Biostatistics (2018) **19**, 3, *pp.* 359–373
doi:10.1093/biostatistics/kxx043
Advance Access publication on September 13, 2017

High-dimensional multivariate additive regression for uncovering contributing factors to healthcare expenditure

RODRIGUE NGUEYEP*
IBM T.J. Watson Research Center, 1101 Kitchawan Road, Yorktown Heights, NY 10598, USA
ngueyep@us.ibm.com

NICOLETA SERBAN
School of Industrial Systems and Engineering, Georgia Institute of Technology,
755 Ferst Drive, NW, Atlanta, GA 30332, USA

Summary

Many studies in health services research rely on regression models with a large number of covariates or predictors. In this article, we introduce novel methodology to estimate and perform model selection for high-dimensional non-parametric multivariate regression problems, with application to many healthcare studies. We particularly focus on *multi-responses or multi-task regression* models. Because of the complexity of the dependence between predictors and the multiple responses, we exploit model selection approaches that consider various level of groupings between and within responses. The novelty of the method lies in its ability to account simultaneously for between and within group sparsity in the presence of non-linear effects. We also propose a new set of algorithms that can identify inactive and active predictors that are common to all responses or to a subset of responses. Our modeling approach is applied to uncover factors that impact healthcare expenditure for children insured through the Medicaid benefits program. We provide important findings on the association between healthcare expenditure and a large number of well-cited factors for two neighboring states, Georgia and North Carolina, which have similar demographics but different Medicaid systems. We also validate our methods with a benchmark cancer data set and simulated data examples.

Keywords: Sparse Additive Models; Healthcare expenditure; High-dimensional data; Multivariate regression; Non-parametric regression; Medicaid health benefits program.

1. Introduction

Data in healthcare are generated at every patient's encounter with the healthcare system resulting in billions of pieces of data every day. Every patient in any medical setting generates an invaluable data point that can contribute to understanding what works, for who and where. Patient data are acquired within various settings, including administrative, clinical, policy, or discovery among others. Particularly, administrative

*To whom correspondence should be addressed.

or claims data are often used to make inference on delivery outcomes of a healthcare system, for example, of great interest to the current health policy making is healthcare expenditure (Neff *and others*, 2004; Reschovsky *and others*, 2011).

In this study, we are particularly interested in understanding what factors are associated with healthcare expenditure for the Medicaid system, a governmental insurance program for low-income population in the United States. The number of potential factors driving healthcare expenditure can be however very large, representing various determinants of health and healthcare, including social and economic environment, demographics, access to care, health indicators, to name a few, and most importantly, utilization of the healthcare system.

Most existing studies in healthcare focus on aggregated expenditure at the county level because many factors of interest are also county aggregates. These aggregations further reduce the sample size, with an increase in the ratio of the number of factors (p) to the observation sample size (n). Moreover, the relationships between the contributing factors and the healthcare expenditure are not necessarily linear as assumed in much of the existing health services research. These relationships may also be different across communities and/or time periods. Addressing these challenges intrinsic to the study of variations in healthcare expenditure or other system outcomes requires development and implementation of high-dimensional models with inference on model selection and non-linear associative relationships.

Estimation of (non)linear associations and model selection in a high-dimensional setting is a common statistical problem encountered in several other fields, including biostatistics, genomics, and imaging among others. Recent research has advanced well-established regularized regression approaches (Tibshirani, 1996) to the analysis of high-dimensional multivariate regression (Obozinski *and others*, 2011).

In multivariate linear regression, the model of interest is given by $Y_i^{(k)} = \sum_{j=1}^{P} B_{jk} x_{ij}^{(k)} + \varepsilon_i^{(k)}$, with $i = 1, \ldots, n_k$ where n_k is the sample size associated with the kth response $Y^{(k)}$ for $k = 1, \ldots, K$ responses. The P covariates $\mathbf{x}_j^{(k)}, j = 1, \ldots, P$ can take the same values across all K responses (referred herein as *multi-responses regression*) or different values (referred herein as *multi-task regression*). In this article, we consider a generalization of the multivariate linear regression model in that it allows modeling non-linear regression functions of the covariates in an additive manner, similarly to the additive regression models (Hastie and Tibshirani, 1986), resulting in a multi-responses/task additive regression model

$$\begin{aligned} Y^{(1)} &= \beta^{(1)} + \overbrace{f_1^{(1)}(X_1^{(1)})}^{\text{Group 1}} + \cdots + \overbrace{f_P^{(1)}(X_P^{(1)})}^{\text{Group P}} + \varepsilon^{(1)} \\ &\vdots \\ Y^{(K)} &= \beta^{(K)} + \underbrace{f_1^{(K)}(X_1^{(K)})}_{\text{Group 1}} + \cdots + \underbrace{f_P^{(K)}(X_P^{(K)})}_{\text{Group P}} + \varepsilon^{(K)} \end{aligned} \tag{1.1}$$

Each model denoted by (k) links a set of predictors $X_1^{(k)}, \ldots, X_P^{(k)}$ to a response $Y^{(k)}$ through univariate additive functions$f_1^{(k)}, \ldots, f_P^{(k)}$ for $k = 1, \ldots, K$. In this non-parametric setting, the objective is to perform model selection on the additive terms while also controlling the smoothness of the selected additive components. We also assume that there is a intercept $\beta^{(k)}$ associated with each responses.

In this article, we will also consider cases where the responses $Y^{(k)}$ are categorical, and "$Y^{(k)} = 1$" if the corresponding observation belongs to category k with $k \in (1, \ldots, K)$. We will use a multi-category additive logistic regression model

$$\mathbb{P}(Y^{(k)} = 1 | X = x) = \frac{\exp(f^{(k)}(x))}{1 + \sum_{l=1}^{K-1} \exp(f^{(l)}(x))}$$

Lin and Zhang (2006), Meier *and others* (2009), Yin *and others* (2012) and Ravikumar *and others* (2009) proposed different methods to simultaneously perform model selection and smoothness of the additive components in the presence of one response ($K = 1$). These methods are not well suited to problems with multiple responses since they solve each model (k) individually. Liu *and others* (2009) and Foygel *and others* (2012) estimate all these models jointly by taking advantage of the fact that responses can share common predictors. Liu *and others* (2009) adapt the ℓ_∞ regularization method of Turlach *and others* (2005) to the non-parametric setting; and Foygel *and others* (2012) use a non-parametric reduced rank regression. These multi-responses regression methods consider that all functions associated with one predictor are in the same group, implying that the same set of predictors is selected to model all responses. These methods however do not allow selection of different predictors for multiple responses, which in turn can result in overfitted models because if a predictor is selected as active for one response, it will automatically be selected for all the other responses. This is because these methods use regularization norms that achieve groupwise selection only.

