Biometrika (2014), **101**, 4, *pp.* 785–797
Printed in Great Britain

doi: 10.1093/biomet/asu031
Advance Access publication 13 August 2014

Variable selection in regression with compositional covariates

BY WEI LIN, PIXU SHI, RUI FENG AND HONGZHE LI

Department of Biostatistics and Epidemiology, Perelman School of Medicine,
University of Pennsylvania, Philadelphia, Pennsylvania 19104, U.S.A.

weilin1@mail.med.upenn.edu pixushi@mail.med.upenn.edu
ruifeng@mail.med.upenn.edu hongzhe@upenn.edu

SUMMARY

Motivated by research problems arising in the analysis of gut microbiome and metagenomic data, we consider variable selection and estimation in high-dimensional regression with compositional covariates. We propose an ℓ_1 regularization method for the linear log-contrast model that respects the unique features of compositional data. We formulate the proposed procedure as a constrained convex optimization problem and introduce a coordinate descent method of multipliers for efficient computation. In the high-dimensional setting where the dimensionality grows at most exponentially with the sample size, model selection consistency and ℓ_∞ bounds for the resulting estimator are established under conditions that are mild and interpretable for compositional data. The numerical performance of our method is evaluated via simulation studies and its usefulness is illustrated by an application to a microbiome study relating human body mass index to gut microbiome composition.

Some key words: Compositional data; Coordinate descent method of multipliers; High-dimensional regression; Lasso; Log-contrast model; Model selection; Regularization; Sparsity.

1. INTRODUCTION

Compositional data, which consist of the proportions or percentages of a composition, appear frequently in a wide range of applications; examples include geochemical compositions of rocks in geology, household patterns of expenditure in economics, species compositions of biological communities in ecology, and topic compositions of documents in machine learning. The fact that the components of a composition must sum to unity renders many standard multivariate statistical methods inappropriate or inapplicable. Since the seminal work of Aitchison (1982), methodological developments for compositional data analysis have given rise to fruitful research, thoroughly surveyed by Aitchison (2003). The increasing availability of large compositional datasets, whose dimensionality is comparable to or much larger than the sample size, poses new challenges to existing methodology. However, little formal effort has been made to develop principled tools of analysis for such data. A typical example arises in metagenomic studies of microbial communities based on 16S rRNA gene sequencing, where the relative abundances of hundreds to thousands of bacterial taxa on a few tens to hundreds of individuals are available for analysis; see, for example, Chen & Li (2013).

The aim of this paper is to address the variable selection problem in high-dimensional regression with compositional covariates. To mitigate the difficulties associated with high dimensionality, it is crucial to select parsimonious models that tend to improve the performance of statistical procedures and the interpretability of the resulting inferences. Regularization methods for simultaneous variable selection and estimation in linear regression and more general contexts have received intense interest recently. In particular, the ℓ_1 regularization or lasso approach (Tibshirani, 1996) has enjoyed widespread popularity, and its theoretical properties in high-dimensional regression are now well-understood; see, for example, Bühlmann & van de Geer (2011) for an overview. Owing to the special nature of compositional data, however, the usual linear regression model is inappropriate for our purposes. In this paper we consider the linear log-contrast model of Aitchison & Bacon-Shone (1984), which is particularly useful for regression analysis with compositional covariates. Under this model, the expected response does not depend on the basis counts from which a composition is obtained. This is the case for our microbiome data example, where the number of sequencing reads varies drastically across samples and should not play a role in predicting the response of interest.

We propose an ℓ_1 regularization method for variable selection and estimation in high-dimensional linear log-contrast models. We formulate the proposed procedure as a constrained convex optimization problem, develop efficient algorithms for computation, and provide strong theoretical guarantees. Since the constraint in the problem couples the parameters, coordinate descent methods for solving ℓ_1-regularized least-squares problems (Friedman et al., 2007) are not directly applicable. We therefore combine coordinate descent with the method of multipliers to introduce an efficient algorithm for solving the optimization problem. To establish model selection consistency and ℓ_∞ bounds for the resulting estimator, we impose conditions analogous to the irrepresentability condition for linear regression of Zhao & Yu (2006). Our conditions, however, differ from those for linear regression models in important ways, which account for the compositional effect and adapt well to the dependence structure of compositional data.

