Received: 10 July 2019 | Accepted: 7 September 2019
DOI: 10.1002/sta4.251

WILEY

ORIGINAL ARTICLE

Re-evaluating composite scores: Adaptive Lasso variable selection for non-linear models

Eli S. Kravitz[1] | **Raymond J. Carroll**[1,2]

[1]Department of Statistics, Texas A&M University, College Station, Texas 77843-3143, USA
[2]School of Mathematical and Physical Sciences, University of Technology Sydney, Sydney, New South Wales, 2006, Australia

Correspondence
Eli S. Kravitz, Department of Statistics, Texas A&M University, College Station, TX 77843-3143, USA.
Email: kravitze@tamu.edu

Funding information
National Cancer Institute, Grant/Award Number: U01-CA057030

In nutrition, epidemiology, and other public health fields, composite scores are a common tool used to assess a health behaviour. These composite scores compare an individual's health behaviour to an idealized standard and provide a number, often between 0 and 100, to indicate their compliance to a health behaviour. Crucially, this measure of health behaviour is applied across populations (gender, smoking status, etc.) and health outcomes (colon cancer, breast cancer, etc.) to create a single interpretable score. One such composite score is the 2005 Healthy Eating Index that breaks diet into 12 components and evaluates nutritional intake by adherence to these components. We provide a general method that can be used to reassess the importance of these 12 components using flexible non-linear models, across populations and diseases, based on an asymptotic least squares approximation. We establish oracle properties of this variable selection technique in our context, which is different from the usual one population and one disease context. Although our methods are motivated by the Healthy Eating Index, they are broad enough to be applied to any composite score and a broad range of non-linear models.

KEYWORDS
adaptive lasso, composite scores, least squares approximation, non-linear regression, nutritional epidemiology

1 | INTRODUCTION

In epidemiology and other public health fields, there is a need to reduce complicated behavioural patterns into simple, interpretable terms. A composite score, also referred to as an index, is commonly used to achieve this end, so as to be applied to different populations and different health outcomes. Composite scoring systems compare an individual's health behaviour to an idealized standard. Based on compliance to a set of health behaviours, an individual is assigned a score between 0 and 100. A score of 0 indicates poor compliance, and 100 is theoretical perfect behaviours.

We look at one particular challenge of working with composite scoring systems derived to be applied to multiple populations and diseases: Given an existing scoring system, are all components in the system necessary? We follow the ideas put forward by Ma, Ma, Wang, Kravitz, and Carroll (2017) and fit a logistic regression model using people from many populations who suffer from many diseases to develop a score. We include non-linear terms in our regression model, which capture the effect of the score on disease risk in a particular population. Ma et al. (2017) perform a similar analysis but include a nonparametric component in their model that we do not. The authors found that the flexibility of their semiparametric model was not needed in their real-world data analysis.

We include an adaptive Lasso penalty (Zou, 2006) to perform variable selection. The literature for the Lasso is well developed but does not apply to our particular model because of identifiability issues discussed in Section 2. Additionally, typical software packages for fitting Lasso problems such as *glmnet* (Friedman, Hastie, & Tibshirani, 2010) or least angle regression (Efron, Hastie, Johnstone, & Tibshirani, 2004) are not able to handle these non-linear models. To remedy this, we use the least squares approximation of Wang and Leng (2007). The least squares approximation allows us to translate our estimation problem into a simpler asymptotically equivalent least squares minimization. We establish that our variable selection technique chooses, asymptotically, the correct subset of components and has the optimal convergence rate. That is, it has oracle properties.

TABLE 1 Description of the Healthy Eating Index 2005 scoring system

Component	Unit	Healthy Eating Index 2005 score calculation
Total fruit	Cups	$\min(5, 5 \times (\text{density}/0.8))$
Whole fruit	Cups	$\min(5, 5 \times (\text{density}/0.4))$
Total vegetables	Cups	$\min(5, 5 \times (\text{density}/1.1))$
DOL	Cups	$\min(5, 5 \times (\text{density}/0.4))$
Total grains	Ounces	$\min(5, 5 \times (\text{density}/3))$
Whole grains	Ounces	$\min(5, 5 \times (\text{density}/1.5))$
Milk	Cups	$\min(10, 10 \times (\text{density}/1.3))$
Meat and beans	Ounces	$\min(10, 10 \times (\text{density}/2.5))$
Oil	Grams	$\min(10, 10 \times (\text{density}/12))$
Saturated fat	% of	if density $\geq$ 15 score = 0
	energy	else if density $\leq$ 7 score = 10
		else if density > 10 score = $8 - (8 \times (\text{density} - 10)/5)$
		else, score = $10 - (2 \times (\text{density} - 7)/3)$
Sodium	Milligrams	if density $\geq$ 2,000 score = 0
		else if density $\leq$ 700 score = 10
		else if density $\geq$ 1,100
		5 mm score = $8 - \{8 \times (\text{density} - 1,100)/(2,000 - 1,100)\}$
		else score = $10 - \{2 \times (\text{density} - 700)/(1,100 - 700)\}$
SoFAAS	% of	if density $\geq$ 50 score = 0
	energy	else if density $\leq$ 20 score = 20
		else score = $20 - \{20 \times (\text{density} - 20)/(50 - 20)\}$

Note. Here, also, "SoFAAS" are calories from solid fats, alcoholic beverages, and added sugars, whereas "DOL" are dark green and orange vegetables and legumes. Except for saturated fat and SoFAAS, density is obtained by multiplying usual intake by 1,000 and dividing by usual intake of kilocalories. For saturated fat, density is 9×100 usual saturated fat (grams) divided by usual calories, that is, the percentage of usual calories coming from usual saturated fat intake. For SoFAAS, the density is the percentage of usual intake that comes from usual intake of calories, that is, the division of usual intake of SoFAAS by usual intake of calories. The total Healthy Eating Index 2005 score is the sum of the individual component scores.