The implication of group wise selection of predictors is that if the jth group of additive functions is selected then all the predictors $X_j^{(1)}, \ldots, X_j^{(K)}$ associated with these functions influence the responses. While such larger models can be useful, it is desirable to reduce the set of predictors further to better understand whether the association of one predictor to a response is an outlier, i.e. only influencing a small number of responses, or whether it is underlying, i.e. present throughout a large number of responses. It can be viewed as the strength of the association between a predictor and the multiple responses. For this, within group selection needs to complement the selection of the group of additive functions. Hence, we can select or discard an entire group of functions, but we can also select a subset of functions within a group. This model selection idea has been considered in the context of linear regression by Simon *and others* (2013). They proposed a regularized scheme, which uses a joint ℓ_1 and $\ell_1 \backslash \ell_2$ penalty.

One contribution of this work is a non-parametric estimation method that can perform joint estimation and selection of multi-responses additive models in high-dimensional settings. To the best of our knowledge, this is the first method that can simultaneously perform group wise and within group selection of additive functions in the presence of multiple responses. We also introduce an efficient algorithm that relies on functional block coordinate descent and functional coordinate descent to estimate additive functions. The main advantage of the algorithm is that each non-null additive function is updated via a closed form formula. We adapt the penalized local scoring algorithm introduced by Liu *and others* (2009) to our estimation method, and thus, we can perform multi-category classification with additive functions.

We apply the proposed methodology to identify the factors associated with healthcare expenditure, including healthcare utilization for services provided by different provider types, determinants of health such as socio-economic factors and demographics, and healthcare access factors among others. We perform the analysis at the county level and with comparison of two states, Georgia and North Carolina. Our focus is on the children population insured under the Medicaid-benefits program. We compare the association between expenditure and relevant factors for five different years, where each year represents a different response in the multi-responses regression model. The first objective is to identify factors that are significantly associated with expenditure across 5 years and to identify outlying associations if any. The second goal is to exploit the use of additive functions to uncover the nature of the non-linear impact of these identified factors on the expenditure across years. We will also be able to assess how these factors associate to healthcare expenditure over time with a comparison between the two states.

To validate our methodology, we also perform multi-category tumor classification and biomarker discovery from a benchmark gene microarray data that has been used in previous articles (Kahn *and others*, 2001; Tibshirani *and others*, 2002; Zhang *and others*, 2008; Liu *and others*, 2009). We find that our method is able to achieve high classification accuracy with a number of genes substantially lower than the previous methods that exploited this data set. We accompany this analysis with a series of simulations. We defer the validation analysis to the Supplemental available at *Biostatistics* online.

The remainder of the article is organized as follows. In Section 2, we review the literature on non-parametric model selection methods in high dimensional settings, and we explain the need for the methodology we introduce in this article. In Section 3, we introduce a new method that jointly leverages a composite $\ell_1 \backslash \ell_2$ functional norm and ℓ_1 functional norm to induce group and within group sparsity across additive functions. In Section 4, we introduce the factors considered in our application, and we also provide their respective sources. We then use our method to uncover factors that impact healthcare expenditure for children insured through the Medicaid benefits program. We present the results of the method applied to synthetic data in the Web Appendix of supplementary material available at *Biostatistics* online. In Section 5, we conclude by providing insights on the application of the proposed method to the motivating applied problem as well as other applied problems. In the web supplements, we present proofs of the conditions presented in Section 3, and the validation analysis performed on a cancer data set, where we compare our method to existing methods and additional results showing the non-linear effects of all the factors on Medicaid related healthcare expenditure.

2. Background

In this section, we provide a brief description of existing methods used to tackle high-dimensional additive models. These methodologies will be used as a building block for the proposed methodology. If we consider a random vector $\mathbf{X} = (X_1, \dots, X_P)$ and a random variable Y, a typical statistical problem is the estimation of additive models.

$Y = \beta + f(\mathbf{X}) + \varepsilon$, where $\mathbb{E}(\varepsilon) = 0$, and $f(\mathbf{X}) = f_1(X_1) + \cdots + f_P(X_P)$ and $\beta \in \mathbb{R}$ is the intercept. For a given random variable X_j with distribution μ_j and a measurable function f_j of x_j, $\|f_j\|$ denotes the $L_2(\mu_j)$ norm $f_j : \|f_j\| = \sqrt{\mathbb{E}\left[f^2(x_j)\right]} = \sqrt{\int_{\mathcal{X}_j} f^2(x_j) d\mu_j}$. For $x \in \mathbb{R}^n$, we denote $\|x\|^2 = \frac{1}{n}\sum_{i=1}^n x_i^2$. For a random variable X_j with $j \in \{1, \dots, P\}$, $\mathcal{H}_j = \{f_j | E(f_j(x_j)) = 0\}$ denotes a Hilbert subspace of functions f_j that are $L_2(\mu_j)$ measurable, with mean zero, and inner product defined as $\langle f_j, g_j \rangle = E(f_j(x_j) g_j(x_j))$. We will focus on P-dimensional functions $f(\mathbf{x})$ that have an additive form $f(\mathbf{x}) = \sum_{j=1}^P f_j(x_j)$, with $f_j \in \mathcal{H}_j$. To estimate the additive functions f_j in their model, Hastie and Tibshirani (1986) find the set of functions that minimize the least squares criterion defined below:

$\underset{\mathbf{f}: f_j \in \mathcal{H}_j}{\text{Min}}\, L(\mathbf{f})$ where $L(\mathbf{f}) = \frac{1}{2}\mathbb{E}\left[\left(Y - \beta - \sum_{j=1}^P f_j(X_j)\right)^2\right]$. They propose to use a backfitting approach to estimate the smooth functions f_j, $f_j = \mathbb{E}\left[\left(Y - \beta - \sum_{l \neq j} f_l\right) | X_j\right] = Q_j\left(Y - \beta - \sum_{l \neq j} f_l\right)$, where $Q_j = \mathbb{E}\left[.|X_j\right]$ is the projection operator onto $\mathcal{H}_j$. The estimate of each smooth function under this setting is given by $\hat{\mathbf{f}}_j \leftarrow S_j(Y - \hat{\beta} - \sum_{l \neq j} \hat{\mathbf{f}}_l)$, where S_j is a smoothing matrix, $\hat{\mathbf{f}}_j \in \mathbb{R}^n$ and $\hat{\mathbf{f}}_j = \left[f_j(x_{1j}), \dots, f_j(x_{nj})\right]$, and $\hat{\beta} = \frac{\sum_{i=1}^n Y_i}{n}$. Other algorithms have been proposed to estimate additive models, for instance, marginal integration was introduced by Linton and Nielsen (1995).

