Constrained groupwise additive index models

Pierre Masselot[1*], Fateh Chebana[2], Céline Campagna[2,3], Éric Lavigne[4,5], Taha B.M.J. Ouarda[2], Pierre Gosselin[2,3,6]

[1]*London School of Hygiene & Tropical Medicine, London, United Kingdom;*

[2]*Institut National de la Recherche Scientifique, Centre Eau-Terre-Environnement, Québec, Canada;*

[3]*Institut National de Santé Publique du Québec, Québec, Canada;*

[4]*School of Epidemiology and Public Health, University of Ottawa, Ottawa, Canada;*

[5]*Air Health Science Division, Health Canada, Ottawa, Canada;*

[6]*Ouranos, Montréal, Canada.*

**Corresponding Author: pierre.masselot@lshtm.ac.uk*

Summary

In environmental epidemiology, there is wide interest in creating and using comprehensive indices that can summarise information from different environmental exposures while retaining strong predictive power on a target health outcome. In this context, the present paper proposes a model called constrained groupwise additive index model (CGAIM) to create easy-to-interpret indices predictive of a response variable, from a potentially large list of variables. The CGAIM considers groups of predictors that naturally belong together to yield meaningful indices. It also allows the addition of linear constraints on both the index weights and the form of their relationship with response variable to represent priori assumptions or operational requirements. We propose an efficient algorithm to estimate the CGAIM, along with index selection and inference procedures. A simulation study shows that the proposed algorithm has good estimation performances, with low bias and variance and is applicable in complex situations with many corelated predictors. It also demonstrates important sensitivity and specificity in index selection, but non-negligible coverage error on constructed confidence intervals. The CGAIM is then illustrated on the construction of heat indices in a health warning system context. We believe the CGAIM could become useful in a wide variety of situations, such as warning systems establishment, and multi-pollutant or exposome studies.

Keywords: dimension reduction; index; additive index models; linear constraints; quadratic programming.

1. Introduction

In environmental epidemiology, there is an increasing recognition of the complexity of the mixture of exposure on human health. The number of exposures, be it pollutants, weather variables or built environment characteristics can be numerous, and interact in a complex fashion to impact human health. The ever-increasing amount of data available allows for complex statistical and machine learning models to be considered. For instance, recent advances such as Bayesian kernel machine regression (Bobb *and others* 2015) , random forests (Breiman 2001) or more generally many nonparametric algorithms can achieve impressive predictive power. However, the mentioned models are often referred to as 'black-boxes' (a recent example is Schmidt 2020) and are challenging to interpret in practice (e.g. Davalos *and others* 2017).

Interpretability of model outputs may be a key component of many real-world applications, especially when they involve decision making or risk assessment (Rudin 2019). Public health scientists or decision makers need clear and easy-to-interpret insights about how the different exposures may impact the given health outcome. Examples include weather-related factors (Chebana *and others* 2013; Pappenberger *and others* 2015) or air quality indices (Lee *and others* 2011; Masselot *and others* 2019). The pool of methods used to create indices is currently limited, as many indices are constructed based on previously estimated univariate risks or created based on a literature review (e.g. Monforte and Ragusa 2018). Another type of studies seeking to summarize large amount of information, exposome-wide association studies (EWAS), usually focus on linear methods selecting a few number of exposure, thus partly discarding the

complexity of exposure mixture (e.g. Nieuwenhuijsen *and others* 2019). Other studies consider index-based methods through the popular weighted quantile sum regression, that still relate the created index linearly to the response variable (Keil *and others* 2020). Therefore, there is a need for methods able to account for complex mixtures of many variables and provide interpretable indices.

Starting from a pool of exposures $\boldsymbol{X} \in \mathbb{R}^d$, indices are defined as a small number $p < d$ of custom predictors Z that are linear combinations of the original predictors, i.e. of the form $Z = \boldsymbol{\alpha}^T \boldsymbol{X}$. In this sense, deriving indices Z can be seen as a dimension reduction problem. The most famous example of a dimension reduction method is principal component analysis (PCA, Jolliffe 2002). However, in the present work, we are especially interested in a regression context, i.e. in deriving indices related to a response of interest. Methods that are suited for this objective include the single-index model (SIM, e.g. Härdle *and others* 1993) in which one index is constructed through the model $Y = g(\boldsymbol{\alpha}^T \boldsymbol{X}) + \epsilon$ or the projection pursuit regression (PPR, Friedman and Stuetzle 1981), also known as the additive-index model, which extends the SIM in the following fashion: $Y = \sum_j g_j(\boldsymbol{\alpha}_j^T \boldsymbol{X}) + \epsilon$. In both models, g and g_j are nonlinear functions representing the relationship between the response Y and the constructed index $Z_j = \boldsymbol{\alpha}_j^T \boldsymbol{X}$.

Although the SIM and PPR models are often used as nonparametric regression models (Wang and Ni 2008; Yuan *and others* 2016; Durocher *and others* 2016; Cui *and others* 2017), their very general and flexible nature results in a lack of interpretability as well as a tendency to overfit the data (Zhang *and others* 2008). The main reasons are: i) the derived indices include all predictors $\boldsymbol{X}$, hence mixing very different variables for which a linear combination makes little sense, ii) the

very general vectors $\boldsymbol{\alpha}_j^T$ do not guarantee interpretability and iii) the flexibility of functions g_j may result in complex functions preventing a clear interpretation of the corresponding index Z_j.

Usually, the predictors at hand can naturally be grouped into variables representing phenomena that jointly impact the response Y. For instance, grouped variables can be naturally interacting variables such as several weather, air pollutants, sociodemographic variables as well as lagged variables (Xia and Tong 2006). Several authors proposed to take advantage of such groupings as a path to improve the interpretability of the derived indices (Li *and others* 2010; Guo *and others* 2015). This leads to the groupwise additive index model (GAIM) expressed as:

$$Y = \sum_{j=1}^{p} g_j(\boldsymbol{\alpha}_j^T \boldsymbol{X}_j) + \epsilon \tag{1}$$

where the $\boldsymbol{X}_j \in \mathbb{R}^{l_j}$ $(j = 1, \dots, p)$ are subsets of variables of $\boldsymbol{X}$, i.e. $l_j < d$. The GAIM in (1) allows deriving more meaningful indices $Z_j = \boldsymbol{\alpha}_j^T \boldsymbol{X}_j$ since they are built from subsets of predictors that logically or naturally belong together. It can be seen as a sparser model since only a subset of variables enters a term in (1), noting that sparsity is a key aspect of interpretability (Rudin 2019).

Although the GAIM in (1) allows an improvement in the indices interpretability, its flexibility can still result in physically or practically incoherent indices. Thus, it is also of interest to be able to constraint the indices weights $\boldsymbol{\alpha}_j$ to yield more meaningful indices. Constraints on the weights $\boldsymbol{\alpha}_j$ can represent additional information included in the model and reflect the expertise of knowledge specific to a given application context or operational requirements for the created indices. For an index to be useful in practice, it is also highly desirable that it relates to the response Y in an easy-to-interpret way. For an air quality index, it is reasonable to expect g_j to be

monotonically increasing. Similarly, a temperature-related index may impose a convexity constraint on g_j, acknowledging a minimum-mortality temperature and increased risks on both sides. A too flexible model for the function g_j might however give implausible or difficult to interpret indices and therefore limit their usefulness for decision making. This means that it is also of interest to impose constraints on the shape of the functions g_j.

In the present paper, we propose a constrained GAIM (CGAIM) as a general model that includes all the constraints discussed above. It is a model of the form (1) in which constraints are added on the weights $\boldsymbol{\alpha}_j$ as well as on the functions g_j depending on the application. Several authors proposed unconstrained GAIMs based on local linear estimation (Li *and others* 2010; Wang *and others* 2015; Guo *and others* 2015; Wang and Lin 2017). Fawzi et al. (2016) proposed the addition of a few constraints on the weights α_j but not on the functions g_j. Chen and Samworth (2015) proposed a PPR with shape-constrained functions g_j, but it is not in a groupwise context. Xia and Tong (2006) and then Kong et al. (2010) proposed a GAIM with constraints on both the weights $\boldsymbol{\alpha}_j$ and functions g_j, but limited to monotonicity and without the possibility to add additional covariates such as confounders. Finally, these methods all lack inference procedures to provide uncertainty assessment or test for specific indices of covariates. Such inference results are important for interpretation purposes. We propose here a general model that encompasses all mentioned ones, with the addition of an efficient estimation procedure, as well as index selection and inference.

2. The constrained groupwise additive index model

In order to present the proposed CGAIM, we rewrite and extend model (1) as:

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j g_j(\boldsymbol{\alpha}_j^T \boldsymbol{X}_j) + \sum_{k=1}^{d} \gamma_k f_k(W_k) + \boldsymbol{U}^T \boldsymbol{\theta} + \epsilon \quad (2)$$

where $\boldsymbol{X}_j \in \mathbb{R}^{l_j}$ $(j = 1, \ldots, p)$ are subsets of all the variables in $\boldsymbol{X}$, $\boldsymbol{\alpha}_j$ is a vector of weights and g_j is a nonlinear function. The coefficients β_j represent the relative importance of each index $Z_j = \boldsymbol{\alpha}_j^T \boldsymbol{X}_j$ in predicting the response Y. The constant β_0 is the intercept of the model.

