAE).

Results

We estimated SMI Models for the mortality data using three different initialisation options: "PPR", "Additive" and "Linear", for comparison purposes. Through our pre-experiments on the new algorithm, we identified that the "PPR" initialisation option has a higher probability of better performance, whereas "Additive" (i.e. Additive Model) and "Linear" (i.e. Single Index Model) are two special cases of the SMI Model. We did not consider "Multiple" and "User Input" initialisations here as both of these two options require user specific inputs to some extend. Further, we tuned the penalty parameters λ_0 and λ_2, over ranges of integers from 1 to 12, and 0 to 12 respectively, through a greedy search based on in-sample MSE. Here, a greedy search is used instead of a grid search to reduce computational time.

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

Model	Predictors	Indices	Test Set 1		Test Set 2	
			MSE	MAE	MSE	MAE
smimodel(12, 0) - PPR	47	5	**80.334**	**6.841**	**99.926**	**7.643**
smimodel(1, 0) - Additive	47	45	151.408	9.816	190.880	11.107
smimodel(12, 5) - Linear	47	2	164.629	10.153	207.040	11.141
Backward Elimination	36	NA	148.387	9.808	162.608	10.034
GAIM	47	3	85.145	7.257	103.494	8.480
PPR	47	3	82.877	7.202	104.217	8.404

The penalty parameter combination ($\lambda_0 = 12, \lambda_2 = 0$) was selected for the model fitted with "PPR" initialisation. The estimated model, ***SMI Model (12, 0) - PPR*** , resulted in five 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 ($\lambda_0 = 12, \lambda_2 = 5$) (***SMI Model (12, 5) - 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 3 months (June, July and August 2014); and
2. ***Test Set 2:*** a test set covering 1 month (June 2014).

Note that in this application, we assumed that the future values of the maximum/minimum temperatures and vapour pressure are known to use in the forecasting model.

The MSE and MAE values for the estimated SMI Models on two different test sets are presented in Table 2. We observe that the SMI Model estimated with "PPR" initialisation, ***SMI Model (12, 0) - PPR*** , shows the best forecasting performance on both test sets, compared to the other two estimated SMI models.

Furthermore, we present the forecasting errors of three benchmark models in Table 2 for comparison with the estimated SMI Models. The first benchmark is a nonparametric additive model formulated through backward elimination, as proposed by Fan & Hyndman (2012) (Section 3). Next, a GAIM (Section 3.2) is also presented. In the case of GAIM, maximum temperature lags, minimum temperature lags, and vapour pressure lags are categorised into three groups, where an index estimated for each group. Finally, we present the forecasting errors of a PPR model. The number of indices in the PPR model was taken as 3, matching the number of indices estimated by the GAIM.

Table 2 shows that ***SMI Model (12, 0) - 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 all benchmark models considered. The actual number of deaths and the predicted values from the ***SMI Model (12, 0) - PPR*** and benchmark models on *Test Set 2* are plotted in Figure 5 for further comparison.