Ravikumar *and others* (2009) extends this model to the high-dimensional setting ($P \gg n$) and create a method labeled Sparse Additive Models (SpAM). They impose sparsity at the function level and thus are able to perform model selection by solving the optimization problem:

$\underset{\mathbf{f}: f_j \in \mathcal{H}_j}{\text{Min}}\, L(\mathbf{f}) + \lambda \Omega(\mathbf{f})$, with $\Omega(\mathbf{f}) = \sum_{j=1}^P \|f_j\|$. They use a sparse backfitting algorithm where at each step the smooth function updates are given by: $f_j = \left[1 - \frac{\lambda}{\|Q_j R_j\|}\right]_+ Q_j R_j$, With $R_j = Y - \beta - \sum_{l \neq j} f_l$ is a partial residual, the operator $[.]_+ = \max(0, .)$.

More recently, Yin *and others* (2012) proposed the GroupSpAM, which adapts the group lasso to the non-parametric setting. They penalize the additive components in a group manner. For a partition $\mathcal{G}$ of $\{1, \dots, p\}$, they solve the following optimization problem:

$\underset{\mathbf{f}:f_j\in\mathcal{H}_j}{\text{Min}}\, L(\mathbf{f}) + \lambda\Omega_{\mathbf{group}}(\mathbf{f})$, with $\Omega_{\mathbf{group}}(\mathbf{f}) = \sum_{g\in\mathcal{G}} \sqrt{d_g}\|\mathbf{f}_g\| = \sum_{g\in\mathcal{G}} \sqrt{d_g}\sqrt{\sum_{j\in g} \mathbb{E}\left[f_j^2(X_j)\right]}$, where $\mathbf{f}_g$ is a set of functions in group g and d_g is the number of predictors in group g. The computational approach to update the group of additive functions is a block coordinate descent algorithm.

In our methodology, we exploit the use of a composite $\ell_2\backslash\ell_1$ functional penalty and ℓ_1 penalty to identify the covariates whose functions are active for at least one of the responses in a multi-task or multi-responses regression model. This allows to select all the additive functions associated with a specific predictor and uniquely determines for which responses this predictor is relevant.

3. $L_2\backslash L_1$ AND L_1 JOINT FUNCTIONAL SPARSITY

As stated in the introduction, we consider a multi-task (or a multi-responses) regression problem with K responses. For each regression problem, the data are $\{(\mathbf{x}_i^{(k)}, y_i^{(k)}), i = 1, \ldots, n_k, k = 1, \ldots, K\}$. Without loss of generality, we assume that $n_1 = \cdots = n_K = n$. Each model has the form:

$$Y^{(k)} = \beta^{(k)} + \sum_{j=1}^{P} f_j^{(k)}(X_j^{(k)}). \tag{3.1}$$

We assume that $\mathbb{E}\left[f_j^{(k)}(X_j^{(k)})\right] = 0\ \forall j \in \{1, \ldots, P\}$. By solving the optimization problem (3.2), we can jointly estimate groups of functions $\mathbf{f}_j = \left(f_j^{(1)}, \ldots, f_j^{(K)}\right)$ or $\mathbf{f}^{(k)} = \left(f_1^{(k)}, \ldots, f_P^{(k)}\right)$. We propose to induce sparsity within the groups of predictors that appear to be relevant, to account for the fact that not all the responses share the same predictors.

$$\underset{\mathbf{f}^{(1)},\ldots,\mathbf{f}^{(K)}}{\text{Min}} \sum_{k=1}^{K} L(\mathbf{f}^{(k)}) + (1-\alpha)\lambda\Omega_{\ell_2\backslash\ell_1}(\mathbf{f}) + \alpha\lambda\Omega_{\ell_1}(\mathbf{f}), \tag{3.2}$$

where $L(\mathbf{f}^{(k)}) = \frac{1}{2}\mathbb{E}\left[\left(Y^{(k)} - \sum_{j=1}^{P} f_j^{(k)} - \beta^{(k)}\right)^2\right]$, $\Omega_{\ell_1}(\mathbf{f}) = \sum_{j=1}^{P}\sum_{k=1}^{K} \|f_j^{(k)}\|$, and $\Omega_{\ell_2\backslash\ell_1}(\mathbf{f}) = \sum_{j=1}^{P}\sqrt{K}\sqrt{\sum_{k=1}^{K}\|f_j^{(k)}\|^2} = \sum_{j=1}^{P}\sqrt{K}\|\mathbf{f}_j\|$. Solving the optimization problem (3.2) is equivalent to solving the problem for the covariate index j while all the others covariates are held constant.

$$\hat{f}_j^{(1)}, \ldots, \hat{f}_j^{(K)} = \underset{f_j^{(1)},\ldots,f_j^{(K)}}{\text{argmin}} \sum_{k=1}^{K} \frac{1}{2}\mathbb{E}\left[\left(R_j^{(k)} - f_j^{(k)}\right)^2\right] + (1-\alpha)\lambda\sqrt{K}\|\mathbf{f}_j\| + \alpha\lambda\sum_{k=1}^{K}\|f_j^{(k)}\|, \tag{3.3}$$

where $R_j^{(k)} = Y^{(k)} - \sum_{\ell\neq j} f_l^{(k)} - \beta^{(k)}$ is the partial residual associated with the kth task and the jth covariate $X_j^{(k)}$. The regularization parameter λ is responsible for inducing sparsity among the additive functions at the group level.When λ is large, many additive functions are shrunk to the null function. The level of sparsity applied to individual additive functions within each group is controlled by $\alpha \in (0, 1)$. So when α is close to 1, we have a model where there is mainly sparsity induced separately on each function, while when α is close to 0 sparsity is induced mainly in a group manner.

A block coordinate descent algorithm can be used to solve the optimization problem (3.3). The first step consists in deriving the stationary conditions that are satisfied when $\mathbf{f}_j$ is optimal (see Web Appendix D of supplementary material available at *Biostatistics* online). This derivation is done by computing the gradients and subgradients of functionals in the Hilbert space $\mathcal{H}_j$. By leveraging these stationary conditions, we are able to find necessary and sufficient conditions under which the optimal solution for problem (3.3) is $\mathbf{f}_j = \mathbf{0}$. The first condition provides the criterion that allows us to determine if all the additive functions associated with variables $X_j^{(k)}$, with $k \in \{1, \ldots, K\}$ are null.