Although our methods are general, we apply them to the 2005 Healthy Eating Index and use the 2005 Healthy Eating Index to motivate our methods. We will refer to the 2005 Healthy Eating Index as the Healthy Eating Index, omitting the year. The Healthy Eating Index is based on the key recommendations of the 2005 Dietary Guidelines for Americans (http://www.health.gov/dietaryguidelines/dga2005/document/default.htm). The index includes ratios of interrelated dietary components to energy (caloric) intakes. The 2005 Healthy Eating Index comprises 12 distinct component scores and a total summary score. See Table 1 for a list of these components and the standards for scoring. See Guenther, Reedy, Krebs-Smith, and Reeve (2008) and Guenther, Reedy, and Krebs-Smith (2008) for details on how the Healthy Eating Index was developed and evaluated.

Intakes of each food or nutrient, represented by one of the 12 components, are expressed as a ratio to energy intake, assessed, and ascribed a score. There are other competing measures of diet such as the 2010 Healthy Eating Index (Guenther et al., 2013), the Modified Mediterranean Diet Score (Trichopoulou et al., 2005), and the MedDietScore (Panagiotakos, Pitsavos, & Stefanadis, 2006), all of which are associated with lowered mortality risk and better overall health. Our aims are (a) to suggest improvements to the dietary guidelines of the Healthy Eating Index and (b) to use model selection techniques to evaluate the relative importance of the 12 components. We find the unexpected fact that empty calories (SoFAAS) are not predictive of increased mortality risk.

2 | A NON-LINEAR MODEL ACROSS POPULATIONS

We will use the term *disease* in a generic way, until our data analysis. The term should be understood to mean a collection of health outcomes, which, for example, could be various combinations of overall mortality, mortality from various diseases, different chronic conditions, or the development of different cancers.

Denote $j = 1, \ldots, J$ as the index of the Healthy Eating Index components. There are $k = 1, \ldots, K$ populations and $\ell = 1, ..., L_K$ diseases in each population. There are $i = 1, \ldots, n_{k\ell}$ individuals be evaluated for disease ℓ in population k. The data observed are

- $Y_{ik\ell}$ is a binary indicator of disease ℓ for the ith person in population k.
- $(X_{i1}, \ldots, X_{iJ})$ is the Healthy Eating Index score for person i with components $j = 1, \ldots, J$. In the 2005 Healthy Eating Index, $J = 12$.
- For each population and disease, there may be different covariates that include terms such as age, ethnicity, education, body mass index, smoking, and physical activity. These covariates are denoted as $Z_{ik\ell}$.

To better capture the risk of a particular disease, we introduce a vector $\boldsymbol{\alpha} = (\alpha_1, \ldots \alpha_J)^T$ that allows flexible rescaling of the Healthy Eating Index components. We model the probability of a subject i of population ℓ having disease k as

$$\text{pr}(Y_{ik\ell} = 1 | X_{ikj}, Z_{ik\ell}) = H(\beta_{k\ell} \sum_{j=1}^{J} X_{ijk\ell}\alpha_j + Z_{ik\ell}^T \theta_{k\ell}) = p_{ik\ell}, \quad (1)$$

where $H(\cdot)$ is the logistic distribution function. The parameter $\boldsymbol{\alpha}$ reweighs each person's original Healthy Eating Index score, $\sum_{j=1}^{J} X_{ijk\ell}$, to create a new modified score, $\sum_j X_{ijk\ell}\alpha_j$. If a particular α_j is greater than 1, this indicates that a particular Healthy Eating Index component should be given more relative importance in the 0-to-100 score than in the Healthy Eating Index whereas a $\alpha_j < 1$ indicates that it should be given less importance. The term $\beta_{k\ell}$ allows the effect of the modified score to vary with population and disease. There is novelty in this modelling approach. We are able to provide a *single* measure of diet for every population and disease. To emphasize this, and for notational convenience, we omit the subscripts over k and ℓ in $\mathbf{X}$. This modelling approach is beneficial to public health practitioners as a single predictor, $\sum X_{ij}\alpha_j$, can be used for any disease/population of interest, and the effect of this predictor, $\beta_{l\ell}$, can be reassessed as needed.

Model (1) results in the composite loglikelihood function

$$L_n(\alpha, \beta, \theta) = \sum_{\ell} \sum_{k} \sum_{i} \{Y_{ik\ell} \log(p_{ik\ell}) + (1 - Y_{ik\ell}) \log(1 - p_{ik\ell})\}.$$

The multiplication of the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in model (1) means that they are not identifiable. To illustrate this, consider multiplying each α_j by a constant c. Then $L_n(c * \alpha, \beta, \theta) = L_n(\alpha, 1/c * \beta, \theta)$, and there is more than one value of the parameters that maximizes $L_n(\cdot)$. Informally, in model (1), the scale of $\boldsymbol{\alpha}$ can be "absorbed" by $\boldsymbol{\beta}$, for example, $\beta_{k\ell} \sum_{j=1}^{J} X_{ijk\ell} C * \alpha_j = 1/C * \beta_{k\ell} \sum_{j=1}^{J} X_{ijk\ell} * \alpha_j$.

The identifiability issue can be fixed by adding constraints to the model. Carroll, Fan, Gijbels, and Wand (1997) give identifiability constraints for single index models, which work for this non-linear model as well. A natural constraint would be to enforce that the maximum value of the new score, $\sum_j X_{ij}\alpha_j$, is 100. In the language of our previous example, $\boldsymbol{\beta}$ would not be able to "absorb" $\boldsymbol{\alpha}$ because the constraint $\sum_j X_{ij}\alpha_j = 100$ fixes the norm of $\boldsymbol{\alpha}$. We eventually use this constraint but do not impose it at first. That particular constraint makes (1) difficult to fit from a computational perspective. Instead, we begin by setting $\beta_{11} = -1$. The remaining parameters, $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and θ, are estimated in an iterative profiling procedure, first fixing $\boldsymbol{\alpha}$ and maximizing L_n with respect to $\boldsymbol{\beta}$ and θ, then fixing $\boldsymbol{\beta}$ and θ, and maximizing over $\boldsymbol{\alpha}$. This processes is repeated until convergence. Ma et al. (2017) provide guarantees that this procedure will converge to the correct value of the parameters.