The W_k and U (with dimension ≥ 1) represent additional covariates that are related to Y but not entering any index. The formers are nonlinearly related to Y through f_k with importance γ_k, which are respective counterparts to g_j and β_j, and the latter are linear. The typical example is confounding variables in environmental epidemiology such as day-of-week or time covariates.

One of the key features of the proposed CGAIM is to allow for any linear constraint on the weights $\boldsymbol{\alpha}_j$, i.e. constraints of the form $\boldsymbol{C}_j \boldsymbol{\alpha}_j \geq 0$ where $\boldsymbol{C}_j$ is a $m_j \times l_j$ matrix, m_j being the number of constraints and l_j the number of variables in the group $\boldsymbol{X}_j$. Linear constraints allow for a large array of constraints. Examples include forcing some or all of the weights in $\boldsymbol{\alpha}_j$ being positive, in which case $\boldsymbol{C}_j$ is the identity matrix, and forcing them to be monotonically decreasing, in which case $\boldsymbol{C}_j$ is an $(l_j - 1) \times l_j$ matrix where $C_{j,pq} = 1$ when $p = q$, $C_{j,pq} = -1$ when $p = q - 1$ and 0 otherwise.

The other key feature of the CGAIM is the possibility to add shape constraints on the functions g_j and f_k. Shape constraints include monotonicity, convexity, concavity and combinations of the former (Pya and Wood 2015). Note that not all functions g_j and f_k need to be constrained or have the same shape constraint.

For identifiability, we assume that the grouping is chosen before model fitting and that no predictor variable enters two indices, i.e. $\boldsymbol{X}_j \cap \boldsymbol{X}_k = \emptyset, \forall j, k$. Regarding the weights $\boldsymbol{\alpha}_j$, identifiability can be ensured by the classical unit norm constraint $\|\boldsymbol{\alpha}_j\| = 1$ with the first element of $\boldsymbol{\alpha}_j$ being positive (Yu and Ruppert 2002; Yuan 2011). However, we can also take advantage of linear constraints to ensure both identifiability and a better interpretability of the resulting indices. For instance, the constraints $\sum_{k=1}^{l_j} \boldsymbol{\alpha}_{jk} = 1$ and $\boldsymbol{\alpha}_{j1} \geq 0$, which represents a weighted average of the variables in $\boldsymbol{X}_j$, are enough to ensure identifiability of $\boldsymbol{\alpha}_j$. As estimation of g_js, f_k and θ for fixed $\boldsymbol{\alpha}_j$ is a generalized additive model (GAM), we consider the classical centering identifiability constraints (Wood 2004; Yuan 2011). Finally, since we allow linear covariates in the model, we assume that no function g_j is linear since it could cause identifiability issues in the groupwise context (a formal proof is provided by Fawzi *and others* 2016).

3. Estimating the CGAIM

In this section, we present an estimation algorithm for the CGAIM based on the general framework of SQP. We first focus on the additive index part of the model for clarity purposes and then extend the estimation to the full model in (2). We also present a generalized cross-validation criterion for model selection and two inference procedures.

3.1. Estimation problem

To fit the CGAIM, given observed data $(y_i, x_{i1}, \dots, x_{id})$, where $i = 1, \dots, n$ and the d predictor variables are partitioned into p groups, we seek to minimize the squared error over coefficients β_0 and β_j, functions g_j and weight vectors $\boldsymbol{\alpha}_j$, $j = 1, \dots, p$, i.e.:

$$\begin{aligned} \min \quad & \sum_{i=1}^{n}\left[y_i - \beta_0 - \sum_{j=1}^{p}\beta_j g_j\left(\boldsymbol{\alpha}_j^T \boldsymbol{x}_{ij}\right)\right]^2 \\ \text{subject to} \quad & \boldsymbol{C\alpha} \geq 0 \\ & g_j \in m \end{aligned} \tag{3}$$

where $\boldsymbol{\alpha} = \left[\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_p^T\right]^T$ and m is one of the shape constraints available for g_j.

Since the $\boldsymbol{\alpha}_j$s do not enter linearly in the squared error (3), this is a nonlinear least squares problem which suggests an approach such as a Gauss-Newton algorithm. However, an additional difficulty arises from the constraints of the model, especially those on the $\boldsymbol{\alpha}_j$s. It is thus appropriate to consider SQP steps, a general algorithm for nonlinear constrained optimization problems (Boggs and Tolle 1995). It has been shown to work well in the context of nonlinear least squares (Schittkowski 1988).

The proposed estimation methods for related models listed in the introduction (Xia and Tong 2006; Li *and others* 2010; Wang *and others* 2015) are all based on local regression to minimize the sum of squares (3). However, it can be computationally intensive and makes the inclusion of constraints more difficult due to the high number of local coefficients to estimate. Here we rather choose an approach based on splines for the function g_j and SQP iterations for the weights $\boldsymbol{\alpha}_j$. Note that smoothing splines were shown to have good performances in a PPR context (Roosen and Hastie 1994).

3.2. Estimation algorithm

Since the minimisation problem (3) is a separable one (Golub and Pereyra 2003), we propose here to estimate the GAIM with an algorithm that iteratively updates the functions g_j and the weight vectors $\boldsymbol{\alpha}_j$. In the first step, with the $\boldsymbol{\alpha}_j$s fixed, we can derive indices values $z_{ij} = \boldsymbol{\alpha}_j^T \boldsymbol{x}_{ij}$. Estimating the functions g_j is thus equivalent to estimating a generalized additive model (GAM,

Hastie and Tibshirani 1986) using the current z_{ij} as predictors. In such a model, g_j can be efficiently estimated by smoothing splines as detailed by Wood (2017). After estimating the functions g_j, they are scaled to have norm one, and the coefficients β_j are adjusted accordingly.

When shape constraints are considered, different corresponding methods can be considered. Pya and Wood (2015) proposed the shape-constrained additive models (SCAM), that estimates reparametrized P-spline coefficients through an iterative reweighted least-squares like algorithm. Meyer (2018) proposed a constrained GAM (CGAM) that uses integrated and convex splines (Ramsay 1988; Meyer 2008) with quadratic programming to enforce shape constraints. Finally, Chen and Samworth (2015) proposed the shape-constrained additive regression that estimates non-smooth shape-constrained functions through maximum-likelihood. All these methods allow for monotonicity, convexity and concavity constraints. Throughout the present paper, we consider SCAM as it allows for a more flexible management of functions g_j smoothness.

In the second step, with the functions g_j estimated, we can update the weights $\boldsymbol{\alpha}_j$ by minimizing the sum of squares function (3) over the $\boldsymbol{\alpha}_j$ only. Let $\boldsymbol{\alpha}^{old}$ be the current value of $\boldsymbol{\alpha} = \left[\boldsymbol{\alpha}_1^T, \dots, \boldsymbol{\alpha}_p^T\right]^T$ and $\boldsymbol{\alpha}^{new}$ the next value to be computed. The update $\delta = \boldsymbol{\alpha}^{new} - \boldsymbol{\alpha}^{old}$ can be conveniently computed by a quadratic program (QP) of the form

$$\begin{array}{ll} \min & \boldsymbol{\delta}^T \boldsymbol{V}^T \boldsymbol{V} \boldsymbol{\delta} - 2\boldsymbol{V}^T \boldsymbol{R} \boldsymbol{\delta} \\ \text{subject to} & \boldsymbol{C}\boldsymbol{\delta} + \boldsymbol{C}\boldsymbol{\alpha}^{old} \geq 0 \end{array} \tag{4}$$

in which $\boldsymbol{V}$ is the matrix containing the partial derivative according to the $\boldsymbol{\alpha}_j$ of the CGAIM equation, i.e., the right-hand side of (2). $\boldsymbol{V}$ contains $\left[\boldsymbol{v}_{i1}, \dots, \boldsymbol{v}_{ip}\right]$ at line i with the vector $\boldsymbol{v}_{ij} = \boldsymbol{x}_{ij}\beta_j g_j'(z_{ij})$. $\boldsymbol{R}$ is the current residual vector that contains $r_i = y_i - \beta_0 - \sum_{j=1}^{p} \beta_j g_j\left(\boldsymbol{\alpha}_j^{old^T} \boldsymbol{x}_{ij}\right)$. The objective function in (4) is a quasi-Newton step in which the Hessian part that involves the

second derivatives of the CGAIM has been discarded to avoid its computational burden, leaving only the term $\boldsymbol{V}^T\boldsymbol{V}$. Thus, the update $\boldsymbol{\delta}$ is guaranteed to be in a descent direction. Discarding the second derivative of the model is a distinctive feature of least squares since it is usually negligible compared to the term $\boldsymbol{V}^T\boldsymbol{V}$ (Hansen *and others* 2013). Note that this is especially true here since both the use of smoothing spline and shape-constraints for $g_j(.)$ results in smooth functions and thus low second derivatives $g_j''(.)$. Finally, the constraints in (4) ensure that the updated vector $\boldsymbol{\alpha}^{new} = \boldsymbol{\alpha}^{old} + \boldsymbol{\delta}$ is still in the feasible region. Note that without these constraints, the problem in (4) reduces to a classical Gauss-Newton step for nonlinear least-squares (Bates and Watts 1988).

The algorithm alternates updating the weights $\boldsymbol{\alpha}_j$ and estimating the functions g_j with the current $\boldsymbol{\alpha}_j$ until convergence. Convergence is usually reached when the least squares function (3) does not evolve anymore after updating the g_j and $\boldsymbol{\alpha}_j$. Note that we can also consider other criteria for convergence such as stopping when the update $\boldsymbol{\delta}$ is very small or the orthogonality convergence criterion of Bates and Watts (1981).