Condition 1: The covariates $X_j^{(k)}$ with $k \in \{1, \ldots, K\}$ are inactive as a group through their additive functions $f_j^{(k)}$, $\mathbf{f}_j = \mathbf{0}$ if and only if

$$\sqrt{\sum_{k=1}^{K} \mathbb{E}\left[\left[\left(1 - \frac{\alpha\lambda}{\|Q_j^{(k)}R_j^{(k)}\|^2}\right)_+ Q_j^{(k)}R_j^{(k)}\right]^2\right]} < (1-\alpha)\lambda\sqrt{K}. \tag{3.4}$$

The proof of this condition is given in Web Appendix D of supplementary material available at *Biostatistics* online. The partial residuals $R_j^{(k)}$ (associated with $f_j^{(k)}$) measures the amount of variation in the responses $Y^{(k)}$ that is left unexplained after accounting for the effects of all the other additive functions, that can influence the response $Y^{(k)}$. If there is a lot of variation left unexplained, condition 1 suggests that the group of additive functions $f_j^{(1)}, \ldots, f_j^{(K)}$ will be selected because the weighted sum of ℓ_2 norms of the projection of the partial residuals on the space spanned by the predictors $X_j^{(1)}, \ldots, X_j^{(K)}$ will be larger than the threshold. If for a covariate $X_j^{(k)}$, $\|Q_j^{(k)}R_j^{(k)}\|^2 < \alpha\lambda$, we obtain $\left(1 - \frac{\alpha\lambda}{\|Q_j^{(k)}R_j^{(k)}\|^2}\right)_+ = 0$ and the term $Q_j^{(k)}R_j^{(k)}$ doesn't contribute to the group condition.

Thus the main difference between our new condition (3.4) and the condition in (1.5) in the Web Appendix of supplementary material available at *Biostatistics* online lies in the fact that the new condition puts emphasis on the within group sparsity.

The second condition provides the criteria for determining which functions $f_j^{(k)}$ should be selected out of an active group of functions $\mathbf{f}_j = \left(f_1^{(1)}, \ldots, f_1^{(K)}\right)$.

Condition 2: For a given index j if the set of covariates $X_j^{(1)}, \ldots, X_j^{(K)}$ is active then a covariate $X_j^{(k)}$ with $k \in \{1, \ldots, K\}$ is inactive through its additive function $f_j^{(k)}, f_j^{(k)} = 0$ if and only if

$$\sqrt{\mathbb{E}\left[\left(Q_j^{(k)}R_j^{(k)}\right)^2\right]} \leq \alpha\lambda. \tag{3.5}$$

To derive this condition, we simply incorporate the explicit derivative of the group penalty $\|\mathbf{f}_j\|$ in the stationary conditions since we know that at least one function out of $\mathbf{f}_j$ is selected. Based on this updated stationary condition, we rely on properties of the subgradient of the function $f_j^{(k)}$ to derive condition 2. The complete proof of condition 2 is given in Web Appendix E of supplementary material available at *Biostatistics* online. The condition simply states that the additive function $f_j^{(k)} = 0$ if the ℓ_2 norm of the projection of partial residual $R_j^{(k)}$ on the space spanned by the covariate $X_j^{(k)}$ is less than the threshold $\alpha\lambda$.

Conditions 1 and 2 are used as thresholding conditions in Algorithm 1, which expands on how the additive functions associated with the same predictors are updated. The main idea is to first check using condition 1, if a group of additive functions has an effect on the responses. If a group is selected, we then test using condition 2 if each additive function in the group has a ℓ_2 norm greater than the defined threshold. The algorithm also shows that the non-null additive functions are updated through a closed-form formula and that the new value depends on the previous updates of the additive functions in the same groups. In the algorithm, the theoretical values of the criteria used in the thresholds are estimated by empirical values that are computed by multiplying the empirical partial residuals by a smoothing matrix associated with a specific predictor. Hastie and Tibshirani (1986) demonstrated the convergence of backfitting when smoothers such as smoothing kernels and splines are used as empirical estimators. This result should hold within our algorithm since it relies mainly on backfitting.

For all the results reported in this article, we used a Gaussian kernel smoothers, and we use a plugin bandwidth $h_j = 0.6 * sd(X_j)n^{-1/5}$. Note that other smoothers such as natural cubic splines and polynomial splines can also be used.

Algorithm 1 Soft-thresholding operator $\text{Algo1}_\lambda[\hat{R}_j^{(1)}, \ldots, \hat{R}_j^{(K)}, S_j^{(1)}, \ldots, S_j^{(K)}]$

```
(1) Input: Smoothing matrices $\mathbf{S}_j^{(k)}$, estimated partial residuals $\hat{\mathbf{R}}_j^{(k)}$,
    for $j$ fixed and $k \in \{1, \ldots, K\}$, and the regularization parameter $\lambda$.
(2) Estimate $\mathbf{Q}_j^{(k)}\mathbf{R}_j^{(k)} = \mathbb{E}\left[R_j^{(k)} | X_j^{(k)}\right]$ as $\hat{Q}_j^{(k)} = S_j^{(k)}\hat{R}_j^{(k)}$ $\forall k \in \{1, \ldots, K\}$
(3) Estimate $\forall k \in \{1, \ldots, K\}$ $g_j^{(k)} = \left(1 - \frac{\alpha\lambda}{\|Q_j^{(k)}R_j^{(k)}\|^2}\right)_+$ as $\hat{g}_j^{(k)} = \left(1 - \frac{\alpha\lambda}{\|\hat{Q}_j^{(k)}\|^2}\right)_+$
(4) Estimate $h_j = \sqrt{\sum_{k=1}^K \mathbb{E}\left[\left[\left(1 - \frac{\alpha\lambda}{\|Q_j^{(k)}R_j^{(k)}\|^2}\right)_+ Q_j^{(k)}R_j^{(k)}\right]^2\right]}$ as $\hat{h}_j = \sqrt{\sum_{k=1}^K \frac{1}{n}\|\hat{g}_j^{(k)}\hat{Q}_j^{(k)}\|^2}$
(5) If $\hat{h}_j < (1-\alpha)\lambda\sqrt{K}$ set $\hat{\mathbf{f}}_j = \mathbf{0}$
(6) Else
    i. For each $k \in \{1, \ldots, K\}$ if $\|\hat{Q}_j^{(k)}\| < \alpha\lambda$ then $\hat{f}_j^{(k)} = 0$
    ii. Else update $\hat{f}_j^{(k)}(i+1) = \frac{\hat{Q}_j^{(k)}}{1 + \frac{(1-\alpha)\lambda\sqrt{K}}{\|\hat{\mathbf{f}}_j(i)\|} + \frac{\alpha\lambda}{\|\hat{f}_j^{(k)}(i)\|}}$, where $i$ is the $i$th iteration
(7) Center all the estimated functions $\hat{f}_j^{(k)} \leftarrow \hat{f}_j^{(k)} - \text{mean}(\hat{f}_j^{(k)})$
(8) Output: estimated additive functions $\hat{\mathbf{f}}_j = \left(\hat{f}_j^{(1)}, \ldots, \hat{f}_j^{(K)}\right)$
```