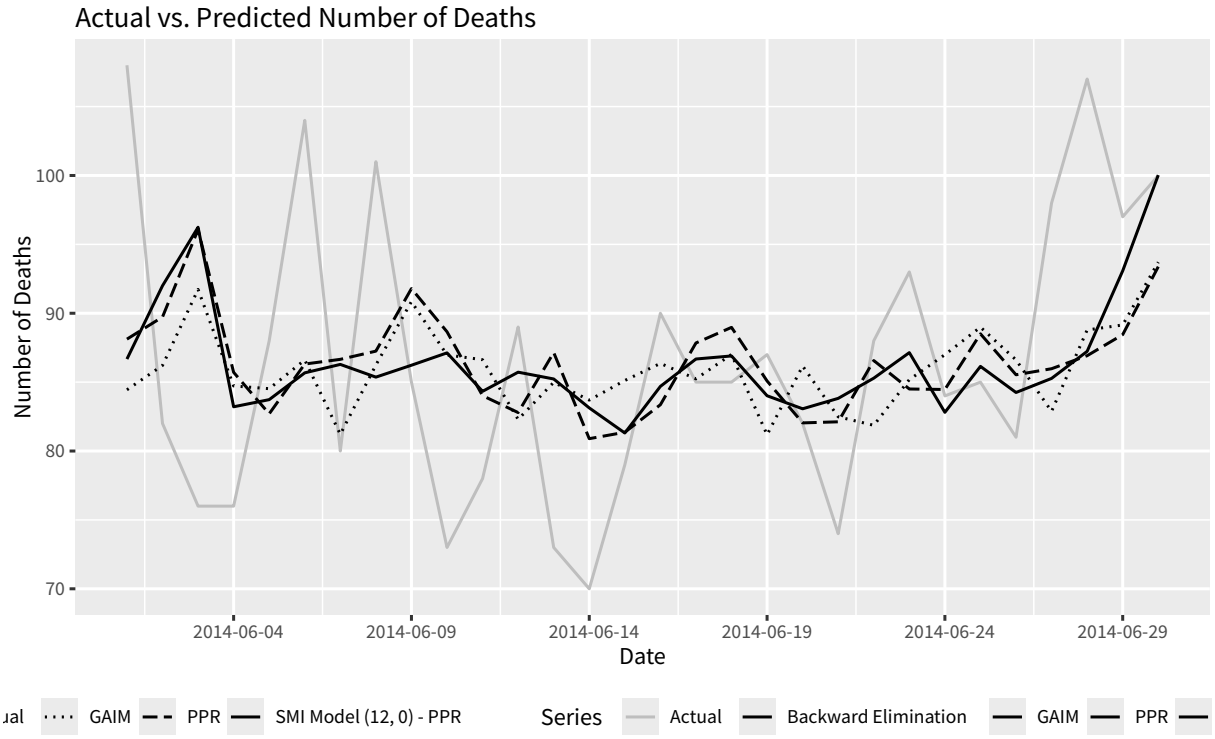

Figure 5: *Actual number of deaths vs. predicted number of deaths from "SMI Model (12, 0) - PPR" and benchmark models for Test Set 2.*

5.2 Forecasting Daily Solar Intensity

Next, we utilise the SMI Modelling algorithm to forecast daily solar intensity, using other weather conditions. As reported by Energy Institute (2023), renewable energy (excluding hydroelectricity) contributed to 7.5% of the world's primary energy consumption in 2022. Solar and wind power saw a combined capacity addition of 266 GW, with solar energy accounting for 72% of the increase. Given this, accurate forecasting of solar power generation, closely linked to solar intensity, is crucial for effective power system planning and management.

Description of the Data

We use solar intensity and other weather variables measured at a 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 21th February 2006 to 27th February 2013, using sensors for measuring temperature, wind chill, humidity, dew point, wind speed, wind direction, rain,

pressure, solar intensity, and UV.

However, the data contained missing entries recorded as "-100000", which we removed from the data set. Moreover, for this analysis, we converted the five minutes data to daily data by averaging each variable over days.

Figure 6 shows the time plot of daily solar intensity for the entire period, which clearly depicts the annual seasonality in the data. As observed in Figure 6, there are days for which the observations were missing. We excluded those days from the analysis, and used only the days for which the data are available.

