Sparse Multiple Index (SMI) Models for High-dimensional Nonparametric Forecasting

Nuwani Palihawadana

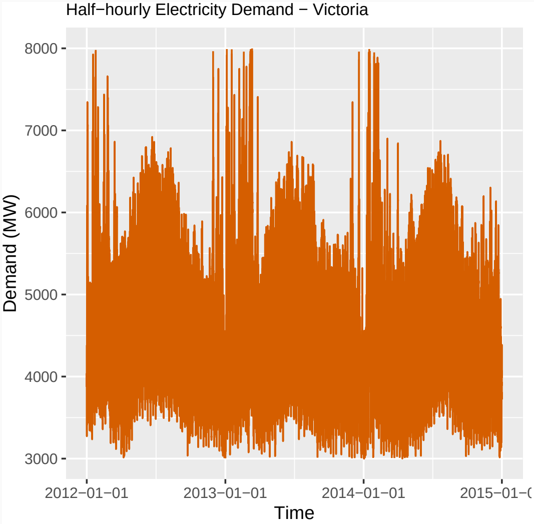
Joint work with : Rob Hyndman, Xiaoqian Wang

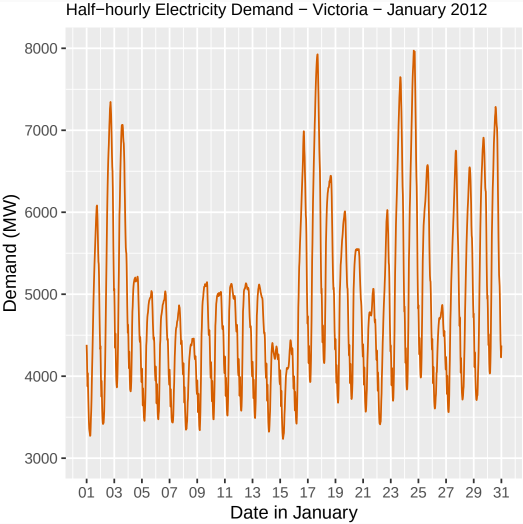
1 July 2024

Outline

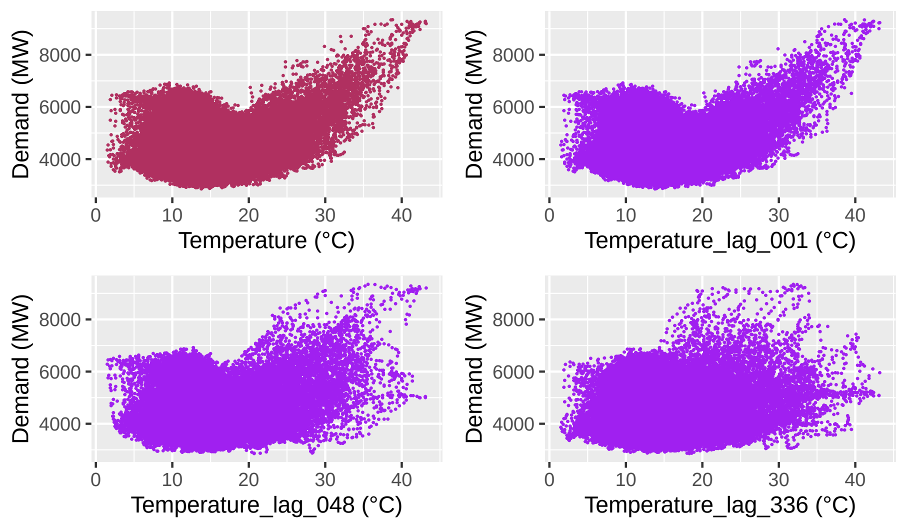
Outline

Electricity Demand Data





Electricity Demand Data



Outline

Background

- ***Nonlinear "Transfer Function" model***

$$y_t = f(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \dots, \boldsymbol{x}_{t-p}, y_1, \dots, y_{t-k}) + \varepsilon_t$$

y_t – variable to forecast

$\boldsymbol{x}_t$ – a vector of predictors

ε_t – random error

Background

- ***Nonlinear "Transfer Function" model***

$$y_t = f(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \dots, \boldsymbol{x}_{t-p}, y_1, \dots, y_{t-k}) + \varepsilon_t$$

y_t – variable to forecast
$\boldsymbol{x}_t$ – a vector of predictors
ε_t – random error