We dubbed the method proposed in this section Group Sparse Multi-Task SpAM (GSMTSpAM). The algorithm described so far assumes that the responses are continuous, however these methods can be adapted to accommodate data sets with categorical responses. When dealing with multi-category classification problems, we use a penalized local scoring algorithm as in Liu *and others* (2009). The method derived from applying the model described in Section 3 to categorical variables is dubbed Group Sparse Multi-category Additive Logistic Regression (GSMALR). In the Web Appendix A of supplementary material available at *Biostatistics* online, we present a penalization scheme that only uses a $\ell_1 \backslash \ell_2$ penalty, this is essentially the GroupSpAM applied to features of multiple responses. We dub this method Group Multi-Task SpAM (GMTSpAM), and present the details of the algorithm in the Web Appendix A of supplementary material available at *Biostatistics* online. The simulations are relegated to the Appendix of supplementary material available at *Biostatistics* online. In the simulation settings, we use a generalized cross validation scheme to find the optimal regularization parameters (α, λ). The optimal regularization parameters minimize the generalized cross validation proposed by Liu *and others* (2009). The criteria is computed as follows: $GCV(\alpha, \lambda) = \sum_{i=1}^n \sum_{k=1}^K \frac{L(\hat{\mathbf{f}})}{(n^2K^2 - nK\,\text{df}(\alpha,\lambda))}$, where $\text{df}(\alpha, \lambda) = \sum_{k=1}^K \sum_{j=1}^P \nu_j^{(k)} I(\|\hat{f}_j^{(k)}\| \neq 0)$ and $\nu_j^{(k)} = \text{trace}(S_j^{(k)})$. A 10-fold cross validation on the validation data sets can also be used. For the cross validation approach, we can select the regularization parameters that minimize the cross validated mean squared error. We generate a grid of regularization parameters. For α we pick 10 equally spaced value in the range $\alpha = (0.05, \ldots, 0.95)$ and for λ, we select among 10 equally spaced values $(0.75, \ldots, 3.0)$. This leads to a selection of a pair of parameters among a grid with a 100 combinations. For each combination (α, λ), train the model on the data associated with the validation data set, and we select the pair of values with minimal GCV. After identification of the optimal regularization parameters, we estimate the additive functions on the training data sets with those optimal parameters.

4. APPLICATION: HEALTHCARE EXPENDITURE IN THE MEDICAID SYSTEM

The objective of this case study is to identify the factors associated with variations in healthcare expenditure. Factors of interest are related to healthcare utilization, healthcare access, demographics, socio-economic, and community environment factors. The primary questions of interest are: *What are the health determinants that are associated with variations in expenditure while controlling for healthcare utilization? What are the differences between the two states?* Particular emphasis is on healthcare access as it has been at the forefront of the current health policy agenda in the United States.

We implement the application framework to data of two neighboring states, North Carolina and Georgia, for years 2005 to 2009. We compare Georgia and North Carolina because of similarities in the demographics (each with approximately 2.3 million children), economic environments (similar median household income and percentage of population below poverty level) while having different Medicaid managed care systems, Medicaid-only CMOs in Georgia, and state-based primary care managed providers (PCMP) in North Carolina as summarized by a report published by Kaiser Family Foundation. We consider only the child population insured under the Medicaid-benefits program but the models apply to other populations and insurance programs.

The relationships between expenditure and factors associated with expenditure may or may not vary throughout the years, thus we consider a different response for each year. Considering each year separately instead of studying the associative relationships to expenditure longitudinally is particularly important because of the year-to-year variations in federal and state-level health policies related to reimbursements, eligibility, and managed care.

We measure expenditure as the Medicaid payments or reimbursements per medicaid eligible member per month (PMPM) aggregated at the county-level using the Medicaid Extract (MAX) files acquired from the Centers for Medicare and Medicaid (CMS) Services for the two states and for the 5 years. In particular, we aggregate yearly expenditure by first summing all the claims associated with Medicaid eligible patients who reside within each county and then dividing this quantity by the total number of months of eligibility in the county to obtain the county level cost PMPM. In our model, we will have five responses where each response corresponds to the expenditure PMPM for all the counties in a given state. We have 100 observations per response in North Carolina and 159 observations per response in Georgia.

We will next describe the factors that will be included as covariates to explain expenditure variations for the populations in this study.

4.1. *Model covariates*

We acquired data on 40 covariates described in this section.

Healthcare utilization. Utilization measures are included in the study since they are directly linked to healthcare expenditure. Grupp-Phelan *and others* (2001); Glynn *and others* (2011) and Harrison *and others* (2012) have analyzed the relationships between utilization and cost of healthcare systems for various medical conditions; the general consensus is that higher utilization of the systems leads to higher cost.

The utilization measures include the number of claims PMPM that are associated with the type of services, for instance inpatient services, outpatient services, dental services, and miscellaneous services. The second set of utilization measures are linked to the place where the medical services were provided. These measures are the number of claims PMPM associated with hospitalizations, visits to the medical practitioner office, emergency room, and outpatient hospital. The data source for these measures is the MAX claims database.

Healthcare access. The main goal of the study is to identify predictors that can be used to intervene and recommend policies that will help reduce healthcare expenditure. The determinants of health that can be used for interventions are mainly related to access, including both financial and geographic access. The proxies used to measure financial access in our analysis are the county-level poverty rates and the percent of adults who reported that they could not see the doctor because of cost. The percent of children (under 19) without insurance is also included since lack of health insurance is a barrier to accessing health care. The geographic access consists of two measures, availability, and accessibility to pediatric primary care, where availability is defined as providers' patient volume or time available for healthcare delivery that patients experience when seeking care and accessibility is defined as the time and/or distance barriers that patients experience in reaching their providers. The estimates of these two measures are derived from Gentili *and others* (2017).

Demographics. Demographic measures have been extensively used in the literature as a control factor to study the disparities in financial and geographical access to care (Nobles *and others*, 2014). The demographics variables included in our models consist of the percentage of claims that are associated with white and non-white patients during the year of interest for the Medicaid-enrolled patients. The demographics variables also include age related variables such as the percentage of claims that are attributed to patients between the age of 0 and 5 years, between the age of 6 and 14 years, and between the age of 15 and 18 years. The data source for these variables is the MAX claims data.

Socio-economics factors. Some socio-economic factors, such as education, economic indicators, crime, and family planning related metrics, have been reported to have an impact on healthcare outcomes (Dooley *and others*, 2006; Egerter *and others*, 2009). To account for the impact of education on healthcare outcomes, we include the county level illiteracy rates, which represents the percentage of the population aged 16 and older that lacks basic literacy skills, the percentage of high school graduates or higher and the percentage of Bachelor degree or higher. The economic factors are the county level per capita income, the unemployment rate, and the percentage of household units with a mortgage with housing costs greater than 30% of income in a given county. High housing costs and high unemployment rates can be associated with poor health outcomes. Additionally, since employer-sponsored health insurance is the most prevalent coverage, unemployment can reduce access to health care. We also include crime related variables such as homicide rates, which represents the homicide rate pre 100,000 in a given county and violent crime rates. Family related social variables are the percent of single parent family households with children and the teen birth rate measured as the number of births per 1000 female population aged 15–19. The data sources for these variables are summarized in Table 3 of the Web Appendix of supplementary material available at *Biostatistics* online.