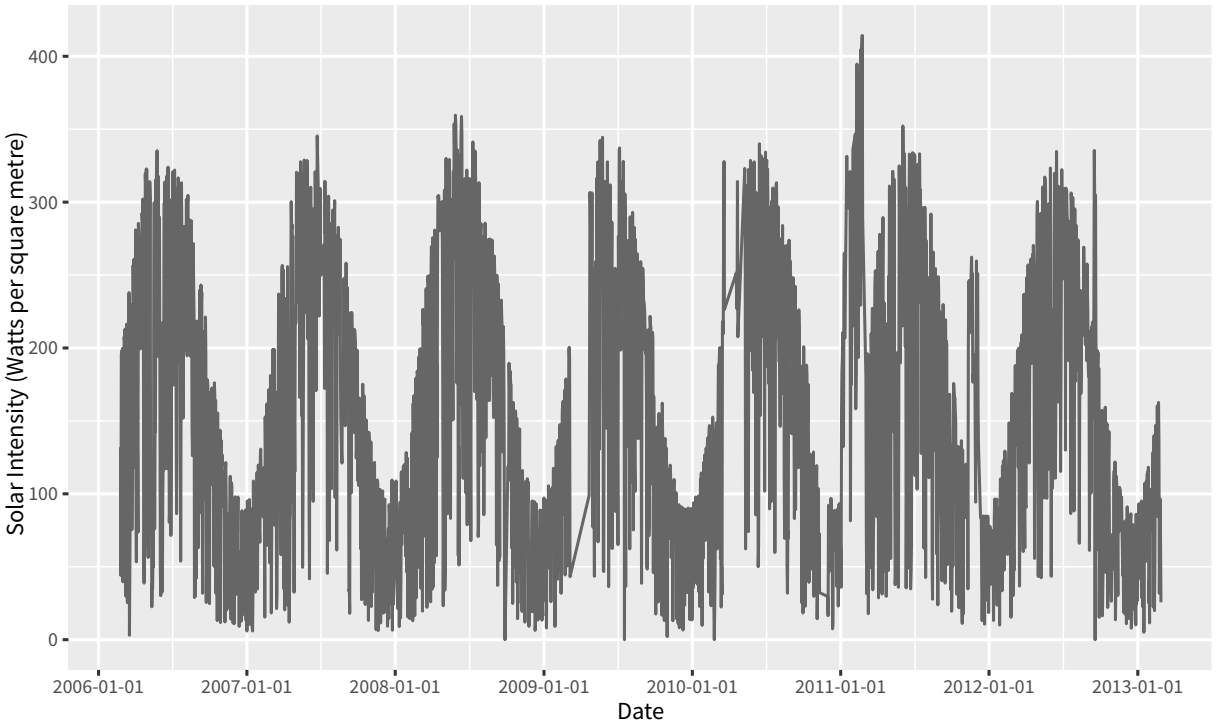

Figure 6: *Daily solar intensity in Amherst, Massachusetts - from February 2006 to February 2013.*

The variables temperature, dew point, wind, rain and humidity were considered to be the main set of predictors in the model. The daily solar intensity is plotted against each of these predictors in Figure 7, where we can observe that the relationships between these predictors and the response are non-linear.

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. Intuitively, the solar intensity of a particular day would have a relationship to the solar intensity of adjacent days.