- Impossible to estimate f for large p – ***curse of dimensionality***

Background

- ***Nonlinear "Transfer Function" model***

$$y_t = f(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \dots, \boldsymbol{x}_{t-p}, y_1, \dots, y_{t-k}) + \varepsilon_t$$

y_t – variable to forecast
$\boldsymbol{x}_t$ – a vector of predictors
ε_t – random error

- Impossible to estimate f for large p – ***curse of dimensionality***
- Reasonable to impose additivity constraints

$$f(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \dots, \boldsymbol{x}_{t-p}) = \sum_{i=0}^{p} f_i(\boldsymbol{x}_{t-i})$$

Background

- ***Nonlinear "Transfer Function" model***

$$y_t = f(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \dots, \boldsymbol{x}_{t-p}, y_1, \dots, y_{t-k}) + \varepsilon_t$$

y_t – variable to forecast
$\boldsymbol{x}_t$ – a vector of predictors
ε_t – random error

- Impossible to estimate f for large p – ***curse of dimensionality***
- Reasonable to impose additivity constraints

$$f(\boldsymbol{x}_t, \boldsymbol{x}_{t-1}, \dots, \boldsymbol{x}_{t-p}) = \sum_{i=0}^{p} f_i(\boldsymbol{x}_{t-i}) \leftarrow \textbf{\textit{Nonparametric Additive Model}}$$

Background

Issues

1. **Challenging to estimate in a high-dimensional setting**
2. **Subjectivity in predictor selection, and predictor grouping to model interactions**

Background

Issues

1. **Challenging to estimate in a high-dimensional setting**
2. **Subjectivity in predictor selection, and predictor grouping to model interactions**

Index Models

- **Mitigate difficulty of estimating a nonparametric component for each predictor**
- **Improve flexibility**

Outline

SMI Model

Semi-parametric model

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

- y_i : univariate response
- $\boldsymbol{x}_{ij} \in \mathbb{R}^{\ell_j}, j = 1, \dots, p$: p subsets of predictors entering indices
- $\boldsymbol{\alpha}_j$: ℓ_j-dimensional vectors of index coefficients

SMI Model

Semi-parametric model

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

- y_i : univariate response
- $\boldsymbol{x}_{ij} \in \mathbb{R}^{\ell_j}$, $j = 1, \dots, p$: p subsets of predictors entering indices
- $\boldsymbol{\alpha}_j$: ℓ_j-dimensional vectors of index coefficients
- Additional predictors :
  - w_{ik} – nonlinear
  - $\boldsymbol{u}_i$ – linear
- g_j, f_k : smooth nonlinear functions

Optimisation Problem

Let q be the *total number of predictors* entering indices.

$$\min_{\beta_0, p, \boldsymbol{\alpha}, \boldsymbol{g}, \boldsymbol{f}, \boldsymbol{\theta}} \quad \sum_{i=1}^{n} \left[y_i - \beta_0 - \sum_{j=1}^{p} g_j(\boldsymbol{\alpha}_j^T \boldsymbol{x}_i) - \sum_{k=1}^{d} f_k(w_{ik}) - \boldsymbol{\theta}^T \boldsymbol{u}_i \right]^2$$

$$+ \lambda_0 \sum_{j=1}^{p} \sum_{m=1}^{q} \mathbb{1}(\alpha_{jm} \neq 0) + \lambda_2 \sum_{j=1}^{p} \|\boldsymbol{\alpha}_j\|_2^2$$

s.t. $$\sum_{j=1}^{p} \mathbb{1}(\alpha_{jm} \neq 0) \in \{0, 1\} \quad \forall m$$

Optimisation Problem

Let q be the *total number of predictors* entering indices.

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

- $\lambda_0 > 0$ – controls the number of selected predictors

Optimisation Problem

Let q be the *total number of predictors* entering indices.

$$\min_{\beta_0, p, \boldsymbol{\alpha}, \boldsymbol{g}, \boldsymbol{f}, \boldsymbol{\theta}} \quad \sum_{i=1}^{n} \left[y_i - \beta_0 - \sum_{j=1}^{p} g_j(\boldsymbol{\alpha}_j^T \boldsymbol{x}_i) - \sum_{k=1}^{d} f_k(w_{ik}) - \boldsymbol{\theta}^T \boldsymbol{u}_i \right]^2$$

$$+ \lambda_0 \sum_{j=1}^{p} \sum_{m=1}^{q} \mathbb{1}(\alpha_{jm} \neq 0) + \lambda_2 \sum_{j=1}^{p} \|\boldsymbol{\alpha}_j\|_2^2$$

$$\text{s.t.} \quad \sum_{j=1}^{p} \mathbb{1}(\alpha_{jm} \neq 0) \in \{0, 1\} \quad \forall m$$