Health factors. Health factors are directly associated with the cost of health care systems. To measure county-level health status of the population, we use the percentage of the population aged 20 and older, that has a BMI greater or equal to 30 kg/m^2 (obesity rates), the diabetes rates, low birth weight defined as the percent of live births for which the infant weighted less than 2500 g. According to County Health Rankings and Roadmaps, (Paneth, 1995), low birth weight is a good predictor of infant mortality and childhood handicap, and they may lead to higher utilization of the systems by affected patients. We also use a self-reported indicator of health, which is the percent of adults reporting in a survey poor or fair health. Nutrition related variables are also added, namely the limited access to healthy foods measured as the percent of population who are low income and do not live close to a grocery store, and the percent of fast food restaurants within each counties. While most of the health factors are derived for the adult population, we believe that they reflect the health status of the population within a community. The data sources for these variables are also presented in Table 3 of the Web Appendix of supplementary material available at *Biostatistics* online.

4.2. *Results*

We apply our model to identify the most important factors and to estimate the potential non-linear relationships between healthcare expenditure and the 40 covariates considered in this study. In this analysis, we demean and scale every predictor by its mean and standard deviation. We also demean and scale the response by the global charges mean (and standard deviation) computed across years and counties. Once the additive functions are estimated, for ease of interpretation they are scaled back to the dollars amount by multiplying them with the response standard deviation.

To identify the variables that are more likely associated with healthcare expenditure in Georgia and in North Carolina, we perform a stability selection analysis. To perform the analysis we follow the steps below:

(1) Form a training data set by randomly selecting 140 counties (respectively 90 counties) out of the 159 counties (respectively 100 counties) present in Georgia (respectively North Carolina).
(2) Generate a grid of regularization parameters. For α we pick 25 equally spaced value in the range $\alpha = (0.001, \ldots, 0.1)$ and for λ we use a logarithmic scale and select 26 values $\left(10^{(-25)}, 10^{(-24)}, \ldots, 10^{(-1)}, 1\right)$.
(3) For each combination (α, λ), train the model on the data associated with the randomly selected counties.
(4) Assess if a predictor is selected for at least one of the years among the 5 years (2005–2009).
(5) Repeat the steps above 100 times.
(6) Display a heat map that shows for each predictor and for all the regularization parameters (α, λ) the number of times the predictor was included in the model.

Model selection. We perform a cross validation predictive scheme, to have a sense of the optimal values of the parameters α and λ. In Figure 7 of the Web Appendix of supplementary material available at *Biostatistics* online, we see that the values of the parameter α with the best predictions are below 0.1 for North Carolina (0.04 for Georgia). This suggests that the ℓ_1 penalty is more strongly enforced than the group penalty. This implies that when a predictor is associated with healthcare cost for 1 year, it is highly likely that it will be associated with the cost of care during the other years. Figure 12 of the Web Appendix of supplementary material available at *Biostatistics* online, shows the full regularization paths for the norm of the additive functions associated with the relevant predictors in North Carolina (with similar results for Georgia but not reported here). For all years and both states, the utilization variables such as the number of claims associated with other services, inpatient services, outpatient services, and inpatient hospitalizations are the first variables to enter in the model. Then the average county level congestion, financial access variables, and some health indicators then become influential.

Utilization factors. We find that utilization variables are the most important factors driving Medicaid healthcare expenditure at the county level in North Carolina and in Georgia.

Let's first enumerate the utilization variables that have a significant effect on cost in both states. Among the utilization variables the number of claims associated with inpatient hospitalizations, outpatient hospitalizations, and visits to the physician office have an effect on Medicaid expenditure in Georgia and in North Carolina. The number of claims issued from inpatient hospitalizations have the strongest positive relationship with the county level expenditure per Medicaid eligible. We find that a standard deviation increase in this predictor leads to an average cost increase of about \$50 in Georgia (see Figure 1) and \$18.25 in North Carolina $x \in [-2, 2]$ (see Figure 2). Similarly, we observe that as the number of claims associated with visits to the physician's office increases Medicaid health expenditure also increases by an average amount of about \$10 in Georgia (see Figure 1) and North Carolina (see Figure 2).

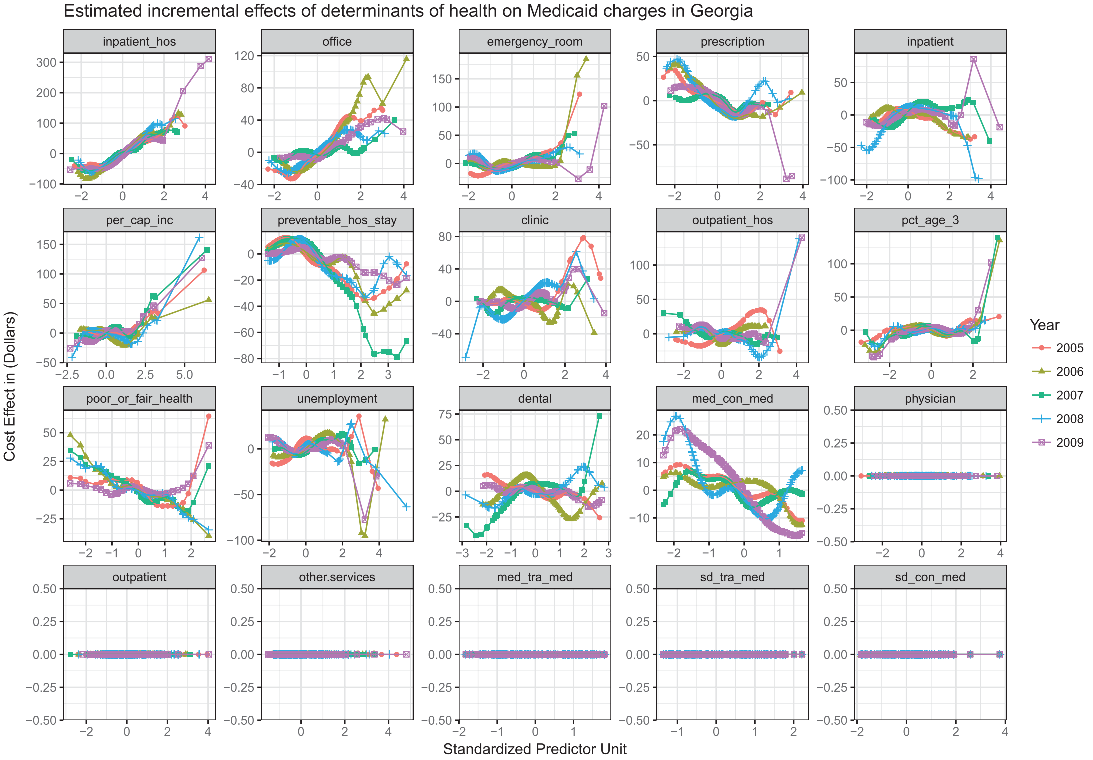

Fig. 1. Effect of determinants of health on county level Medicaid charges in Georgia.