To start the algorithm, a constrained linear regression of the $[\boldsymbol{x}_{i1}, \dots, \boldsymbol{x}_{ip}]$ on y_i should provide an initial guess $\boldsymbol{\alpha}_0 = [\boldsymbol{\alpha}_1^T, \dots, \boldsymbol{\alpha}_p^T]$ close to the optimal solution (Wang *and others* 2015). This constrained linear regression can be implemented as a QP of the form (4) by replacing $\boldsymbol{V}$ with the design matrix $\boldsymbol{X}$, and $\boldsymbol{R}$ with the response vector $\boldsymbol{Y}$. Alternatively, $\boldsymbol{\alpha}_0$ can be initiated randomly, using constrained random number generators (Van den Meersche *and others* 2009). Key steps of the estimation procedure are summarized in Algorithm 1.

Algorithm 1: Constrained GAIM estimation

0. Initialize $\boldsymbol{\alpha} = \left[\boldsymbol{\alpha}_1^T, \dots, \boldsymbol{\alpha}_p^T\right]$ either by a QP as in (4) or randomly.
1. Functions g_j update:
    a. Estimate the g_j by a SCAM with y_i as the response and the $z_{ij} = \boldsymbol{\alpha}_j^T \boldsymbol{x}_{ij}$ as predictor.
    b. Scale the estimated g_j to have unit norm and adjust the coefficients β_j consequently.
2. Weights $\boldsymbol{\alpha}_j$ update:
    a. Compute the update $\boldsymbol{\delta} = \left[\boldsymbol{\delta}_1, \dots, \boldsymbol{\delta}_p\right]$ through the QP (4).
    b. Set $\boldsymbol{\alpha} = \boldsymbol{\alpha} + \boldsymbol{\delta}$.
    c. Scale each $\boldsymbol{\alpha}_j$ to have unit norm.
3. Iterate steps 1 and 2 until convergence.

3.3. Additional covariates

The integration of the covariates W_k and U in the estimation procedure is straightforward since they only intervene in the update of functions g_j (step 1 of algorithm 1). In this step, they are simply added as covariates in the SCAM (or GAM in the unconstrained case), along the current indices Z_j. Shape constraints can be applied on the functions f_k as well. These terms do not intervene in the $\boldsymbol{\alpha}_j$ update step, since they are considered constants with respect to $\boldsymbol{\alpha}_j$, which mean that they disappear from the derivative matrix $\boldsymbol{V}$. Finally, note that the coefficients γ_k are obtained as the norm of functions f_j.

3.4. Model selection

As the number of indices and covariates grow in the model, it is of interest to select a subset that are the most predictive of the response Y. To this end, we propose here a generalized cross-validation (GCV) type criterion of the form (Golub *and others* 1979):

$$GCV = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/n}{(1 - edf/n)^2} \tag{5}$$

where the numerator represents the residual error with $\hat{y}_i$ the fitted value from the CGAIM, and the denominator is a penalization that depends on the effective degrees of freedom (edf). For a large number of indices p, we can perform the selection as a forward stepwise algorithm in which, at each step, the index minimizing the GCV is added to the model.

When a model can be reformulated linearly, the edf term in (5) can be estimated as the trace of the hat matrix, but it is not the case here. Instead, we consider a similar approximation as proposed by Roosen and Hastie (1994) for PPR, i.e.

$$edf = p + d + \sum\left(edf_g\right) + \sum(edf_\alpha) \tag{6}$$

where $p + d$ charge one degree of freedom per index and covariate for the coefficients β_j and γ_k, $\sum\left(edf_g\right)$ represent the sum of edfs for each ridge function smoothing, and $\sum(edf_\alpha)$ is the sum of edfs for each index weight vector estimation. Estimation of edf_g is well described in Meyer and Woodroofe (2000) and corresponds to the number of basis functions used in the smooth, to which we subtract the number of active constraints multiplied by a constant usually specified at $c \approx 1.5$ to account for the smoothing penalization (see also Meyer 2018). Similarly, edf_α can be estimated as the number of coefficients to which we subtract the number of active constraints (Zhou and Lange 2013).

3.5. Inference

Inference for the ridge functions g_j is well described elsewhere (Pya and Wood 2015; Meyer 2018) and inference for the coefficients β_j is straightforward as they can be treated like regular regression coefficients using the $g_j(\boldsymbol{\alpha}_j^T \boldsymbol{x}_{ij})$ as predictors. Here, we describe inference for the vector of weights $\boldsymbol{\alpha} = \left[\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_p^T\right]^T$ only. If one assumes normality of the residuals, then the transformed vector $\boldsymbol{\xi} = \boldsymbol{C}(\boldsymbol{\alpha} - \widehat{\boldsymbol{\alpha}})$ follows a truncated multivariate normal with null mean, covariance matrix $\boldsymbol{C}\boldsymbol{\Sigma}_\alpha\boldsymbol{C}^T$ where $\boldsymbol{\Sigma}_\alpha$ is the covariance matrix of $\boldsymbol{\alpha}$ for an unconstrained model, and lower bound $\boldsymbol{C}(\boldsymbol{b} - \widehat{\boldsymbol{\alpha}})$ (Geweke 1996). We can efficiently simulate a large number of vectors $\boldsymbol{\xi}^*$ from this truncated multivariate normal, and back-transform them as $\boldsymbol{\alpha}^* = \widehat{\boldsymbol{\alpha}} + \boldsymbol{C}^{-1}\boldsymbol{\xi}^*$ to obtain an estimate of the distribution of the vector $\boldsymbol{\alpha}$ (Botev 2017). Empirical confidence intervals or other inference can then be obtained from the simulated $\boldsymbol{\alpha}^*$.

The unconstrained covariance matrix $\boldsymbol{\Sigma}_\alpha$ can be obtained through the classical nonlinear least-squares approximation $\boldsymbol{\Sigma}_\alpha = s^2(\boldsymbol{V}^T\boldsymbol{V})^{-1}$ where s^2 is an estimate of the residual variance of the model (Bates and Watts 1988). In this instance, s^2 should be estimated using the effective degrees of freedom formula devised in section 3.4. Note also that since it needs to be inverted, the constraint matrix $\boldsymbol{C}$ should be a square matrix. If this is not the case, it can be augmented by a matrix $\boldsymbol{C}_0$ spanning the row null space of $\boldsymbol{C}$ while the vector $\boldsymbol{b}$ is augmented with $-\infty$ (Tallis 1965).

Without the normality assumption, inference and confidence intervals can be obtained through a bootstrap procedure (DiCiccio and Efron 1996), with the following procedure. We start by extracting the residuals $\hat{\epsilon}_i$ of the CGAIM fit. We then draw from the $\hat{\epsilon}_i$ with replacement to obtain a new sample ϵ_i^* that is then added to the fitted values to obtain a new response vector

$y_i^* = \hat{y}_i + \epsilon_i^*$ on which the CGAIM can be fitted (Efron and Tibshirani 1993). We repeat this a large number B of times to obtain a bootstrap distribution of any parameter from the CGAIM, including the weights $\boldsymbol{\alpha}_j$, the ridge functions g_j and the coefficients β_j.

4. Simulation study

In this section, we analyze the performances of the CGAIM on different types of simulated data. We test the ability of the proposed CGAIM to estimate accurately weights α_j, by comparing it with other methods, its ability to find the most relevant predictors in the context of an important number of exposures, the ability of the GCV criterion to find the correct model and the coverage of the confidence intervals applied as described above.

4.1. Index estimation

In this setting, three predictor matrices are generated following a multivariate normal distribution of dimension $p_j = 4$ ($j = 1,2,3$), with null means and covariance matrices having unit diagonal and non-diagonal elements equal to a predefined ρ value. The first index is composed of sharply decreasing weights with a log function g_1 to emulate the effect of air pollution on mortality. The second includes moving average weights with a sigmoid function g_2 that represent a soft threshold on the index typical of logistic models. The third index represents a classical mortality-temperature relationship with weights representing a delayed impact and a U-shaped relationship. The linear predictor is then the sum of the three ridge functions, i.e. with magnitudes $\beta_j = 1$ for $j = 1,2,3$ and an intercept $\beta_0 = 5$. A large number $n_s = 1000$ datasets are generated by adding gaussian white noise to the linear predictor described here. More details on the simulation setup are given in Supplementary Materials.

From the basic mechanism described above, various scenarios are implemented. In these scenarios we change the sample size of simulated data with $n = 100, 200, 500, 1000$, the correlation between predictor variables with non-diagonal elements of the covariance matrix in $\rho = 0, 0.25, 0.50, 0.75$, and the noise level with standard deviations in $\sigma = 0.2, 0.5, 1$. The unconstrained GAIM and the CGAIM are applied on each of the generated datasets. The CGAIM is applied with the constraints that all weights are positives, that $\boldsymbol{\alpha}_1$ is increasing and $\boldsymbol{\alpha}_2$ decreasing. The functions g_1 and g_2 are constrained to be both monotonically increasing and g_3 is constrained to be convex. The specific constraints applied to each index are summarized in Supplementary Table 1. The GAIM is only applied with identifiability constraints, i.e. that non-negativity of the first element of each weight vector $\boldsymbol{\alpha}_{j1} \geq 0$ and unit norm for $\boldsymbol{\alpha}_j$. To test the model with wrongly specified constraints, we fit a mis-specified model (MGAIM) constraining $\boldsymbol{\alpha}_1$ to be decreasing and $\boldsymbol{\alpha}_2$ increasing. For CGAIM and MGAIM, we fix the smoothness of g_j to an equivalent of 10 degrees of freedom. This avoids the computational burden of smoothness optimization in SCAM, while keeping enough flexibility for model fitting.