- $\lambda_0 > 0$ – controls the number of selected predictors
- $\lambda_2 \geq 0$ – controls the strength of the additional shrinkage

MIQP Formulation

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

MIQP Formulation

$$\min_{\beta_0, p, \boldsymbol{\alpha}, \boldsymbol{g}, \boldsymbol{f}, \boldsymbol{\theta}, \boldsymbol{z}} \quad \sum_{i=1}^{n} \left[y_i - \beta_0 - \sum_{j=1}^{p} g_j(\boldsymbol{\alpha}_j^T \boldsymbol{x}_i) - \sum_{k=1}^{d} f_k(w_{ik}) - \boldsymbol{\theta}^T \boldsymbol{u}_i \right]^2$$

$$+ \lambda_0 \sum_{j=1}^{p} \sum_{m=1}^{q} z_{jm} + \lambda_2 \sum_{j=1}^{p} \sum_{m=1}^{q} \alpha_{jm}^2$$

$$\text{s.t.} \quad |\alpha_{jm}| \leq M z_{jm} \quad \forall j, \forall m,$$

$$\sum_{j=1}^{p} z_{jm} \leq 1 \quad \forall m,$$

$$z_{jm} \in \{0, 1\} \quad \leftarrow \quad z_{jm} = \mathbb{1}(\alpha_{jm} \neq 0)$$

MIQP Formulation

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

- $M < \infty$: If $\boldsymbol{\alpha}^*$ is an optimal solution, then $\max\left(\{|\alpha_{jm}^*|\}_{j \in [p], m \in [q]}\right) \leq M$

Estimation Algorithm

Step 1: Initialise Index Structure and Index Coefficients

- Obtain a feasible initialisation :

Estimation Algorithm

Step 1: Initialise Index Structure and Index Coefficients

- Obtain a feasible initialisation :

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

Estimation Algorithm

Step 1: Initialise Index Structure and Index Coefficients

- Obtain a feasible initialisation :

1. **PPR:** Projection Pursuit Regression Based Initialisation
2. **Additive:** Nonparametric Additive Model Based Initialisation

Estimation Algorithm

Step 1: Initialise Index Structure and Index Coefficients

- Obtain a feasible initialisation :

1. **PPR:** Projection Pursuit Regression Based Initialisation
2. **Additive:** Nonparametric Additive Model Based Initialisation
3. **Linear:** Linear Regression Based Initialisation

Estimation Algorithm

Step 1: Initialise Index Structure and Index Coefficients

- Obtain a feasible initialisation :

  1. **PPR:** Projection Pursuit Regression Based Initialisation
  2. **Additive:** Nonparametric Additive Model Based Initialisation
  3. **Linear:** Linear Regression Based Initialisation
  4. **Multiple:** Picking One From Multiple Initialisations

- Scale each $\hat{\alpha}_j$ to have unit norm

Estimation Algorithm

Step 2: Estimate Nonlinear Functions

- Estimate a GAM :

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

where

- y_i – response
- $\hat{h}_{ij} = \hat{\boldsymbol{\alpha}}_j^T \boldsymbol{x}_i, j = 1, \dots, p$ – estimated indices

Estimation Algorithm

Step 3: Update Index Coefficients

$$\min_{\boldsymbol{\alpha}^{\text{new}}, \boldsymbol{z}^{\text{new}}} (\boldsymbol{\alpha}^{\text{new}} - \boldsymbol{\alpha}^{\text{old}})^T \boldsymbol{V}^T \boldsymbol{V} (\boldsymbol{\alpha}^{\text{new}} - \boldsymbol{\alpha}^{\text{old}}) - 2(\boldsymbol{\alpha}^{\text{new}} - \boldsymbol{\alpha}^{\text{old}})^T \boldsymbol{V}^T \boldsymbol{r}$$

$$+ \lambda_0 \sum_{j=1}^{p} \sum_{m=1}^{q} z_{jm}^{\text{new}} + \lambda_2 \sum_{j=1}^{p} \sum_{m=1}^{q} (\alpha_{jm}^{new})^2$$

$$\text{s.t. } |\alpha_{jm}^{\text{new}}| \leq M z_{jm}^{\text{new}} \quad \forall j, \forall m,$$

$$z_{jm}^{\text{new}} \in \{0, 1\},$$

$$\sum_{j=1}^{p} z_{jm}^{\text{new}} \leq 1 \quad \forall m,$$

$\boldsymbol{V}$ – matrix of partial derivatives of RHS of SMI model equation, with respect to $\boldsymbol{\alpha}_j$