The impact of some utilization variables is more consistent in only one state. For instance, the number of claims issued from emergency rooms or during inpatient care delivery clearly lead to increase in expenditure in Georgia; but their relationships to North Carolina Medicaid expenditure is highly non-linear and thus more ambiguous.

Some determinants of health associated with utilization tend to only be influential in one state and not the other. We find that predictors such as the number of claims issued from prescriptions, dental care visits and visits to the clinic are influential in Georgia but not in North Carolina. Services categorized as “other services” appear to be significantly associated with expenditure in North Carolina but not in Georgia.

Healthcare access. We find evidence that availability of services is associated with the healthcare expenditure in Georgia, but not in North Carolina. This association is negative in Georgia, so a decrease in congestion leads to a minor increase in expenditure if all the other predictors are held constant. Access to care measured by accessibility or travel distance has a minor positive impact on expenditure in North Carolina, but no effect in Georgia. A standard deviation increase in travel distance led on average to a $10 increase in health expenditure in 2007 and 2009 for counties with the highest travel distances ($x \in [0, 2]$, see Figure 2).

Determinants of health associated with financial access have an insignificant effect on healthcare expenditure in both states. In North Carolina, we find that the percent of those who reported not being able to see the doctor because of cost is marginally impacting cost in counties with very high or very low financial access (see Figure 2).

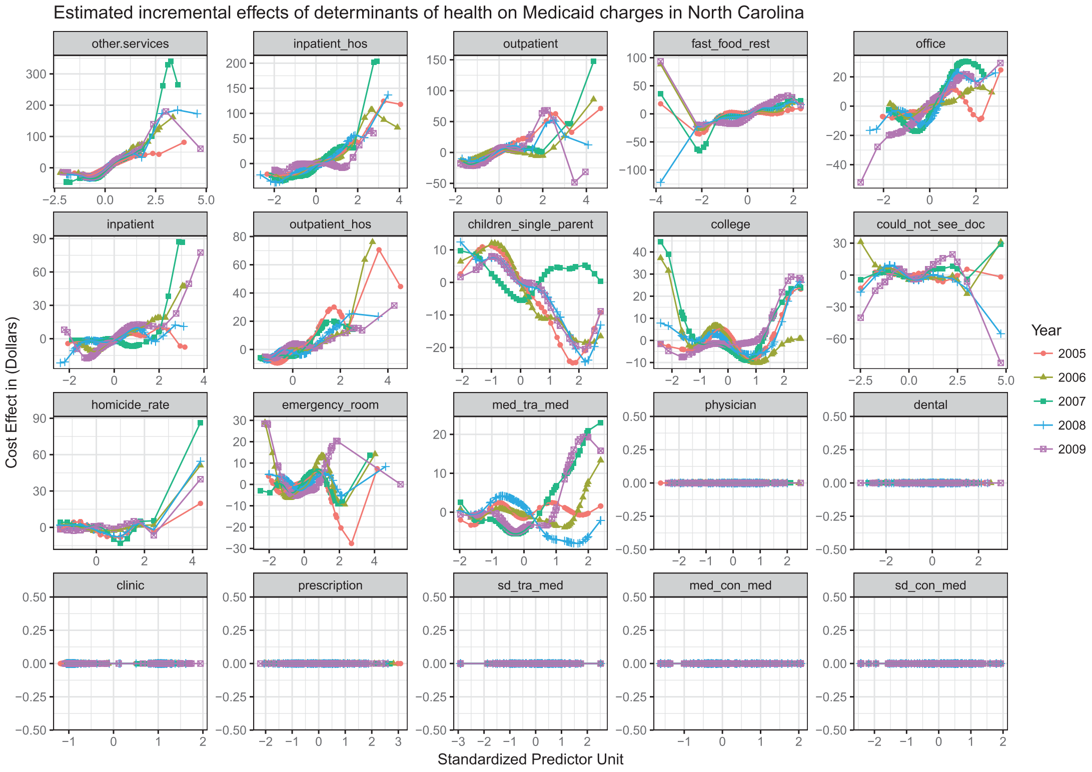

Fig. 2. Effect of determinants of health on county level Medicaid charges in North Carolina.

Demographics factors. In North Carolina, we do not find strong evidence suggesting that demographics are significantly associated with healthcare expenditure, consistently across all years. In Georgia, counties with a high percentage of teenagers (aged 15–18 years) tend to have lower healthcare expenditure (see Figure 1), and among the socio-economic factors, the percent of unemployed adults and the per capita income are significantly associated with expenditure, however they are less often included in the model than the most relevant utilization measures.

Health factors. For North Carolina, we find that an increase in the percentage of fast food restaurants is associated with a marginal increase of the county level health care expenditure (see Figure 2). And in Georgia, if the number of preventable hospital stay increases by one standard deviation, Medicaid expenditure decreases by about \$20 (Figure 1). The percentage of individuals reporting poor or fair health in Georgia also mildly influences Medicaid expenditure.

In the Web Appendix of supplementary material available at *Biostatistics* online, we display the predictors that do not affect Medicaid healthcare expenditure (see Figures 13 and 14 supplementary material available at *Biostatistics* online).

5. Concluding remarks

In this article, we have presented a new methodology for variable selection when dealing with multivariate additive non-parametric regression or classification. The methodology introduces a novel joint penalty, by combining a functional $\ell_2\backslash\ell_1$ norm with a functional ℓ_1 norm applied to additive terms in the multivariate additive regression model. By deriving the subdifferentials of these penalties, we propose a series of backfitting algorithms that can update each additive functions with a closed form solution. We illustrate the merit of the method by performing a series of simulations. In these simulations, the benefits of the group sparsity and the within group sparsity in the presence of non-linear relationships between the responses and the predictors are characterized by a better ability to select relevant variables in the presence or absence of within group sparsity. We also see in that the predictive performance of the proposed methods are superior to the sparse linear methods and SpAM (see Tables 1 and 2 supplementary material available at *Biostatistics* online).

We validated the proposed methodology using the children cancer data set introduced in Kahn *and others* (2001). The method is used to classify the small round blue cell tumors (SRBCTs) into four categories of cancers. We find that the method proposed in this article can achieve a 100 % classification rate with only 12 genes out of 2308, while method such as MT-SpAM of (Liu *and others*, 2009) achieves this classification rate with 20 genes. More details are provided in the Web Appendix G of supplementary material available at *Biostatistics* online.