Besides the three models described above, three benchmark models are applied on the generated datasets. The first one is the PPR as the most general additive index model available. Comparing the (unconstrained) GAIM to the PPR allows assessing the benefits of defining groups of variables *a priori*. The second benchmark is the groupwise minimum average variance estimator (gMAVE) of Li et al. (2010), as representative of groupwise dimension reduction methods. It allows the evaluation of the estimation method without constraint. Finally, we also apply the functional additive cumulative time series (FACTS) model of Kong et al. (2010), that contains a groupwise additive structure and monotonicity constraints on both index weights $\boldsymbol{\alpha}_j$ and ridge functions g_j. We only apply the monotonicity constraints applied to CGAIM to FACTS, as its

extension to other type of constraints in not trivial. The performances are evaluated by comparing the estimated weights $\widehat{\boldsymbol{\alpha}}_j$ to the true values $\boldsymbol{\alpha}_j$. The quality of estimated $\widehat{\boldsymbol{\alpha}}_j$ are evaluated using the classical root mean squared errors (RMSE) that aggregates information about both the bias and standard error of the estimators.

Figure 1 shows the RMSE for each model for different sample sizes, correlation coefficient between the predictor variables, and noise levels. There is overall a clear hierarchy between the compared models, with the GAIM and CGAIM having the lowest errors, the gMAVE having slightly higher errors and being more sensitive to the sample size, and then the FACTS. PPR have overall much higher errors being in addition extremely variable . The methods based on the proposed algorithm on the other hand, show important stability with robustness to variation in all explored parameters. As expected, the MGAIM shows low performances because of the mis-specified constraints preventing the model to converge to the true $\boldsymbol{\alpha}_j$. Note however that it displays very low variance, as it converges towards the best solution within the feasible region (see Supplementary Materials).

4.2. Index selection

In this experiment, we evaluate the ability of the GCV criterion (5) to retrieve the correct model. We consider the structure detailed in the previous experiment with $n = 1000$ and $\rho = 0$, as well as three noise levels $\sigma = 0.2, 0.5$ and 1. For each realisation, we randomly select p^* indices $J^* \subset \{1,2,3\}$ and attribute them a unit coefficient $\beta_{j\in J^*} = 1$. We attribute $\beta_{j\notin J^*} = 0$ to the remaining indices, thus discarding them from the generated model. At each realization, we choose the best model by GCV and compute the sensitivity and specificity. Sensitivity is defined as the

proportion of indices in j^* that are in the model selected by GCV, and specificity the proportion of indices not in j^* that are discarded by the GCV.

Figure 2 shows the average sensitivity and specificity on $n_s = 1000$ realizations for the two number of non-null indices and various noise levels. Sensitivity is equal to one in all simulations, meaning that the GCV always selects the true indices in the model. Specificity indicates that the GCV select only the true indices most of the time, i.e. around 95 % of the time for the CGAIM and around 80% of the time for GAIM. The GAIM might then be prone to slight overfitting, while the constraints in the CGAIM allow achieving more parsimonious models. However, specificity is still high in all cases and is not sensitive to the noise level. The proposed GCV criterion is therefore mostly successful for model selection.

4.3. Coverage

To evaluate the inference procedures proposed for the CGAIM, we perform simulations to assess the coverage achieved by confidence intervals for the $\boldsymbol{\alpha}_j$ weights. We generate datasets following the same mechanism described above, with $n = 1000$ and $\rho = 0$, as well as three noise levels $\sigma = 0.2, 0.5$ and 1. We then fit a CGAIM model as in the two first experiments and estimate its 95% confidence interval using both the normal approximation and residual bootstrap. In both cases, the number of created samples is fixed to $B = 500$. As the constraints can create some bias, especially for coefficients involved in active constraints, we compute the bias-corrected coverage, as the proportion of confidence intervals containing the average estimated values $\hat{\alpha}_j$ from the simulations (Morris *and others* 2019).

Figure 3 shows the bias-corrected coverage for both method and the three noise levels. The residual bootstrap shows constant reasonable coverage with values around 96% for all noise

levels. In contrast, the normal approximation method is widely affected by the noise level and shows important coverage errors, with major underestimation for the highest noise level. This coverage error can also significantly vary between the various $\boldsymbol{\alpha}_j$ with low coverages for $\boldsymbol{\alpha}_3$ specifically, while the variation between indices is lesser for the residual bootstrap (see Supplementary Figure 4).

4.4. Exposome

In this experiment we depart from the structure of the previous experiments and apply the GAIM and CGAIM to estimate the most important predictors in a simulation study typical of exposome studies. We modify the simulation study proposed by Agier *and others* (2016), using the structure of the HELIX cohort (Robinson *and others* 2018). We generate a matrix of $d = 28$ predictors with $n = 1200$ subjects from the correlation matrix provided in Robinson et al. (2018). In each realization, we select $p^* = 5{,}10{,}15$ predictors $K^* \subset \{1, \dots, 28\}$ to have non-null weights, while the remaining predictors are attributed null weights, and have therefore no association with the response. We then generate the response vector through the model $Y^* = \sum_{k \in K^*} \alpha_k X_k + \epsilon$ where α_k is either -1 or 1 with equal probability to evaluate the ability of the model estimate the direction of the association. Response vectors are then generated such that the R^2 of the model is $3p^*/100$ (Agier *and others* 2016).

As the correlation matrix used to generate the predictors X_k represents environmental stressors, five groups naturally arise (Nieuwenhuijsen *and others* 2019): climatic ($l_1 = 4$), air pollution ($l_2 = 5$), traffic-related ($l_3 = 4$), natural environment ($l_4 = 5$) and built environment ($l_5 = 10$) variables. We apply the (unconstrained) GAIM and CGAIM on 1000 realisations of the above-described mechanism with these groups of variables. The CGAIM is applied with the constraints

$|\alpha_k| \leq 1 \ \forall k \in \{1, \ldots, 28\}$, convexity constraint on g_1 (representing the effect of climate) and increasing monotonicity on other g_j ($j \neq 1$). We then compare the estimated $\hat{\alpha}_k$ ($k \in \{1, \ldots, 28\}$) to the true value α_k.

Figure 4 shows that the $\hat{\alpha}_k$ are on average close to the true α_k, successfully discriminating null weights, but also the direction of non-null weights. They are closer to the true value for the CGAIM compared to GAIM, with also much lower variability in the estimated weights. The difference between estimated and true weights also slightly decreases with the number of non-null weights α_k. Therefore, the CGAIM, performs well with many predictors and complex correlation patterns, especially when constraints are considered.

5. Application

This section presents an example of application in environmental epidemiology in which the CGAIM is used to construct multiple indices representing heat-related mortality risks. A second application on air pollution is presented in Supplementary materials. This application considers daily mortality and exposure data spanning the months of June-August for the period 1990 – 2014 ($n = 2300$) from the Metropolitan Area of Montreal in the province of Quebec, Canada, which are described in detail in, e.g., Masselot et al. (2018, 2019). Briefly, daily all-cause mortality data are provided by the province of Quebec National Institute of Public Health, while daily temperature and humidity data are extracted from the 1x1 km gridded dataset DayMet (Thornton *and others* 2020).

We apply the CGAIM to find optimal weights for temperature indices that represent potentially adverse effects. Indices created include lagged averages of $Tmin$ and $Tmax$, following the current indices in Montreal (Chebana *and others* 2013) and we also include the vapor pressure

(Vp) variable to represent humidity, since it is also sometimes considered a determinant of summer mortality (e.g. Barreca 2012). The objective here is to give an example of the possibilities offered by the CGAIM. Thus, to estimate these indices, the following full model is considered:

$$Y = \beta_0 + \beta_1 g_1(\boldsymbol{\alpha}_1^T Tmin_{3d}) + \beta_2 g_2(\boldsymbol{\alpha}_2^T Tmax_{3d}) + \beta_3 g_3(\boldsymbol{\alpha}_3^T Vp_{3d}) + \gamma_1 f_1(DOS) + \gamma_2 f_2(year) + \epsilon \quad (7)$$

where Y is the all-cause daily mortality, $Tmin_{3d}$, $Tmax_{3d}$ and Vp_{3d} represent matrices of lags 0, 1 and 2 days of corresponding variables, meaning that the $\boldsymbol{\alpha}_j$ $(j = 1, \dots, 3)$ are vectors of length 3. The two additional covariates are the day-of-season (DOS) and year variables to control for the seasonality, inter-annual trend and residual autocorrelation as commonly done in time series study in environmental epidemiology (Bhaskaran *and others* 2013).

We consider a CGAIM and an (unconstrained) GAIM. The CGAIM model includes constraints for positive and decreasing weights with the lag, i.e. $\alpha_{j0} \geq \alpha_{j1} \geq \alpha_{j2} \geq 0, \forall j$. This is encoded by the following constraint matrix

$$\boldsymbol{C}_j = \begin{bmatrix} 0 & 0 & 1 \\ 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix} \quad (8)$$

For the indices to directly represent a measure of heat risk, and because the data are restricted to the hottest months of the year with little exposure to cold, we add the constraint that the relationship g_j is monotone increasing for all j. As in the simulation study, we fix the smoothness to the equivalent of 10 degrees of freedom. Confidence intervals are computed through the residual bootstrap.