Once the estimates have converged, we rescale the α coefficients, so the new score is between 0 and 100. Define $\mathbf{c}_{\text{max}} = (c_{\text{max},1}, \ldots, c_{\text{max},J})^T$as the maximum value that the original Healthy Eating Index assigns to a particular dietary component. Each element of $\boldsymbol{\alpha}$ is set to $\alpha_j^* = \alpha_j / \boldsymbol{\alpha}^T \mathbf{c}_{\text{max}}$. This puts the newly assigned score on a scale from 0 to 100, and the constraint on $\boldsymbol{\alpha}$ allows β_{11} to be estimated by refitting (1) with $\boldsymbol{\alpha}^*$ in place of $\boldsymbol{\alpha}$.

3 | VARIABLE SELECTION

3.1 | Least squares approximation and adaptive Lasso

In our context of multiple diseases and populations, we next establish which Healthy Eating Index components have no effect on disease risk. To test this, we add an adaptive Lasso penalty (Zou, 2006) to our likelihood (2). Like all Lasso-style penalties, the adaptive Lasso can perform variable selection by maximizing a likelihood function which may force some coefficients to take on a value of 0. The adaptive Lasso has a number of desirable properties that are explored in Section 3.2.

In our problem, we focus only on penalization of the $\boldsymbol{\alpha}$ parameters. In principle then, we would minimize

$$-L_n(\beta, \alpha, \theta) + \lambda \sum_{j=1}^{J} |\hat{\alpha}_{full,j}|^{-\gamma} |\alpha_j|, \quad (2)$$

with respect to $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, and θ, where λ is the tuning parameter, γ is a prespecified positive number, and $\hat{\alpha}_{full,j}$ is an estimate of α_j that has not been subject to any penalization, found by maximizing the likelihood (2). If a component in $\boldsymbol{\alpha}$ is shrunk to 0 by the adaptive Lasso method, we take this as an indication that a particular Healthy Eating Index component is unnecessary in predicting outcomes of interest.

However, in practice, there is a computational problem. Typical tools for fitting Lasso problems such as *glmnet* (Friedman et al., 2010) or least angle regression (Efron et al., 2004) are designed for standard linear and generalized linear models and do not handle the term $\sum_j X_{ij}\alpha_j$ correctly, because they cannot penalize the $\boldsymbol{\alpha}$ coefficient without also penalizing the $\boldsymbol{\beta}$ coefficient. For a conceptually simple and computationally fast solution, Wang and Leng (2007) proposed a least squares approximation for unifying computation of all Lasso models. Consider a problem with parameters $\boldsymbol{\Psi} = (\psi_1, \ldots, \psi_p)$ and a loss function $L_n(\boldsymbol{\Psi})$. Let $\widetilde{\Psi}$ be the minimizer of $L_n(\cdot)$. The authors show that any reasonable loss function, in parameters denoted as $\boldsymbol{\Psi}$,

$$L_n(\Psi) + \sum_{j=1}^{d} \lambda_j |\psi_i|,$$

can be expressed as an asymptotically equivalent least squares problem

$$Q(\Psi) = (\tilde{\Psi} - \Psi)^{\mathsf{T}} \hat{\Sigma}^{-1} (\tilde{\Psi} - \Psi) + \sum_{j=1}^{d} |\psi_j|,$$

where $\tilde{\Psi}$ is the vector that minimizes $L_n(\cdot)$ and $\hat{\Sigma}$ is an asymptotically consistent estimate of the covariance matrix of $\tilde{\Psi}$.

We approximate our penalized negative loglikelihood as

$$L_n(\Theta) + \lambda \sum_{j=1}^{J} |\hat{\alpha}_{full,j}|^{-\gamma} |\alpha_j| \approx (\tilde{\Theta} - \Theta)^{\mathsf{T}} \hat{\Sigma}^{-1} (\tilde{\Theta} - \Theta) + \lambda \sum_{j=1}^{J} |\hat{\alpha}_{full,j}|^{-\gamma} |\alpha_j|, \tag{3}$$

where $\Theta = (\beta, \alpha, \theta)$.

The least squares approximation in (3) can be solved with standard optimization software or fit as a Gaussian family adaptive Lasso model using the *glmnet* R package. Denote $\hat{\Theta}_{LSA}(\lambda)$ as the value that minimizes the right hand side of (3) as a function of λ. In our computation, we take $\gamma = 2$, a typical choice, though it can be set to any value satisfying Lemma 1 and Lemma 2 in Section 3.2.

3.2 | Model selection and oracle properties

Variable selection procedures ideally should possess the *oracle* property. Fan and Li (2001) give an overview for what it means for a selection procedure to have the oracle property. Denote $A = \{j : \Theta_j \neq 0\}$ and $\hat{A}(\lambda) = \{j : \hat{\Theta}(\lambda)_{LSA,j} \neq 0\}$. The procedure should have

- **Selection consistency**: $\text{pr}\{\hat{A}(\lambda) = A\} \to 1$.
- **Optimal estimation rate**: $\sqrt{n}(\hat{\Theta}_{\delta,\hat{A}_\delta} - \Theta_A) \to N(0, \Sigma_A)$ in distribution, where Θ_A are the nonzero components of Θ and Σ_A is the covariance matrix knowing the true subset of predictors.

We derive the selection consistency and optimal estimation rate of the least squares approximation in our problem in a similar manner as Zhang, Cao, and Carroll (2015). The following result provides the selection consistency for $\hat{\theta}_{LSA}(\lambda)$. The proof of this theorem, as well as all the proofs in Section 3.2, is provided in the appendix.