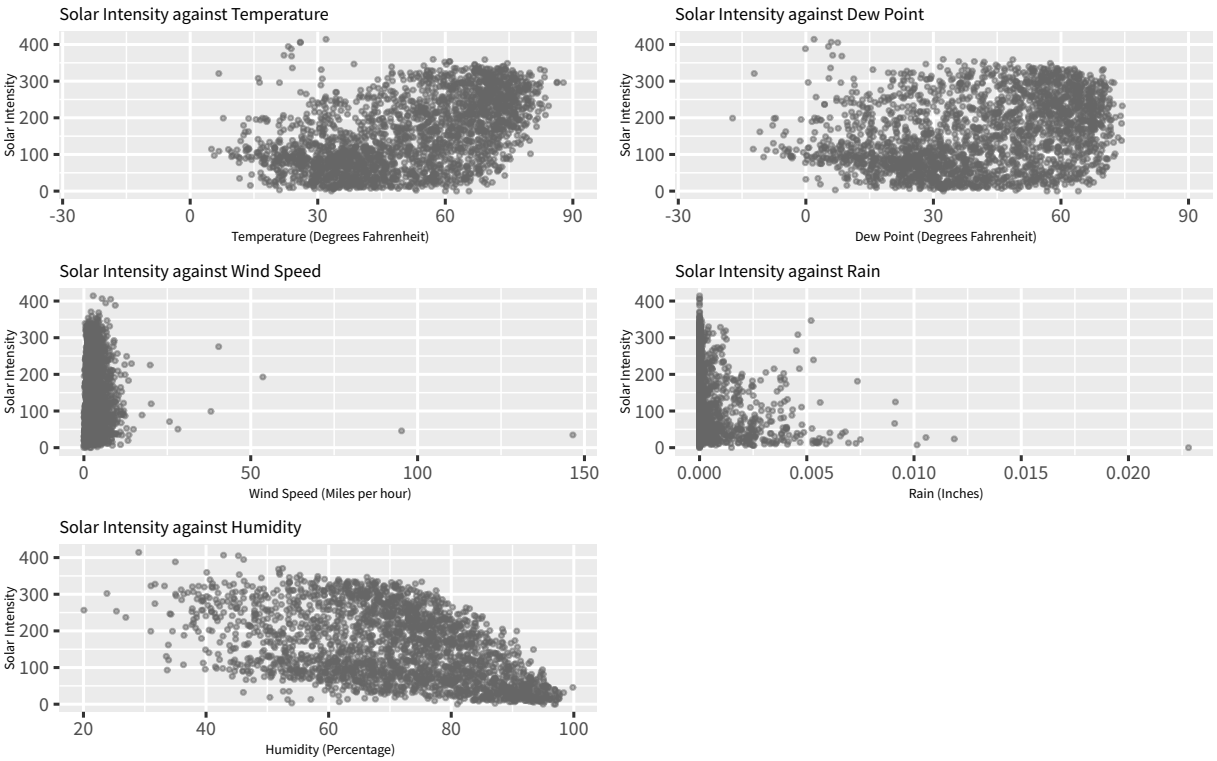

Figure 7: *Daily solar intensity against other weather variables.*

2) Current weather variables and lags:

In addition to current temperature, dew point, wind speed, rain and humidity, the measurements of three previous days (i.e. 0^{th} to 3^{rd} lag) for each of the these weather variables are also included as predictors in the forecasting model.

3) Calendar effects:

Finally, a couple of calendar variables; *Month* (12 months of the year) and *Season* (the four seasons: Spring, Summer, Autumn and Winter), are incorporated into the model to capture annual seasonality, and control for autocorrelation in residuals.

Modelling Framework

The lags of solar intensity, and the lags of weather variables are considered as predictors that are entering indices. The two calendar variables, *Month* and *Season*, are included into the model as linear (categorical) predictor variables.

Hence, the relevant SMI Model can be written as

$$\mathbf{Solar} = \beta_0 + \sum_{j=1}^{p} g_j(\boldsymbol{X}\boldsymbol{\alpha}_j) + \theta_1 \mathbf{Month} + \theta_2 \mathbf{Season} + \boldsymbol{\varepsilon}, \tag{8}$$

where

- **Solar** is the vector containing daily observations of solar intensity;
- β_0 is the model intercept;
- p is the unknown number of indices that will be estimated through the algorithm;
- $\boldsymbol{X}$ is the matrix containing the predictor variables that are entering 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 lenth equal to the number of predictors entering indices ($q = 23$);
- $g_j, j = 1, \ldots, p$ are unknown nonparametric functions;
- θ_1 and θ_2 are the two coefficients corresponding to the two linear predictor variables; 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 the months January and February 2013 are separated to be the test set to evaluate the forecasting performance. The data from the months November and December 2013 are kept aside as a validation set, which is required to estimate some of the benchmark models for comparison.

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.

Results

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

The penalty parameter combination ($\lambda_0 = 12, \lambda_2 = 0$) was selected for the model fitted with "PPR" initialisation. The estimated model, ***SMI Model (12, 0) - 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 nonparametric 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 *Month* and *Season*. Notice that all three estimated SMI Models have $\lambda_2 = 0$, indicating that all three models have omitted the ℓ_2-penalty in the estimation process.