For both the CGAIM and GAIM, we use the GCV criterion (5) to determine the best set of indices to predict summer mortality. Bets models include $Tmax$ and Vp for both the GAIM and CGAIM. Among these two best models, the GCV of the CGAIM one is slightly lower being at 90.2 compared to 90.6 for the GAIM.

The indices and their association with mortality are shown in Figure 5. The CGAIM attributes a slightly decreasing weights with lags of $Tmax$ resulting in an index that has a large association with mortality for extreme values, especially above the value 0.9 of the standardized index (around 32°C). Note that this value is slightly below the current Tmax threshold in Montreal. The GAIM attributes a slightly larger weight to lag 2 of $Tmax$ compared to lag 0 and 1, but with larger confidence intervals compared to the CGAIM. The g curve is similar to the one of the CGAIM but less smooth. Regarding Vp, the results from the CGAIM is very similar to those of Tmax, the weights roughly spread across lags and with a relationship sharply increasing at highest values of the index. In contrast, the GAIM attributes two opposite weights for lag 0 and 1 of Vp and a null weight for lag 2 with a ridge function oscillating around the zero line. Given the flatness of the ridge function, such a contrast between GAIM or CGAIM could either suggest the influence of unmeasured confounding, or some overfitting from the models. Indeed, evidence regarding the role of humidity in heat-related mortality is overall weak and inconsistent (Armstrong *and others* 2019).

6. Discussion

Following the growing need of understanding the impact of mixtures of environmental exposures on human health, the present paper proposes a method to construct indices with constraints under the form of a constrained groupwise additive index model (CGAIM). The CGAIM is expected to

be of use both for modelling and creating comprehensive indices for public health stakeholders. Its strengths include the possibility to include a high number of predictors $\boldsymbol{X}_j$ (including lags), include additional prior information from public health experts, and construct multiple indices simultaneously. Compared to previous work on the subject, the key novelties of the work are thus: i) the possibility to add any linear constraints on the index weights $\boldsymbol{\alpha}_j$, ii) the inclusion of constrained smoothing in the model to improve the indices usefulness, iii) a simple and efficient algorithm to estimate the indices, and iv) a criterion for index selection.

The constraints allow the proposed model to integrate additional information reflecting prior assumptions about the studied associations as well as integrate operational limitations to constructed indices. Examples of useful prior assumption include constraining indices and function shape to be convex for temperature-related mortality studies, or increasing for air-pollution-related mortality studies, for which usual flexible methods may fail (Armstrong *and others* 2020). Constraints can also force coefficients towards a specific feasible region to better control for unmeasured confounding causing issues such as the reversal paradox (Nickerson and Brown 2019). Adding such constraint for prior information, if correctly specified, also results in quicker convergence as shown by the timings reported in Supplementary Materials. On the other hand, operational constraints force constructed indices to have specific desirable properties. For instance, it is desirable that monitored heat indices reflect two constraints: i) decreased influence of higher lags to account for increased uncertainty in weather forecasts, and ii) a monotonic association with mortality for ease of interpretation. Such constraints might be desirable even at the expense of more optimal solutions. Although most applications displayed in this paper include non-negativity constraints, this is not a specificity of the method, and constraints with

negative coefficients are possible, for instance to construct exposure representing differences between variables.

A simulation study shows that the CGAIM can accurately estimate the index weights as well as the index relationship with the response variable compared to other advanced and recent models which is a step further in obtaining representative indices for practical applications. It shows that constraints help the model recover the true coefficient values. The simulation study also shows the model is robust to low sample sizes, highly correlated predictors, low signal-to-noise ratio, and high dimension with complex correlation patterns. The CGAIM is also compared to the PPR to evaluate the benefits of grouping variables, to the gMAVE as well as FACTS algorithms. Comparisons suggest that the CGAIM is more stable than these algorithms. In fact, even without any constraint, the proposed algorithm is efficient and converges quickly to an optimal solution, as shown by the comparison between the GAIM and gMAVE (see Supplementary Materials for a comparison of computational burden). In addition, simulation studies of sections 4.2 and 4.3 show that the model can efficiently recover the indices and variables that are the most predictive of the response.

Another strength of the work is in proposing and evaluation two inference procedures, an aspect of multiple index models that is often neglected in multiple index models, except in recently proposed Bayesian methods (McGee *and others* 2022). One proposed procedure is based on a normal approximation of constrained nonlinear least-squares, and one based on bootstrap resampling. Both methods however display non-negligible coverage error for confidence intervals. The normal approximation can especially widely underestimate the uncertainty. This is mainly related to the covariance matrix constructed from nonlinear least-squares that have been shown to significantly underestimate coverage even in far simpler settings (Donaldson and

Schnabel 1987). In contrast, bootstrap-based confidence intervals provide more satisfactory results although section 4.3 shows that they tend to overestimate uncertainty which is also consistent with previous work on bootstrap confidence intervals (Carpenter and Bithell 2000). Inference in constrained settings often presents mixed results (Meyer 2018), and further work is necessary to improve this aspect of the method.

The proposed method assumes that the variables and their grouping is selected *a priori*, with the idea that in many cases, the researcher has a clear idea of the relevant variables to be included. This assumption is reasonable in many applications in which a natural grouping of variables arises. For instance, in environmental epidemiology, exposure variable can often be grouped into category such as climate, air pollution or built-environment variables. Common tools to determine which variables to include in a study such as directed acyclic graphs (Greenland *and others* 1999) or clustering (Song and Zhu 2016) can also be used to determine a grouping *a priori*. When a limited number of concurrent groupings are investigated, the GCV criterion proposed in the present work can be used to decide. However, there may also be a need for a more automated selection procedure and an area for future research is thus to propose a flexible grouping mechanism. This is a difficult problem as the number of possible classifications increases dramatically with the number of variables.

Another limit of the proposed CGAIM is that it is currently restricted to continuous responses. Although this encompasses many situations, including counts when they are large enough such as in the applications above, it is of interest to extend this work to special cases such as logistic regression or survival analysis to increase its applicability. It is thus of interest to develop a generalized version of the CGAIM, in the same fashion as the generalized extension of the PPR (Roosen and Hastie 1993; Lingjærde and Liestøl 1998).

Supplementary material

Supplementary material includes full details on the data-generating mechanisms and additional results from the simulation study. A second application of the CGAIM to an air pollution index is also presented in Supplementary Materials. An R package cgaim implementing the method and the reproducible code for the simulations and applications are freely available on the first author's GitHub (https://github.com/PierreMasselot).

Acknowledgements

This work was supported by the Ouranos consortium; and the Quebec government's Fonds Vert under the Climate change Action Plan 2013-2020 financial contribution. The authors also want to thank the *Institut national de santé publique du Québec* for access to mortality data. The authors also acknowledge the important role of Jean-Xavier Giroux and Christian Filteau for extracting and managing the data and Anas Koubaa for helping in obtaining preliminary results. Finally, the authors thank the co-editors Professor Dimitris Rizopoulos & Professor Sherri Rose as well as two anonymous referees for their helpful comments and suggestions.

Bibliography

Agier, L., Portengen, L., Chadeau, -Hyam Marc, Basaga, ña X., Giorgis, -Allemand Lise, Siroux, V., Robinson, O., Vlaanderen, J., Gonz, ález J. R., Nieuwenhuijsen, M. J., et al. (2016). A Systematic Comparison of Linear Regression–Based Statistical Methods to Assess Exposome-Health Associations. *Environmental Health Perspectives* **124**, 1848–1856.

Armstrong, B. G., Gasparrini, A., Tobias, A. and Sera, F. (2020). Sample size issues in time series regressions of counts on environmental exposures. *BMC Medical Research Methodology* **20**, 15.

Armstrong, B., Sera, F., Vicedo-Cabrera, A. M., Abrutzky, R., Åström, D. O., Bell, M. L., Chen, B.-Y., de Sousa Zanotti Stagliorio Coelho, M., Patricia Matus, C., Dang Tran, N., et al. (2019). The Role of Humidity in Associations of High Temperature with Mortality: A Multiauthor, Multicity Study. *Environmental Health Perspectives* **127**, 097007.

Barreca, A. I. (2012). Climate change, humidity, and mortality in the United States. *Journal of Environmental Economics and Management* **63**, 19–34.

Bates, D. M. and Watts, D. G. (1981). A Relative Off set Orthogonality Convergence Criterion for Nonlinear least Squares. *Technometrics* **23**, 179–183.

Bates, D. M. and Watts, D. G. (1988). Nonlinear Regression Analysis and Its Applications. Wiley.

Bhaskaran, K., Gasparrini, A., Hajat, S., Smeeth, L. and Armstrong, B. (2013). Time series regression studies in environmental epidemiology. *International Journal of Epidemiology* **42**, 1187–1195.

Bobb, J. F., Valeri, L., Claus Henn, B., Christiani, D. C., Wright, R. O., Mazumdar, M., Godleski, J. J. and Coull, B. A. (2015). Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics* **16**, 493–508.

Boggs, P. T. and Tolle, J. W. (1995). Sequential Quadratic Programming. *Acta Numerica* **4**, 1–51.