Lemma 1. *As $n \to \infty$, if $n^{1/2}\lambda \to 0$, and $n^{(1+\gamma)/2}\lambda \to \infty$, then*

$$\text{pr}\{\hat{A}(\lambda) = A\} \to 1.$$

Wang and Leng's proof of oracle properties relies extensively on what they call the *covariance assumption*. The covariance assumption specifies a strict relationship between the asymptotic covariance matrix of the full model and the asymptotic covariance matrix of an overfitted model. The exact assumption is stated as follows: Let Σ denote the variance of the limiting distribution of the parameters of the full model. Denote $\Omega = \Sigma^{-1}$, and $\Omega^{(S)}$ as the submatrix of Ω corresponding to the submodel $\mathcal{S}$. Let Σ_S denote the variance of the limiting distribution of model $\mathcal{S}$, and $\Omega_S = \Sigma_S^{-1}$. The covariance assumption states that $\Omega^{(S)} = \Omega_S$ for any overfitted $\mathcal{S}$.

The variance–covariance matrix of model (1) is fit using a sandwich estimator. A sandwich estimator has the form $\Sigma = J^{-1}HJ^{-T}$ where $J = J_n(\Theta) = \nabla L_n(\Theta)$, $H = H_n(\Theta) = \nabla^2 L_n(\Theta)$, and $J^{-T} = (J^{-1})^{\mathsf{T}}$. See the Carroll, Ruppert, Crainiceanu, and Stefanski (2006) section A.3.1 for a detailed treatment of sandwich estimators. In general, $\Sigma^{(S)} = (J^{-1}HJ^{-\mathsf{T}})^{(S)} \neq J_S^{-1}H_S J_S^{-T} = \Sigma_S$, and therefore, covariance matrices derived from sandwich estimators will not satisfy the covariance assumption of Wang and Leng.

The selection consistency of Theorem 1 does not rely on the Wang and Leng's covariance assumption, but the optimal estimation rate does. Therefore, Wang and Leng's theory will not guarantee asymptotically consistent parameter or variance estimates. However, we can get parameter and variance estimates by fitting the model (1) using only the selected components. This is explained in the following theorem.

Lemma 2. *Let A denote the set of nonzero covariates, θ_A denote these nonzero covariates, Σ_A denote the covariance matrix of the nonzero covariates, and $\hat{\theta}(A)$ denote the estimates of θ_A found by fitting the logistic model from Section 2. As $n \to \infty$, if $n^{1/2}\lambda \to 0$ and $n^{(1+\gamma)/2}\lambda \to \infty$, then*

$$\sqrt{n}\{\hat{\Theta}(A) - \Theta_A\} \to N(0, \Sigma_A).$$

The results of Lemma 1 and Lemma 2 rely the proper choice of λ. Like all Lasso methods, the least squares approximation provides a solution for any λ; however, the optimal value of λ must be selected. For finding the best fitting penalized model, Wang and Leng propose a BIC-style criterion, namely,

$$BIC(\lambda) = \{\hat{\Theta}_{\text{LSA}}(\lambda) - \tilde{\Theta}_{\text{full}}\}^{\mathsf{T}} \hat{\Sigma}^{-1} \{\hat{\Theta}_{\text{LSA}}(\lambda) - \tilde{\Theta}_{\text{full}}\} + g_n / \{n \log(n)\},$$

where g_n is the number of nonzero coefficients in $\hat{\Theta}_{LSA}(\lambda)$. Define A_{true} as the set of nonzero coefficients. The interval $(0, \infty)$ can be partitioned into three disjoint sets depending on whether $\hat{A}(\lambda)$ is overfit, underfit, or equal to the true model:

$$\mathbb{R}_- = \{\lambda \in (0,\infty) : \widehat{A}(\lambda) \subset A_{true}\},$$
$$\mathbb{R}_0 = \{\lambda \in (0,\infty) : \hat{A}(\lambda) = A_{true}\},$$
$$\mathbb{R}_+ = \{\lambda \in (0,\infty) : \widehat{A}(\lambda) \supset A_{true}, \widehat{A}(\lambda) \neq A_{true}\}.$$

Letting $\lambda^* \propto n^{-2/3}$, which satisfies Theorem 1, then we have,

$$\text{pr}\{\hat{A}(\lambda^*) = A\} \to 1.$$

Additionally, we have the following result for any $\lambda \in R_-$ and $\lambda \in R_+$.

Lemma 3. *As $n \to \infty$, if $\hat{\Sigma}$ is a consistent estimate of the variance–covariance matrix of the limiting distribution of the full model, then*

$$\text{pr}\{\inf_{\lambda \in \mathbb{R}_- \cup \mathbb{R}_+} BIC(\lambda) > BIC(\lambda^*)\} \to 1. \tag{4}$$

Theorem 3 tell us that any λ that produces the incorrect model, that is, $\lambda \in \mathbb{R}_-$ and $\lambda \in \mathbb{R}_+$, will not be selected by the BIC criterion as the optimal tuning parameter. Wang and Leng's BIC criterion is consistent in selecting the optimal tuning parameter.

4 | DATA ANALYSIS

4.1 | Background

Of particular interest to nutritionists and epidemiologists is the relationship between diet and cancer as well as diet and mortality. We conduct our analysis on the 2005 NIH-AARP Study of Diet and Health. This longitudinal study tracks incidence of lung, colorectal, prostate, breast, and ovarian cancer in adults between the ages of 51–75, as well as cause of death for those who died while the study was conducted. Table 2 lists the number of adults surveyed as well as the breakdown of cancer by men and women, and Table 3 lists mortality. The study follows mortality caused by cancer, cardiovascular disease, and all other causes of mortality for both men and women.

We consider three events of interest: cancer occurrence, mortality, and all-cause mortality. Cancer occurrence is defined as diagnosis of any of the five types of cancer in Table 2, mortality is defined as mutually exclusive outcome of one of the three causes in Table 3, and all-cause mortality is the aggregation of *any*type of mortality. We consider these outcomes separately and fit separate models for each outcome. For each outcome, the analysis is as follows.