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

Model	Predictors	Indices	Test Set	
			MSE	MAE
SMI Model (12, 0) - PPR	25	5	1745.030	33.246
SMI Model (1, 0) - Additive	25	23	1112.181	26.975
SMI Model (1, 0) - Linear	2	0	2009.847	35.346
Backward Elimination	16	NA	911.570	25.035
GAIM	25	6	2203.530	37.788
PPR	23	6	**796.779**	**22.455**

Note that similar to the previous application of heat related mortality forecasting, we assumed that the future values of the weather variables are known to use in the forecasting model.

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.

Similar to Section 5.1, we also present forecasting errors of 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 a different 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 two categorical calendar variables were excluded when estimating the PPR model.

According to Table 3, the forecasting errors of ***SMI Model (1, 0) - Additive*** is lower than the GAIM. However, the ***SMI Model (1, 0) - Additive*** is unable to outperform both the semi-parametric additive model with backward elimination and the PPR model, where in this case, the estimated PPR model has resulted in the best forecasting accuracy.

Here, it is worth considering the differences between the SMI Model and the benchmark models that show superior forecasting performance. The method proposed by Fan & Hyndman (2012) formulates a semi-parametric additive model using a backward elimination of predictors. When estimating a PPR model, both the number of indices and the predictors within each index (each index includes all provided predictors that are entering indices) are pre-determined. In contrast, the SMI Model takes a more general and objective approach, 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.

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.