Botev, Z. I. (2017). The normal law under linear restrictions: simulation and estimation via minimax tilting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79**, 125–148.

Breiman, L. (2001). Random forests. *Machine Learning* **45**, 5–32.

Carpenter, J. and Bithell, J. (2000). Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine* **19**, 1141–1164.

Chebana, F., Martel, B., Gosselin, P., Giroux, J.-X. and Ouarda, T. B. (2013). A general and flexible methodology to define thresholds for heat health watch and warning systems, applied to the province of Québec (Canada). *International journal of biometeorology* **57**, 631–644.

Chen, Y. and Samworth, R. J. (2015). Generalized additive and index models with shape constraints. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, n/a-n/a.

Cui, H.-Y., Zhao, Y., Chen, Y.-N., Zhang, X., Wang, X.-Q., Lu, Q., Jia, L.-M. and Wei, Z.-M. (2017). Assessment of phytotoxicity grade during composting based on EEM/PARAFAC combined with projection pursuit regression. *Journal of Hazardous Materials* **326**, 10–17.

Davalos, A. D., Luben, T. J., Herring, A. H. and Sacks, J. D. (2017). Current approaches used in epidemiologic studies to examine short-term multipollutant air pollution exposures. *Annals of Epidemiology* **27**, 145-153.e1.

DiCiccio, T. J. and Efron, B. (1996). Bootstrap Confidence Intervals. *Statistical Science* **11**, 189–212.

Donaldson, J. R. and Schnabel, R. B. (1987). Computational Experience with Confidence Regions and Confidence Intervals for Nonlinear Least Squares. *Technometrics* **29**, 67–82.

Durocher, M., Chebana, F. and Ouarda, T. B. M. J. (2016). Delineation of homogenous regions using hydrological variables predicted by projection pursuit regression. *Hydrology and Earth System Sciences* **20**, 4717–4729.

Efron, B. and Tibshirani, R. (1993). An introduction to the bootstrap. Chapman & Hall/CRC.

Fawzi, A., Fiot, J., Chen, B., Sinn, M. and Frossard, P. (2016). Structured Dimensionality Reduction for Additive Model Regression. *IEEE Transactions on Knowledge and Data Engineering* **28**, 1589–1601.

Friedman, J. H. and Stuetzle, W. (1981). Projection Pursuit Regression. *Journal of the American Statistical Association* **76**, 817–823.

Geweke, J. F. (1996). Bayesian Inference for Linear Models Subject to Linear Inequality Constraints. In Modelling and Prediction Honoring Seymour Geisser. Eds J. C. Lee, W. O. Johnson and A. Zellner. New York, NY: Springer. pp 248–263.

Golub, G. H., Heath, M. and Wahba, G. (1979). Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter. *Technometrics* **21**, 215–223.

Golub, G. and Pereyra, V. (2003). Separable nonlinear least squares: the variable projection method and its applications. *Inverse Problems* **19**, R1–R26.

Greenland, S., Pearl, J. and Robins, J. M. (1999). Causal diagrams for epidemiologic research. *Epidemiology* **10**, 37–48.

Guo, Z., Li, L., Lu, W. and Li, B. (2015). Groupwise Dimension Reduction via Envelope Method. *Journal of the American Statistical Association* **110**, 1515–1527.

Hansen, P. C., Pereyra, V. and Scherer, G. (2013). Least Squares Data Fitting with Applications. Johns Hopkins University Press. https://orbit.dtu.dk/en/publications/least-squares-data-fitting-with-applications [online; last accessed January 24, 2020].

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

Hastie, T. and Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science* **1**, 297–310.

Jolliffe, I. T. (2002). Principal Component Analysis. Springer Science & Business Media.

Keil, A. P., Buckley, J. P., O'Brien, K. M., Ferguson, K. K., Zhao, S. and White, A. J. (2020). A Quantile-Based g-Computation Approach to Addressing the Effects of Exposure Mixtures. *Environmental Health Perspectives* **128**, 047004.

Kong, E., Tong, H. and Xia, Y. (2010). Statistical modelling of nonlinear long-term cumulative effects. *Statistica Sinica* **20**, 1097–1123.

Lee, D., Ferguson, C. and Scott, E. M. (2011). Constructing representative air quality indicators with measures of uncertainty. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **174**, 109–126.

Li, L., Li, B. and Zhu, L.-X. (2010). Groupwise Dimension Reduction. *Journal of the American Statistical Association* **105**, 1188–1201.

Lingjærde, O. and Liestøl, K. (1998). Generalized projection pursuit regression. *SIAM Journal on Scientific Computing* **20**, 844–857.

Masselot, P., Chebana, F., Lavigne, É., Campagna, C., Gosselin, P. and Ouarda, T. B. M. J. (2019). Toward an Improved Air Pollution Warning System in Quebec. *International Journal of Environmental Research and Public Health* **16**, 2095.

Masselot, P., Chebana, F., Ouarda, T. B. M. J., Bélanger, D., St-Hilaire, A. and Gosselin, P. (2018). A new look at weather-related health impacts through functional regression. *Scientific Reports* **8**, 15241.

McGee, G., Wilson, A., Webster, T. F. and Coull, B. A. (2022). Bayesian multiple index models for environmental mixtures. *Biometrics* **n/a**. doi:10.1111/biom.13569.

Meyer, M. C. (2018). A Framework for Estimation and Inference in Generalized Additive Models with Shape and Order Restrictions. *Statistical Science* **33**, 595–614.

Meyer, M. C. (2008). Inference using shape-restricted regression splines. *Annals of Applied Statistics* **2**, 1013–1033.

Meyer, M. and Woodroofe, M. (2000). On the degrees of freedom in shape-restricted regression. *The Annals of Statistics* **28**, 1083–1104.

Monforte, P. and Ragusa, M. A. (2018). Evaluation of the air pollution in a Mediterranean region by the air quality index. *Environmental Monitoring and Assessment* **190**, 625.

Morris, T. P., White, I. R. and Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine* **0**. doi:10.1002/sim.8086.

Nickerson, C. A. and Brown, N. J. L. (2019). Simpson's Paradox is suppression, but Lord's Paradox is neither: clarification of and correction to Tu, Gunnell, and Gilthorpe (2008). *Emerging Themes in Epidemiology* **16**, 5.

Nieuwenhuijsen, M. J., Agier, L., Basagaña, X., Urquiza, J., Tamayo-Uria, I., Giorgis-Allemand, L., Robinson, O., Siroux, V., Maitre, L., de Castro, M., et al. (2019). Influence of the Urban Exposome on Birth Weight. *Environmental Health Perspectives* **127**, 47007.

Pappenberger, F., Jendritzky, G., Staiger, H., Dutra, E., Di Giuseppe, F., Richardson, D. S. and Cloke, H. L. (2015). Global forecasting of thermal health hazards: the skill of probabilistic predictions of the Universal Thermal Climate Index (UTCI). *International journal of biometeorology* **59**, 311–323.

Pya, N. and Wood, S. N. (2015). Shape constrained additive models. *Statistics and Computing* **25**, 543–559.

Ramsay, J. O. (1988). Monotone Regression Splines in Action. *Statistical Science* **3**, 425–441.

Robinson, O., Tamayo, I., de, C. M., Valentin, A., Giorgis, -Allemand Lise, Hjertager, K. N., Marit, A. G., Ambros, A., Ballester, F., Bird, P., et al. (2018). The Urban Exposome during Pregnancy and Its Socioeconomic Determinants. *Environmental Health Perspectives* **126**, 077005.

Roosen, C. B. and Hastie, T. J. (1994). Automatic Smoothing Spline Projection Pursuit. *Journal of Computational and Graphical Statistics* **3**, 235–248.

Roosen, C. B. and Hastie, T. J. (1993). Logistic Response Projection Pursuit. AT&T Bell laboratories.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* **1**, 206–215.

Schittkowski, K. (1988). Solving Constrained Nonlinear Least Squares Problems by a General Purpose SQP-Method. In Trends in Mathematical Optimization: 4th French-German Conference on Optimization. Eds K.-H. Hoffmann, J. Zowe, J.-B. Hiriart-Urruty and C. Lemarechal. Basel: Birkhäuser Basel. pp 295–309.

Schmidt, C. W. (2020). Into the Black Box: What Can Machine Learning Offer Environmental Health Research? *Environmental Health Perspectives* **128**, 022001.

Song, S. and Zhu, L. (2016). Group-wise semiparametric modeling: A SCSE approach. *Journal of Multivariate Analysis* **152**, 1–14.

Tallis, G. M. (1965). Plane Truncation in Normal Populations. *Journal of the Royal Statistical Society: Series B (Methodological)* **27**, 301–307.

Thornton, M. M., Shrestha, R., Wei, Y., Thornton, P. E., Kao, S.-C. and Wilson, B. E. (2020). Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 4. *ORNL DAAC*. doi:10.3334/ORNLDAAC/1840.

Van den Meersche, K., Soetaert, K. and Van Oevelen, D. (2009). xsample(): An R Function for Sampling Linear Inverse Problems. *Journal of Statistical Software* **30**, 1–15.

Wang, K. and Lin, L. (2017). Robust and efficient direction identification for groupwise additive multiple-index models and its applications. *TEST* **26**, 22–45.

Wang, S.-J. and Ni, C.-J. (2008). Application of Projection Pursuit Dynamic Cluster Model in Regional Partition of Water Resources in China. *Water Resources Management* **22**, 1421–1429.