- Model (1) is fit using all the components of the 2005 Healthy Eating Index.
- The least squares approximation with an adaptive Lasso penalty is used to identify the relevant subset of Healthy Eating Index components.
- Model (1) is refit using only components selected by the least squares approximation.

This results in three sets of selected components and three sets of parameter estimates.

TABLE 2 Summary of the NIH-AARP data for cancer occurrence

	Men		Women	
Description	# Cases	Percentage	# Cases	Percentage
Sample size	294,673		199,285	
Breast cancer			7,736	3.88
Ovarian cancer			759	0.38
Prostate cancer	23,477	7.97		
Colorectal cancer	4,693	1.59	2,291	1.15
Lung cancer	6,135	2.08	3,630	1.82

TABLE 3 Summary of the NIH-AARP data for mortality

	Men	Women
Description	# Cases	# Cases
Sample size	219,612	169,480
CVD mortality	8,112	4,028
Cancer mortality	12,247	7,344
Other cause mortality	10,821	6,547

Abbreviation: CVD, cardiovascular disease.

TABLE 4 Results from Section 4.2 when the outcome of interest is cancer occurrence, various types of mortality, and aggregated all-cause mortality

	Cancer	Mortality	All-cause mortality
Whole grains	3.98 [3.18, 4.78]	5.59 [5.04, 6.13]	5.61 [5.03, 6.18]
Total grains	1.34 [0.59, 2.09]	0.87 [0.41, 1.33]	0.93 [0.47, 1.40]
Whole fruit	1.56 [0.74, 2.37]	0.22 [−0.12, 0.57]	0.39 [0.02, 0.08]
Total fruit	2.71 [1.87, 3.55]	**0**	**0**
Total veg.	1.37 [0.47, 2.26]	2.13 [1.57, 2.68]	2.04 [1.46, 2.62]
DOL veg.	1.52 [0.65, 1.14]	0.81 [0.39, 0.68]	0.73 [0.30, 1.17]
Dairy	0.90 [0.65, 1.14]	0.53 [0.39, 0.68]	0.58 [0.43, 0.74]
Meat and beans	**0**	1.06 [0.81, 1.31]	0.98 [0.71, 1.25]
Oils	0.58 [0.30, 0.86]	0.59 [0.42, 0.77]	0.63 [0.44, 0.81]
Sodium	1.28 [0.95, 1.61]	1.96 [1.78, 2.13]	1.85 [1.66, 2.04]
Saturated fats	1.00 [0.72, 1.28]	1.04 [0.88, 1.22]	1.09 [0.91, 1.27]
Empty calories	**0**	**0**	**0**

Note. Provided estimates are found by fitting the logistic regression model from Section 2 using only the subset of components chosen by the least squares approximation. Parentheses are 95% confidence intervals. Bold 0s indicate components that are set to 0 by the least squares approximation.

The $\boldsymbol{\alpha}^*$ coefficients, which correspond to the rescaled Healthy Eating Index described in Section 2, are provided for cancer occurrence, mortality, and all-cause mortality in Table 4. The variance of the unscaled $\boldsymbol{\alpha}$ coefficients is calculated with sandwich estimator, and variance of the rescaled $\boldsymbol{\alpha}^*$ is calculated using the delta method. This derivation is provided in Appendix 0.4. We do not provide confidence intervals for components in $\boldsymbol{\alpha}^*$, which are set to 0 by the least squares approximation.

4.2 | Results

The Healthy Eating Index puts a large penalty on diets high in empty calories. Empty calories, referred to as SoFAAS in Table 1 and made up of solid fats, alcohol, and added sugars, make up 20 points of the Healthy Eating Index score. This means that someone with a diet high in empty calories will always be assigned a score below 80 regardless of their other nutritional intake. This is the largest contribution by a single component. This is in stark contrast to our results. In each analysis, the least squares approximation forces empty calories to take a value of 0. That is, we find that empty calories are not very predictive of mortality.

Total grains appear to be undervalued by the Healthy Eating Index. For example, a person receiving a perfect score of 5 for whole grains in the original Healthy Eating Index would be reassigned a score of $5.61 \times 5 = 28.05$ if all-cause mortality was of interest. Similarly, our assessment gives total vegetables over twice its original weight when predicting mortality. It is also apparent that for any kind of mortality, the 2005 Healthy Eating Index may overstate the importance of fruit in general.

To some, the 2005 Healthy Eating Index may appear more appropriate for predicting cancer than it does for predicting mortality. This possible concern may have prompted the development of the Alternative Healthy Eating Index by McCullough et al. (2002).

5 | SIMULATIONS

We justify the numeric results from Section 4.2 and the theory from Section 3.1 with two sets of simulations. The first simulation examines the stability of the estimation procedure as the sample size changes. In the second simulation, we generate data from a similar model to the data example in Section 4.

5.1 | Subset analysis

To ensure that the results presented in Section 4.2 are robust and not an artefact of the large sample size, we split the data by a factor of 1/4 and 1/8 and rerun the analysis. We present the results when all-cause mortality is the outcome of interest. We want to ensure that the estimation procedure and the variable selection procedure are roughly similar for each subset. These results are given in Table 5. The results are fairly stable.