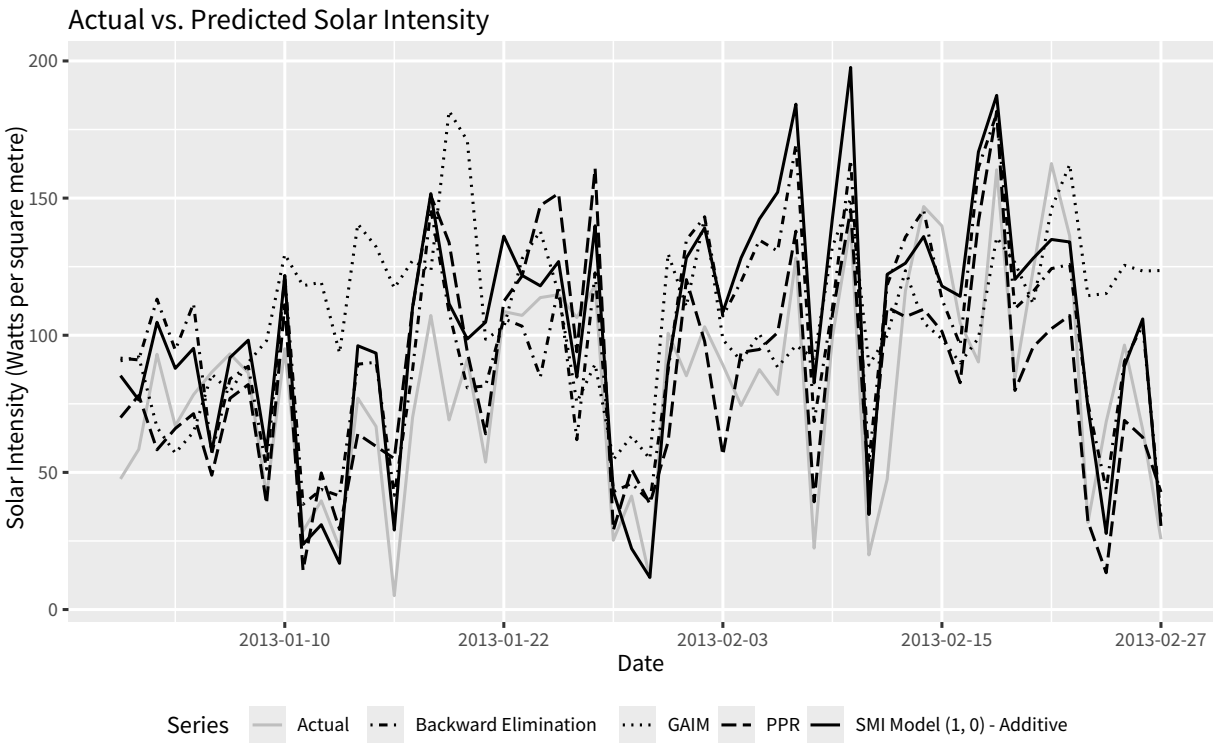

Figure 8: *Actual solar intensity vs. predicted solar intensity from "SMI Model (1, 0) - Additive" and benchmark models.*

In summary, the two empirical applications presented above highlight the challenge of finding a universally applicable initialisation option for the SMI Model across various applications. As mentioned in Section 4, we encourage users to follow a trial-and-error procedure to identify the most effective initialisation option for their specific application.

The two empirical applications were performed using R statistical software (R Core Team 2023), and the Rstudio integrated development environment (IDE, Posit 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*** (v1.9.1, Wood 2011).

6 Conclusions and Further Research

In this paper, we presented a novel algorithm for estimating a nonparametric additive index model with optimal predictor selection, 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 non-liner and linear predictors in the model that are not entering any indices, allowing the estimation of semi-parametric additive models.

Due to the limited input information provided to the algorithm, the estimation of a SMI Model is a challenging problem. We demonstrated the performance of the proposed algorithm through a simple simulation and two empirical applications. Since we observed that the final estimated model changes with the choosen initialisation, one limitation of the proposed algorithm is the difficulty of specifying an initialisation that works in general. Hence, an interesting future research problem would be to explore the potential for determining a generalised initialisation for the SMI Modelling algorithm that will work across various applications.

Moreover, we admit that the empirical examples presented in the paper may not be diverse enough to draw definitive conclusions about the unique strengths or weaknesses of the proposed algorithm. This study should be viewed as an 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 joining the 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.

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.

Energy Institute (2023). *Statistical Review of World Energy*. https://www.energyinst.org/__data/assets/pdf_file/0004/1055542/EI_Stat_Review_PDF_single_3.pdf.

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 & JW Tukey (1974). A Projection Pursuit Algorithm for Exploratory Data Analysis. *IEEE Transactions on Computers* **C-23**(9), 881–890.

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.

Huber, PJ (1985). Projection Pursuit. *Annals of Statistics* **13**(2), 435–475.

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

Ibrahim, S, R Mazumder, P Radchenko & E Ben-David (2022). "Predicting Census Survey Response Rates via Interpretable Nonparametric Additive Models with Structured Interactions". https://arxiv.org/abs/2108.11328.

Kruskal, JB (1969). "Toward a practical method which helps uncover the structure of a set of multivariate observations by finding the linear transformation which optimizes a new "index of condensation"". In: *Statistical Computation*. Ed. by Roy C. Milton and John A. Nelder. Academic Press, pp.427–440.

Masselot, P, F Chebana, C Campagna, É Lavigne, TBMJ Ouarda & P Gosselin (2022). Constrained groupwise additive index models. *Biostatistics* **00**(00), 1–19.

Mazumder, R, P Radchenko & A Dedieuc (2022). Subset Selection with Shrinkage: Sparse Linear Modeling When the SNR Is Low. *Operations Research*, 1–19.

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

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

Posit team (2024). *RStudio: Integrated Development Environment for R*. Posit Software, PBC. Boston, MA. http://www.posit.co/.

R Core Team (2023). *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*. https://traces.cs.umass.edu/index.php/Sensors/Sensors.

Wang, T, P Xu & L Zhu (2015). Variable selection and estimation for semi-parametric multiple-index models. *Bernoulli* **21**(1), 242–275.

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.

Zhang, X, L Liang, X Tang & HY Shum (2008). L1 regularized projection pursuit for additive model learning. In: *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp.1–8.