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

Wood, S. N. (2017). Generalized Additive Models: An Introduction with R. 2nd ed. Chapman and Hall/CRC.

Wood, S. N. (2004). Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models. *Journal of the American Statistical Association* **99**, 673–686.

Xia, Y. and Tong, H. (2006). Cumulative effects of air pollution on public health. *Statistics in Medicine* **25**, 3548–3559.

Yu, Y. and Ruppert, D. (2002). Penalized Spline Estimation for Partially Linear Single-Index Models. *Journal of the American Statistical Association* **97**, 1042–1054.

Yuan, J., Xie, C., Zhang, T., Sun, J., Yuan, X., Yu, S., Zhang, Y., Cao, Y., Yu, X., Yang, X., et al. (2016). Linear and nonlinear models for predicting fish bioconcentration factors for pesticides. *Chemosphere* **156**, 334–340.

Yuan, M. (2011). On the Identifiability of Additive Index Models. *Statistica Sinica* **21**, 1901–1911.

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

Zhou, H. and Lange, K. (2013). A Path Algorithm for Constrained Estimation. *Journal of computational and graphical statistics : a joint publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America* **22**, 261–283.

Figures

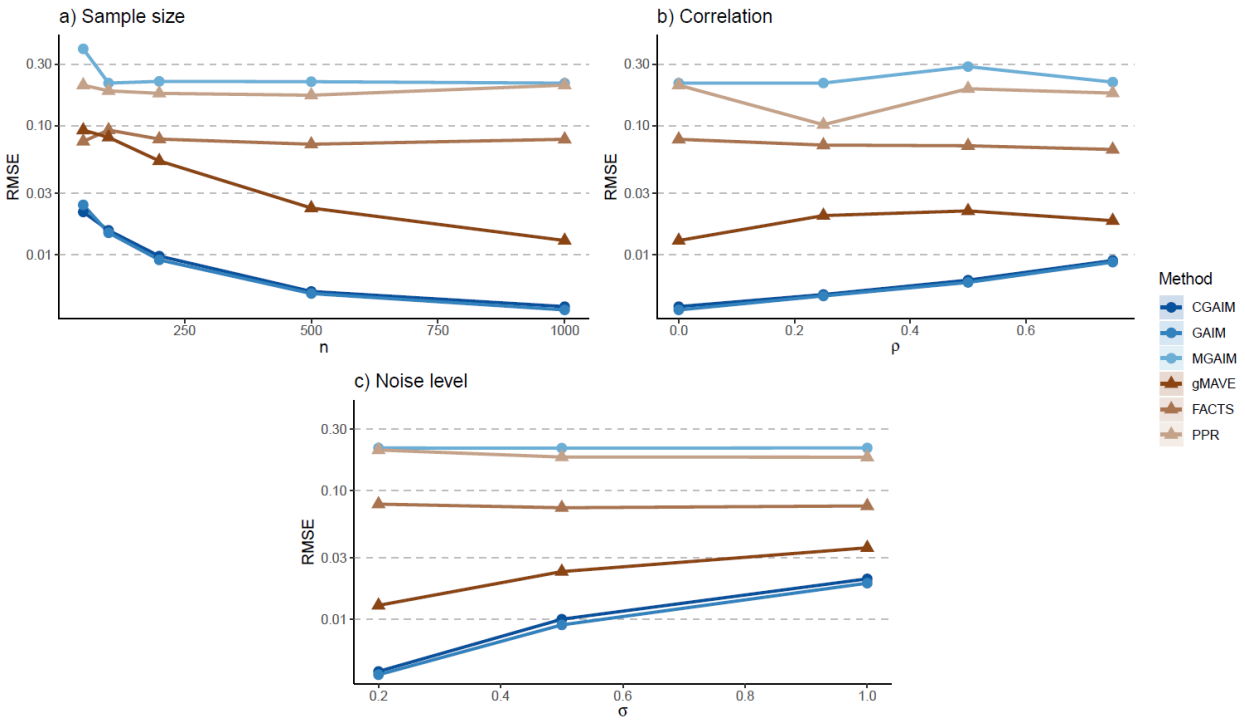

Figure 1: Estimated RMSE for different scenarios, varying the sample size (a), the correlation between variables (b) and the noise level (c). Note that the CGAIM and GAIM curves overlap each other at the bottom. Note the log scale.

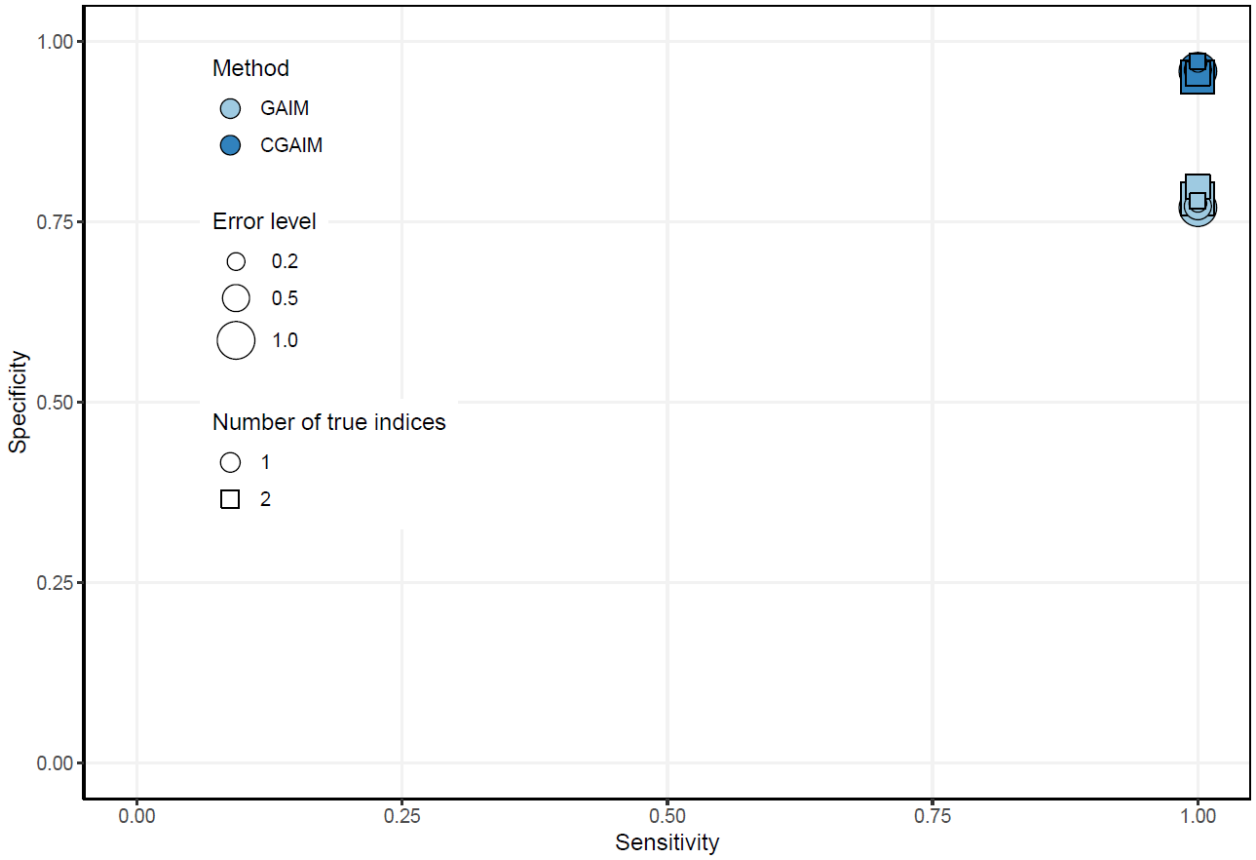

Figure 2: Average sensitivity and specificity of index selection computed on the 1000 simulations for various number of true indices and error level.