5.2 | Variable selection and coverage

We simulate from the model

$$\text{pr}(Y_{ik\ell}|X_{ikj}, Z_{ik\ell}) = H(\beta_{k\ell} \sum_{j=1}^{J} X_{ijk\ell}\alpha_j + Z_{ik\ell}^{\mathsf{T}}\theta_{k\ell}) = p_{ik\ell}, \tag{5}$$

where there are $k = 2$ populations, $L_1 = 3$ diseases in the first population, and $L_2 = 4$ diseases in the second population. We set $\beta_1 = (-1, -0.08, -0.04)^{\mathsf{T}}$ and $\beta_2 = (-0.09, -0.06, -0.03, -0.01)^{\mathsf{T}}$. The Healthy Eating Index measurements, X_{ikj}, and covariates, $Z_{ik\ell}$, are sampled without being replaced from the NIH-AARP Study of Diet and Health. The Healthy Eating Index measurements for total fruit and whole fruit, as

	1/8 ($N = 48{,}636$)	1/4 ($N = 97{,}273$)	1 ($N = 389{,}092$)
Total grains	1.37 [−0.05, 2.8]	1.39 [0.76, 2.03]	0.93 [0.47, 1.40]
Total fruit	**0**	**0**	**0**
Whole fruit	0.29 [−0.83, 1.41]	0.47 [0.03, 0.97]	0.39 [0.46, 1.40]
Whole grains	5.26 [3.56, 6.96]	5.11 [4.34, 5.87]	5.60 [5.02, 6.18]
Total veg.	1.64 [0.13, 3.4]	2.08 [1.29, 2.87]	2.04 [1.46, 2.62]
DOL veg.	0.87 [−0.43, 2.17]	0.60 [0.01, 1.19]	0.73 [0.30, 1.17]
Dairy	0.6 [0.14, 1.06]	0.56 [0.35, 0.77]	0.58 [0.43, 0.74]
Meat and beans	0.89 [0.10,1.68]	0.97 [0.61, 1.33]	0.98 [0.71, 1.25]
Oils	0.81 [0.26, 1.36]	0.65 [0.40, 0.90]	0.63 [0.44, 0.81]
Sodium	1.9 [1.33, 2.48]	2.04 [1.79, 2.30]	1.85 [1.66, 2.04]
Saturated fats	1.08 [0.53, 1.63]	0.95 [0.71, 1.19]	1.09 [0.91, 1.27]
Empty calories	**0**	**0**	**0**

Note. The fraction refers to the proportion of the original data set used to fit the model. It is followed in parenthesis by the sample size used in the analysis. All results refer to refitting the stratified model of Section 2 using the subset of components identified by the least squares approximation. Point estimates are given, and 95% confidence intervals follow in parenthesis. Bold 0s indicate parameters set to 0 by the least squares approximation. The results are stable through across all subsets.

TABLE 5 Results for the subset analysis in Section 5.1 when all-cause mortality is the outcome of interest

Sample size	% Selected	V1	V2	V3	V4	V5
10,000	97.9	94.6	95.6	96.0	94.7	96.8
5,000	94.6	94.6	94.5	95.0	95.3	96.4
2,500	94.4	94.4	94.9	94.6	94.6	96.0
1,000	93.6	93.6	93.6	93.0	92.3	93.2

Note. The second column gives the proportion of the 1,000 simulations that identified the correct five predictor subsets. The remaining columns, V1 through V5, give the approximate coverage of 95% confidence intervals for the five nonzero predictors.

TABLE 6 Simulation results from Section 5.2

well as total grains and whole grains, are summed together to create two components representing fruit and grains. This is done because the measurements are highly correlated. This means that the dimension of $\mathbf{X}$ is 10, instead of 12 as in the real data set. We set $\boldsymbol{\alpha}$ to be a vector of length 10 with five nonzero components. The nonzero components of $\boldsymbol{\alpha}$ are set to $3, 3, 2.5, 2.5$, and 2. Two continuous covariates are selected from $\mathbf{Z}$, and the parameters θ_{kl} are drawn from uniform $[-2, 2]$ distribution. We simulate $N = 1,000$ data sets with a range of sample sizes. The γ tuning parameter is set to 2.

The theory in Section 3.1 makes two guarantees: The probability of selecting the correct subset of predictors approaches 1, and the asymptotic covariance matrix of the nonzero parameters, Σ_T, is correctly estimated. We demonstrate these claims by simulation. We test for variable selection with

$$N^{-1}\sum_{i=1}^{N}\mathbb{I}(\widehat{\alpha}_{LSA,i} = \alpha_T),$$

where $\widehat{\boldsymbol{\alpha}}_{\text{LSA},i}$ is the subset of $\boldsymbol{\alpha}$ chosen by the least squares approximation on the ith simulation, α_T is the set of true nonzero predictors, and "=" denotes set equality.

We demonstrate the asymptotic consistency of $\widehat{\Sigma}_T$ by showing that we can construct confidence intervals for the nonzero components that have proper coverage. Coverage is tested separately for each component of $\boldsymbol{\alpha}$ with

$$N^{-1}\sum_{i=1}^{N}\mathbb{I}\{\alpha_j \in (\widehat{\alpha}_j - z_{\alpha/2}\widehat{\Sigma}_{jj}^{1/2}, \widehat{\alpha}_j + z_{\alpha/2}\widehat{\Sigma}_{jj}^{1/2})\},$$

where α_j is the jth nonzero component of $\boldsymbol{\alpha}$, $\Sigma_{jj}^{1/2}$ is the standard error of $\widehat{\alpha}_j$, and $z_{\alpha/2}$ is the upper α percentile of the standard normal distribution. The estimates $\widehat{\alpha}$ and $\widehat{\Sigma}$ refer to the estimates obtained after refitting (5) with only the variables selected by the least squares approximation.

The results are given in Table 6. We point out that although the coverage probabilities are close to nominal at $n = 1,000$, consistent variable selection requires larger sample sizes. Proper variable selection is seen at $n = 10,000$, though acceptable results can be seen at smaller sample sizes. A large sample size is likely required because we use two asymptotic approximations: a sandwich estimator for the covariance matrix and the least squares approximation for variable selection. Regardless of the sample size requirements, the NIH-AARP Study of Diet and Health is more than large enough for consistent variable selection and proper confidence interval coverage.

6 | DISCUSSION

Using non-linear models and adaptive Lasso penalization, we have introduced a method to continually reassess composite score techniques. Our method produces parameter estimates and covariance estimates that are asymptotically consistent. Asymptotically, the penalization method chooses the correct subset of coefficients. Our empirical results are similar to Ma et al. (2017), though they included a nonparametric component in their model that we do not. The authors found that the flexibility of their semiparametric model was not needed in their actual data analysis. Although highly technical, it is possible to expand our analysis into their framework.