We applied the proposed method to uncover factors associated with Medicaid healthcare expenditure in Georgia and in North Carolina. It is not surprising that one major factor associated with the Medicaid expenditure in both Georgia and North Carolina is the utilization of inpatient hospitalization. While controlling for this major utilization, the association of the utilization of other type of providers is not consistently significant. For example, in our study, we find that utilization of the clinics and dental services are not significantly associated with healthcare expenditure in North Carolina while it is significant for Georgia. On the other hand, utilization of other services is not statistically significant in Georgia, while it is significant in North Carolina. These findings suggest that targeted interventions for healthcare utilization will not have the same impact across all states.

While healthcare access has been spotlighted in the current health policies, we find that it has a marginal association to the Medicaid expenditure while controlling for utilization. Among the two measures of geographic access, accessibility, and availability, only availability of services has some marginal impact on expenditure in Georgia, while only accessibility has some marginal effect in North Carolina. This finding indicates that reductions in expenditure can be only be slightly impacted by policies to improve geographic access.

Moreover, demographics and health factors have also an insignificant association with expenditure except for the percentage of fast food restaurants, a health factor which has commonly been associated with high obesity prevalence.

Last, we find some non-linearities in the relationships described above, which can further help design policies that will leverage influential variables in a more targeted manner (i.e. by focusing on ranges where the costs exhibit sharp increases).

In this article, we have focused solely on estimation and did not explore inference, as this is beyond the scope of our analysis. The estimation of uncertainty and inference in high dimensional settings has recently been developed for linear methods. Assigning confidence intervals directly to our method would be inappropriate, since our estimates (just like lasso) have a non-continuous distribution that cannot easily be characterize in high dimensional settings. The current methods proposed for building confidence intervals around sparse estimates either perform a de-biasing of the original estimates or split the sample in multiple subsets to separately perform estimation then inference. It would be interesting to explore inference pertaining to the factors we have identified as drivers of medicaid costs.

Supplementary material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

FUNDING

National Science Foundation Grant CMMI-0954283.

REFERENCES

DOOLEY D., FIELDING J. AND LEVI L. (1996). Health and unemployment. *Annual Review of Public Health* **17**, 449–465.

EGERTER S., BRAVEMAN P., SADEGH-NOBARI T., GROSSMAN-KAHN R. AND DEKKER M. (2009). *Education Matters for Health*. Princeton, NJ: RWJF Commission to Build a Healthier America, Issue Brief **6**.

FOYGEL, R., HORRELL, M., DRTON, M. AND LAFFERTY, J. (2012). Nonparametric reduced rank regression. *Advances in Neural Information Processing Systems* **25**, 1628–1636.

GENTILI, M., SERBAN, N., O'CONOR, J. AND SWANN, J. (2017). Quantifying disparities in accessibility and availability of pediatric primary care with implications for policy. *Health services research* 10.1111/1475-6773.12722.

GLYNN, L. G., VALDERAS, J. M., HEALY, P., BURKE, E., NEWELL, J., GILLESPIE, P. AND MURPHY, A. W. (2011). The prevalence of multimorbidity in primary care and its effect on health care utilization and cost. *Family Practice* **28**, 516–523.

GRUPP-PHELAN, J., LOZANO, P. AND FISHMAN, P. (2001). The prevalence of multimorbidity in primary care and its effect on health care utilization and cost. *Family Practice* **38**, 363–373.

HARRISON, P. L., POPE, J. E., COBERLEY, C. R. AND RULA, E. Y. (2012). Evaluation of the relationship between individual well-being and future health care utilization and cost. *Population Health Management* **15**, 325–330.

HASTIE, T. AND TIBSHIRANI, R., (1986). Generalized additive models. *Statistical Science* **1**, 297–318.

KHAN, J., WEI, J. S., RINGNER, M., SAA, L. H., LADANYI, M., WESTERMANN, F., BERTHOLD, F., SCHWAB, M., ANTONESCU, C. R., PETERSON, C. AND MELTZER, P. S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine* **7**, 673–679.

LIN, Y. AND ZHANG, H. H. (2006). Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics* **34**, 2272–2297

LINTON, O. AND NIELSEN, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika* **82**, 93–100.

LIU, H., LAFFERTY, J. AND WASSERMAN, L. (2009). Nonparametric regression and classification with joint sparsity constraints. *Advances in Neural Information Processing Systems* **21**, 969–976.

MEIER, L., VAN DE GEER, S. AND BÜHLMANN, P. (2009). High-dimensional additive modeling. *The Annals of Statistics* **37**, 3779–3821.

NEFF, J. M., SHARP, V. L., MULDOON, J., GRAHAM, J. AND MYERS, K. (2004). Profile of medical charges for children by health status group and severity level in a Washington State health plan. *Health Services Research* **39**, 73–89.

NOBLES, M., SERBAN, N. AND SWANN, J. L. (2014). Measurement and inference on pediatric healthcare accessibility. *Annals of Applied Statistics* **8**, 1922–1946.

OBOZINSKI, G., WAINWRIGHT, M. J. AND JORDAN M. I. (2011) Union support recovery in high-dimensional multivariate regression. *Annals of Statistics* **39**, 1–47.

PANETH, N.S. (1995). The problem of low birth weight. *Future Child* **5**, 19–34.

RAVIKUMAR, P., LIU, H., LAFFERTY, J. AND WASSERMAN, L. (2009). SpAM: sparse additive models. *Journal of the Royal Statistical Society, Series B* **71**, 1009 –1030.

RESCHOVSKY, J. D., HADLEY, J., SAIONTZ-MARTINEZ, C. B. AND BOUKUS, E.R. (2011). Following the money: factors associated with the cost of treating high-cost medicare beneficiaries. *Health Services Research* **46**, 997–1021.

SIMON, N., FRIEDMAN, J., HASTIE, T., TIBSHIRANI, R., (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics* **22**, 231–245.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society, Series B* **58**, 267–288.

TIBSHIRANI, R., HASTIE, T., NARASIMHAN, B., CHU, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 6567–6572.

TURLACH, B., VENABLES, W. N. AND WRIGHT, S. J. (2005). Simultaneous variable selection. *Technometrics* **27**, 349–363.

YIN, J., CHEN X. AND XING P. E. (2012). Group sparse additive models. *Proceedings of the 29th International Conference on Machine Learning (ICML-12)* 871–878.

YUAN, M. AND LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of Royal Statistical Society, Series B* **68**, 49–67.

ZHANG, H. H., LIU, Y., WU, Y. AND ZHU, J. (2008). Variable selection for the multicategory SVM via adaptive sup-norm regularization. *Electronic Journal of Statistics* **2**, 149–1167.

[*Received March 6, 2016; revised July 31, 2017; accepted for publication August 15, 2017*]