$\boldsymbol{r}$ – current residual vector

Estimation Algorithm

Step 4: Iterate steps 2 and 3

Until:

- convergence
- loss increases for three consecutive iterations or
- reaching maximum iterations

Estimation Algorithm

Step 4: Iterate steps 2 and 3

Until:

- convergence
- loss increases for three consecutive iterations or
- reaching maximum iterations

Step 5: Stop or repeat step 4

- No dropped predictors – **Stop**
- Otherwise, include a new index consisting of dropped predictors – **Repeat step 4**

Estimation Algorithm

Step 6: Increase p by 1 in each iteration of step 5

Until:

- number of indices reaches q – **output = final fitted model**
- loss increases after the increment – **output = previous iteration model** or
- solution maintains same number of indices as previous iteration, and absolute difference of index coefficients between two successive iterations is not larger than a pre-specified tolerance – **output = model with a smaller loss**

Outline

Data Generation

Predictor variables

- x_0 – Uniform [0,1]
- z_0 – Normal (5, 4)
- Construct lagged series of both x_0 and z_0 up-to lag 5

Data Generation

Predictor variables

- x_0 – Uniform [0,1]
- z_0 – Normal (5, 4)
- Construct lagged series of both x_0 and z_0 up-to lag 5

Response variables

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

Experiment Setup

- Three different sets of predictors:

1. All x variables
2. All x variables and all z variables
3. First three x variables (i.e. x_0, x_1 and x_2) and all z variables

- under each initialisation option
- for each response and noise level combinations

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 $\boldsymbol{z}$	$(x_0, x_1, z_0, z_3, z_4)\,(x_2)$	$(x_0, x_1, z_0, z_1, z_3, z_4)\,(x_2)$	$(x_0, x_1, x_2, z_0, z_3, z_4)$	$(x_0, x_1, z_0, z_1, z_3, z_4)\,(x_2)$

Outline

Forecasting Heat Exposure Related Mortality

Variables

- **Response:** **Daily deaths** – 1990 to 2014 – Montreal, Canada
- **Index Variables:**
  - Death lags
  - Max temperature lags
  - Min temperature lags
  - Vapor pressure lags
- **Nonlinear:** DOS, Year

Data Split

- **Training Set:** 1990 to 2012
- **Validation Set:** 2013
- **Test Set:** 2014

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

Results

Model	Predictors	Indices	Test Set 1 MSE	Test Set 1 MAE	Test Set 2 MSE	Test Set 2 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	NA	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

- **Test Set 1:** Three months (June, July and August 2014)
- **Test Set 2:** One month (June 2014)

Forecasting Solar Intensity

Variables

- **Response:** **Daily solar intensity** – February 2006 to February 2013 – Amherst, Massachusetts
- **Index Variables:**
  - Solar intensity lags
  - Temperature, dew point, wind speed, rain and humidity lags
- **Linear:** DOY fourier terms

Data Split

- **Training Set:** February 2006 to October 2012
- **Validation Set:** November and December 2012
- **Test Set:** January and February 2013

$$\textbf{Solar} = \beta_0 + \sum_{j=1}^{p} g_j(\boldsymbol{X}\boldsymbol{\alpha}_j) + \sum_{k=1}^{8}(\theta_k \textbf{DOY\_S}_{\boldsymbol{k}} + \delta_k \textbf{DOY\_C}_{\boldsymbol{k}}) + \varepsilon,$$

Results

Model	Predictors	Indices	Test Set MSE	Test Set MAE
SMI Model (1, 0) - PPR	39	4	764.087	21.916
SMI Model (6, 0) - Additive	39	23	981.033	24.617
SMI Model (1, 0) - Linear	16	0	2065.762	34.105
Backward Elimination	36	NA	819.429	22.998
GAIM	39	6	1972.777	37.130
PPR	23	6	**723.067**	**21.731**

Outline

Conclusion

Key features:

- Automatic selection of number of indices and predictor grouping
- Automatic predictor selection
- A wide spectrum: from single index models to additive models
- Flexibility to include separate nonlinear and linear predictors

Conclusion

Key features:

- Automatic selection of number of indices and predictor grouping
- Automatic predictor selection
- A wide spectrum: from single index models to additive models
- Flexibility to include separate nonlinear and linear predictors

Things to improve:

- Initialisation: we encourage trial-and-error
- Applicability: more applications are needed
- Computational time: increases with number of predictors and indices