If researchers suspect that a particular composite score does not apply well to their population of interest, they use the methods outlined in this paper to reweigh the relative importance of each score component and see if each component is necessary. Analysing composite scores in this way can lead to important new finding. Our analysis of the Healthy Eating Index suggests that the negative effects of empty calories may be overstated. Similarly, the relative importance of fruit and whole grains can change dramatically when considering mortality risk instead of cancer risk.

There is considerable scope for future work. Empirically, future work should address the correlation between many of the dietary components. It is, for example, impossible to consume total grains without also consuming whole grains. The components selected and parameter estimates may change if the collinearity is addressed. The methodology in this work may be extended to variable selection while using a nonparametric or semiparameteric model for diet. Diet may be modelled with a single index model (Carroll et al., 1997), letting the reweighted sum of Healthy Index scores vary freely as in Ma et al. (2017), or with a generalized additive model (Wood, 2017), modelling each dietary component separately with a smooth function.

ACKNOWLEDGEMENT

Carroll's research was supported by a grant from the National Cancer Institute (U01-CA057030).

FINANCIAL DISCLOSURE

None reported.

CONFLICT OF INTERESTS

The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

The R programs used in the simulations of Section 5 and the data analysis of Section 4 are available by request or on Github at https://github.com/kravitzel/Composite_Scores.

We do not have permission to distribute the data involved in Section 4: Such data can be obtained via a data transfer agreement with the National Cancer Institute, see https://epi.grants.cancer.gov/Consortia/cohort_projects.html.

ORCID

Eli S. Kravitz https://orcid.org/0000-0003-4975-5239

REFERENCES

Carroll, R. J., Fan, J., Gijbels, I., & Wand, M. P. (1997). Generalized partially linear single-index models. *Journal of the American Statistical Association*, *92*(438), 477–489.

Carroll, R. J., Ruppert, D., Crainiceanu, C. M., & Stefanski, L. A. (2006). *Measurement error in nonlinear models: A modern perspective*: Chapman and Hall/CRC: New York, New York.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, *32*(2), 407–499.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1.

Guenther, P. M., Casavale, K. O., Reedy, J., Kirkpatrick, S. I., Hiza, H. A., Kuczynski, K. J., ..., & Krebs-Smith, S. M. (2013). Update of the healthy eating index: HEI-2010. *Journal of the Academy of Nutrition and Dietetics*, *113*(4), 569–580.

Guenther, P. M., Reedy, J., & Krebs-Smith, S. M. (2008). Development of the healthy eating index-2005. *Journal of the American Dietetic Association*, *108*, 1896–1901.

Guenther, P. M., Reedy, J., Krebs-Smith, S. M., & Reeve, B. B. (2008). Evaluation of the healthy eating index-2005. *Journal of the American Dietetic Association*, *108*, 1854–1864.

Ma, S., Ma, Y., Wang, Y., Kravitz, E. S., & Carroll, R. J. (2017). A semiparametric single-index risk score across populations. *Journal of the American Statistical Association*, *112*(520), 1648–1662.

McCullough, M. L., Feskanich, D., Stampfer, M. J., Giovannucci, E. L., Rimm, E. B., Hu, F. B., ..., & Willett, W. C. (2002). Diet quality and major chronic disease risk in men and women: Moving toward improved dietary guidance. *American Journal of Clinical Nutrition*, *76*(6), 1261–1271.

Panagiotakos, D. B., Pitsavos, C., & Stefanadis, C. (2006). Dietary patterns: A mediterranean diet score and its relation to clinical and biological markers of cardiovascular disease risk. *Nutrition, Metabolism and Cardiovascular Diseases*, *16*(8), 559–568.

Trichopoulou, A., Orfanos, P., Norat, T., Bueno-de Mesquita, B., Ocké, M. C., Peeters, P. H., ..., & Trichopoulos, D. (2005). Modified mediterranean diet and survival: EPIC-elderly prospective cohort study. *BMJ*, *330*(7498), 991.

Wang, H, & Leng, C (2007). Unified lasso estimation by least squares approximation. *Journal of the American Statistical Association*, *102*(479), 1039–1048.

Wood, S. N. (2017). *Generalized additive models: An introduction with R*: Chapman and Hall/CRC: New York, New York.

Zhang, X., Cao, J., & Carroll, R. J. (2015). On the selection of ordinary differential equation models with application to predator-prey dynamical models. *Biometrics*, *71*(1), 131–138.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429.

How to cite this article: Kravitz ES, Carroll RJ. Re-evaluating composite scores: Adaptive Lasso variable selection for non-linear models. *Stat*. 2019;8:e251. https://doi.org/10.1002/sta4.251

APPENDIX A: SKETCH OF TECHNICAL ARGUMENTS

A.1 | Proof of Theorem 1

By Theorem 2 of Wang and Leng,

$$\lim_{n\to\infty} \mathrm{pr}(\hat{A} \subseteq A_T) = 1, \tag{A1}$$

and by Theorem 1 of Wang and Leng,

$$\lim_{n\to\infty} \mathrm{pr}(\hat{A} \subset A_T) = 0. \tag{A2}$$

Then (A1) and (A2) imply that

$$\lim_{n\to\infty} \mathrm{pr}(\hat{A} = A_T) = 1. \tag{A3}$$

A.2 | Proof of Theorem 2

Denote $\hat{\Theta}(A)$ as the estimates from regressing Y on the subset of Θ specified by A and $\hat{\Theta}(A_T)$ as the estimates from regressing Y on the true subset of coefficients. Then (A3) means that

$$\lim_{n\to\infty} \mathrm{pr}\{\hat{\Theta}(\hat{A}) = \hat{\Theta}(A_T)\} = 1. \tag{A4}$$