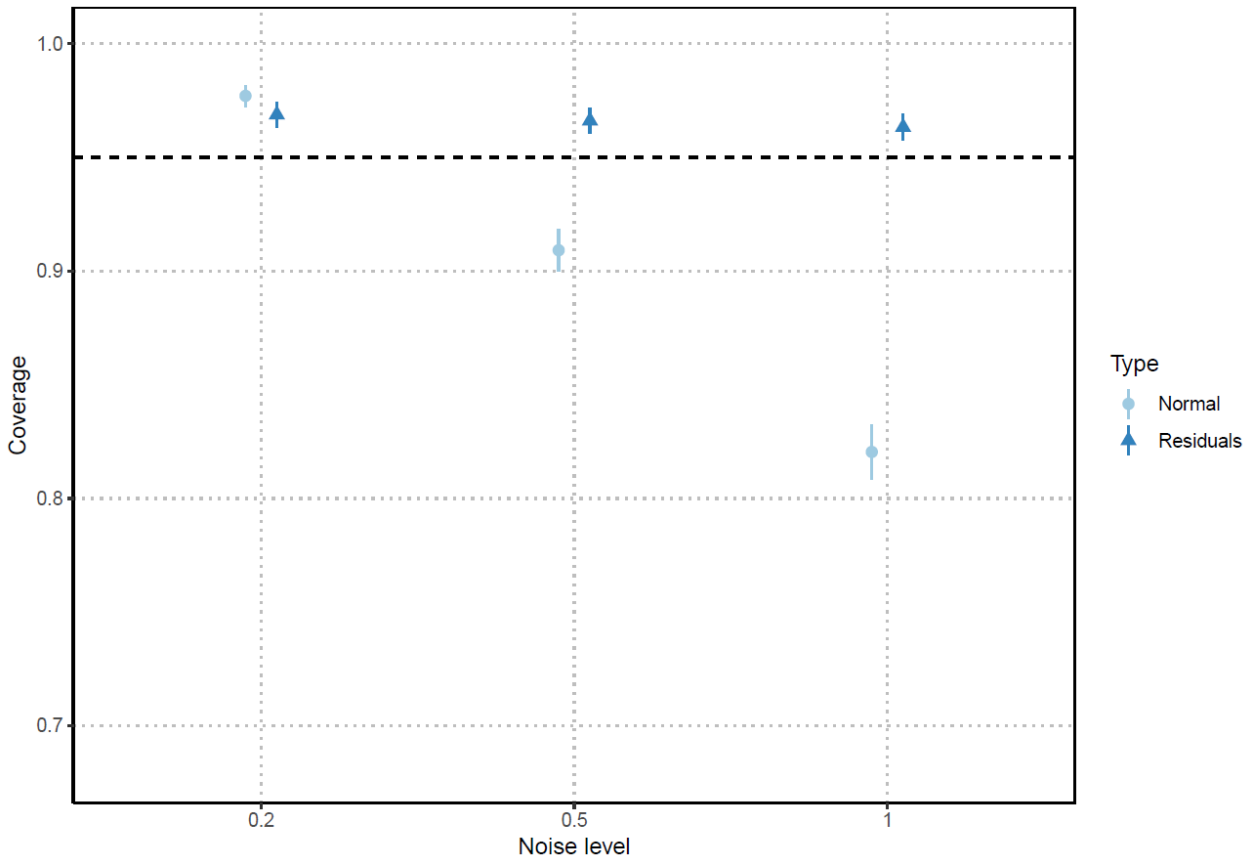

Figure 3: Estimated coverage for both inference methods and various noise level. Vertical segments indicate +/- 1 standard error of the coverage.

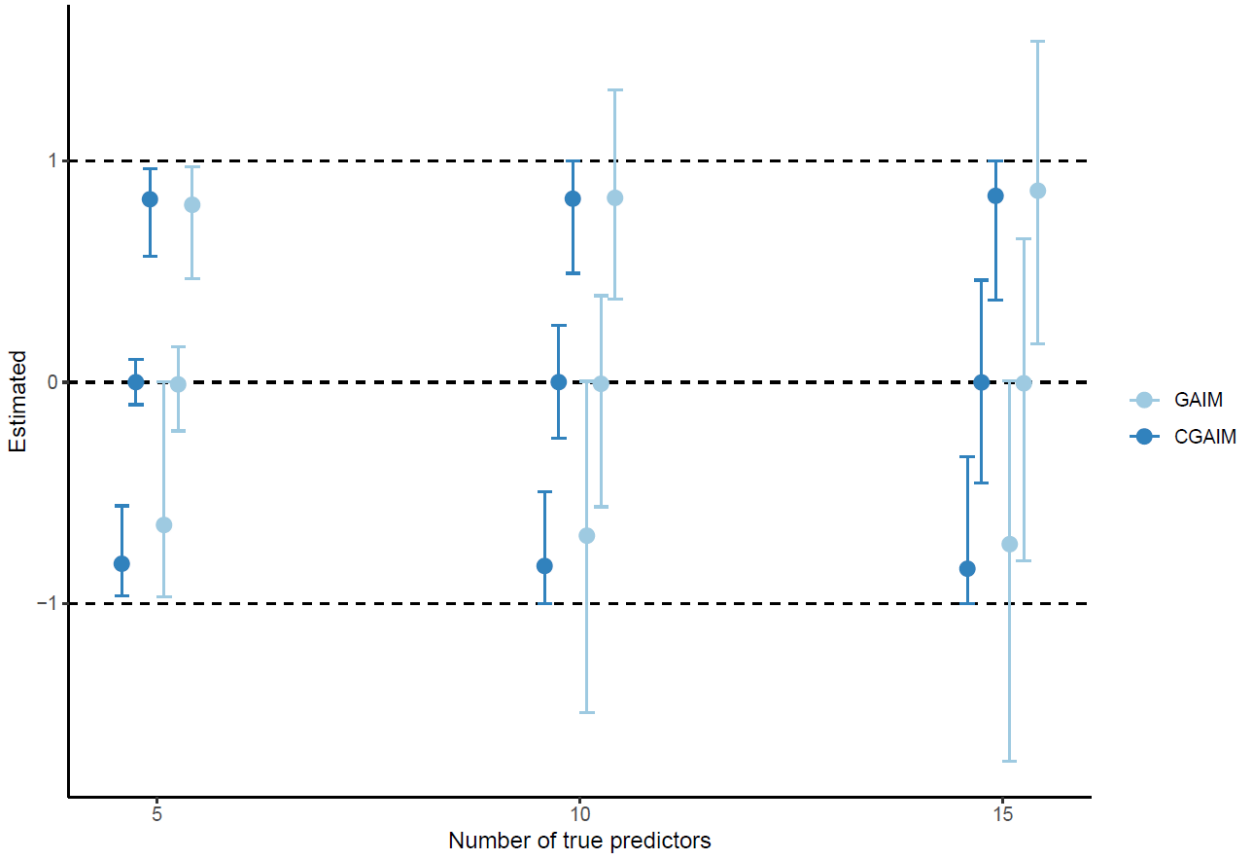

Figure 4: Average estimated $\hat{\alpha}_j$ for according to the true value α_j. Segments indicate 2.5th and 97.5th percentile of estimated α_j.

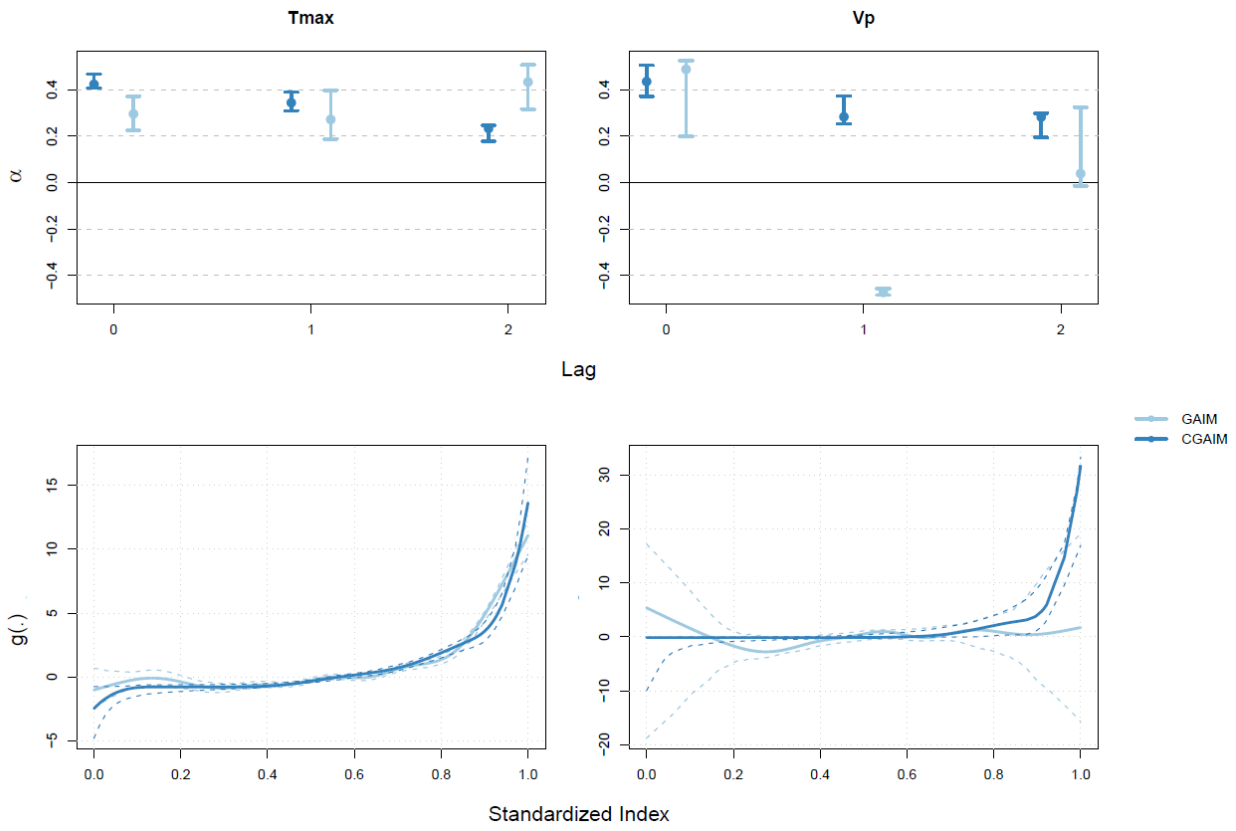

Figure 5: Resulting indices created in Montreal. Top row: weights $\boldsymbol{\alpha_j}$ for each selected index; bottom row: functions $\boldsymbol{g_j}$. Indices have been standardized over the range $[0-1]$ for ease of comparison. Each column corresponds to one index. Vertical segments and dotted lines represent block-bootstrap 95% confidence intervals.