We have

$$n^{1/2}\{\hat{\Theta}(A_T) - \Theta_T\} \to N(0, \Sigma_A).$$

Thus for any vector a,

$$n^{1/2}[a^{\mathsf{T}}\{\hat{\Theta}(A_T) - \Theta_T\}] \to N(0, \sigma^2).$$

where $\sigma^2 = a^{\mathsf{T}}\Sigma_A a$. Thus,

$$\mathrm{pr}[n^{1/2}a^{\mathsf{T}}\{\hat{\Theta}(A_T) - \Theta_T\}/\sigma \le z] \to \Phi(z).$$

where $\Phi(\cdot)$ is the normal cumulative distribution function. Because (A4) holds, we have

$$\mathrm{pr}[n^{1/2}a^{\mathsf{T}}\{\hat{\Theta}(\hat{A}) - \Theta_T\}/\sigma \le z] \to \Phi(z).$$

This can be expressed as

$$\begin{aligned}
&\mathrm{pr}[n^{1/2}a^{\mathsf{T}}\{\hat{\Theta}(\hat{A}) - \Theta_T\} \quad \sigma \le z, \hat{A} = A_T]\} + \mathrm{pr}[n^{1/2}a^{\mathsf{T}}\{\hat{\Theta}(\hat{A}) - \Theta_T\} \quad \sigma^2 \le z, \hat{A} \ne A_T] \\
&= \mathrm{pr}[n^{1/2}a^{\mathsf{T}}\{\hat{\Theta}(A_T) - \Theta_T\} \quad \sigma \le z, \hat{A} = A_T]\} + \mathrm{pr}[n^{1/2}a^{\mathsf{T}}\{\hat{\Theta}(\hat{A}) - \Theta_T\} \quad \sigma^2 \le z, \hat{A} \ne A_T]\} \\
&= \mathrm{pr}[n^{1/2}a^{\mathsf{T}}\{\hat{\Theta}(A_T) - \Theta_T\}/\sigma \le z, \hat{A} = A_T] + o_p(1) \\
&= \mathrm{pr}[n^{1/2}a^{\mathsf{T}}\{\hat{\Theta}(A_T) - \Theta_T\}/\sigma \le z] - \mathrm{pr}[n^{1/2}a^{\mathsf{T}}\{\hat{\Theta}(A_T) - \Theta_T\}/\sigma \le z, \hat{A} \ne A_T] + o_p(1) \\
&= \Phi(z) + o_p(1).
\end{aligned}$$

A.3 | Proof of Theorem 3

Theorem 3 is a direct application of Wang and Leng's Theorem 4.

A.4 | Variance calculation of rescaled coefficients

Let $H = (h_1, \ldots, h_J)^\mathsf{T}$, where $h_j(\boldsymbol{\alpha}) = \alpha_j/(\mathbf{c}_{\max}^\mathsf{T}\boldsymbol{\alpha})$. Define $D = (d_1, \ldots, d_J,)^\mathsf{T}$ where $d_j = (\partial h_j/\partial\alpha_1, \ldots, \partial h_j/\partial\alpha_J)$. The diagonals of matrix D are given by $D_{ii} = (\sum_{k\neq i} c_k\alpha_k)/(c_{\max}^\mathsf{T}\alpha)^2$. For $i \neq j$, $D_{ij} = (c_j\alpha_i)/(c_{\max}^\mathsf{T}\alpha)^2$. The delta method states that the covariance of H is given by $\text{cov}\{H(\alpha)\} \approx D\Sigma_\alpha D^\mathsf{T}$.

Now, we move on to calculating the variance of $\beta_{k\ell}$. For $k = 1, \ldots, K$ and $\ell = 1, \ldots, L$, the logits are

$$\beta_{k\ell}X^\mathsf{T}\alpha + Z^\mathsf{T}\theta_{k\ell},$$

with the initial constraint is that $\beta_{11} = -1$ for identifiability.

Denote $c(\boldsymbol{\alpha}) = c_{\max}^\mathsf{T}\boldsymbol{\alpha}$. After model (1) converges, replace α with $\boldsymbol{\alpha}^* = \boldsymbol{\alpha}/c(\boldsymbol{\alpha})$. Then the logits become

$$\beta_{k\ell}^* X^\mathsf{T}\alpha^* + Z^\mathsf{T}\theta_{k\ell} = \beta_{k\ell}d(\alpha)X^\mathsf{T}\alpha_* + Z^\mathsf{T}\theta_{k\ell}.$$

For $k = \ell = 1$, this means that $\beta_{k\ell}^* = -c(\boldsymbol{\alpha})$. By the delta method, $\text{var}(\widehat{\beta}_{k\ell}^*) \approx \text{cov}\{c(\widehat{\alpha})\} \approx c_\alpha(\widehat{\alpha})^\mathsf{T}\text{cov}(\widehat{\alpha})d_\alpha(\widehat{\alpha})$, where the subscript α indicates the $J \times 1$ vector of derivatives of $d(\alpha)$.

If $(k, \ell) \neq (1, 1)$, we have that $\widehat{\beta}_{k\ell}^* = \widehat{\beta}_{k\ell}c(\widehat{\alpha})$. We can express this as $\beta_{k\ell}^* = g(\beta_{k\ell}, \boldsymbol{\alpha})$ and use the delta method to get $\text{var}(\widehat{\beta}_{k\ell}^*) \approx \nabla g(\widehat{\beta}_{k\ell}, \widehat{\boldsymbol{\alpha}})^\mathsf{T}\text{cov}(\widehat{\beta}_{k\ell}, \widehat{\boldsymbol{\alpha}})\nabla g(\widehat{\beta}_{k\ell}, \widehat{\boldsymbol{\alpha}};)$. Here, $\nabla g(\widehat{\beta}_{k\ell})$ is the gradient of $g(\cdot)$ with respect to $\beta_{k\ell}$ and $\boldsymbol{\alpha}$.