2. Variable selection in the linear log-contrast model

Log-contrast models were originally introduced by Aitchison & Bacon-Shone (1984) for modelling experiments with mixtures, and have proved to be useful for a wide variety of regression problems with a composition playing the role of covariate. Suppose that we observe an n-vector y of responses and an $n \times p$ matrix $X = (x_{ij})$ of covariates, with each row of X lying in the $(p-1)$-dimensional positive simplex $S^{p-1} = \{(x_1, \ldots, x_p) : x_j > 0$ $(j = 1, \ldots, p),\ \sum_{j=1}^p x_j = 1\}$. Because of the unit-sum constraint, the p components of a composition cannot vary freely; therefore traditional methodology often requires the omission of certain components to ensure identifiability, and so encounters intrinsic difficulties in providing sensible interpretations for the regression parameters. To resolve these difficulties, Aitchison & Bacon-Shone (1984) proposed applying the log-ratio transformation (Aitchison, 1982) to compositional covariates, resulting in the linear log-contrast model

$$y = Z^p \beta^*_{\backslash p} + \varepsilon, \tag{1}$$

where $Z^p = \{\log(x_{ij}/x_{ip})\}$ is the $n \times (p-1)$ log-ratio matrix whose pth component is the reference component, $\beta^*_{\backslash p} = (\beta^*_1, \ldots, \beta^*_{p-1})^{\mathrm{T}}$ is the corresponding $(p-1)$-vector of regression coefficients, and ε is an n-vector of independent noise distributed as $N(0, \sigma^2)$. By introducing a new coefficient $\beta^*_p = -\sum_{j=1}^{p-1} \beta^*_j$, model (1) can be more conveniently expressed in the symmetric form

$$y = Z\beta^* + \varepsilon, \quad \sum_{j=1}^p \beta^*_j = 0, \tag{2}$$

where $Z = (z_1, \ldots, z_p) = (\log x_{ij})$ is the $n \times p$ design matrix and $\beta^* = (\beta^*_1, \ldots, \beta^*_p)^{\mathrm{T}}$ is the p-vector of regression coefficients. We do not include an intercept in the model, since it can be eliminated by centring the response and predictor variables. We are concerned with the high-dimensional sparse setting, where the dimensionality p is comparable to or much larger than the sample size n, while only a few of the regression coefficients are nonzero.

Applying the ℓ_1 regularization approach to model (2), we consider the constrained convex optimization problem

$$\hat{\beta} = \arg\min_{\beta} \left(\frac{1}{2n} \|y - Z\beta\|_2^2 + \lambda \|\beta\|_1 \right), \quad \text{subject to} \sum_{j=1}^{p} \beta_j = 0, \tag{3}$$

where $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$, $\lambda > 0$ is a regularization parameter, and $\|\cdot\|_2$ and $\|\cdot\|_1$ denote the ℓ_2 and ℓ_1 norms, respectively. The zero-sum constraint in problem (3) is crucial for the resulting estimator to enjoy interpretive advantages over a standard lasso estimator. Specifically, the proposed estimator possesses the following desirable properties.

Property 1. Scale invariance: the estimator is unchanged under the transformation $X \mapsto TX$ for an arbitrary diagonal matrix $T = \mathrm{diag}(t_1, \ldots, t_n)$ with all $t_i > 0$.

Property 2. Permutation invariance: the estimator is invariant under any permutation π of the p components, meaning that it is unchanged if π is applied to both the columns of X and the components of $\hat{\beta}$.

Property 3. Selection invariance: the estimator is unchanged if one knew in advance which components would be estimated as zero and applied the procedure to the subcomposition formed by the remaining components.

Properties 1 and 3 are due to the zero-sum constraint; they ensure that the inferences are independent of an arbitrary scaling of the basis from which a composition is obtained, and remain unaffected by correctly excluding some or all of the zero components. Property 2 is immediately seen from the symmetric formulation of problem (3), but would not be guaranteed by first transforming the p components into a $(p-1)$-dimensional feature space and then applying a standard variable selection procedure.

Upon eliminating the constraint by using $\beta_p = -\sum_{j=1}^{p-1} \beta_j$, we can rewrite problem (3) as the unconstrained problem

$$\hat{\beta}_{\backslash p} = \arg\min_{\beta_{\backslash p}} \left(\frac{1}{2n} \|y - Z^p \beta_{\backslash p}\|_2^2 + \lambda \|D\beta_{\backslash p}\|_1 \right),$$

where $\beta_{\backslash p} = (\beta_1, \ldots, \beta_{p-1})^{\mathrm{T}}$ and $D = (I_{p-1}, -1_p)^{\mathrm{T}} \in \mathbb{R}^{p \times (p-1)}$, with I_r and 1_r denoting the $r \times r$ identity matrix and the r-vector of ones, respectively. This asymmetric form can be recognized as an instance of the generalized lasso problem considered by Tibshirani & Taylor (2011), but existing results do not specialize in our case to give an appropriate algorithm or theory for several reasons. First, eliminating one arbitrary component and applying a generic algorithm to the $(p-1)$-dimensional problem generally does not yield numerical solutions that are permutation invariant. Second, a coordinate descent algorithm that is fast and applicable to a prespecified set of λ values is not yet available. Third, theory for the generalized lasso problem does not provide useful insights into the compositional constraint and its effect on variable selection. All these limitations call for the development of computational methods and theoretical results that are relevant in particular to the analysis of compositional data.

3. Computation

3·1. *Optimization algorithm*

Coordinate descent algorithms have been shown to be very efficient for solving large-scale ℓ_1 regularization problems (Friedman et al., 2007). They are not directly applicable to problem (3), however, because the nondifferentiable ℓ_1 terms are inseparable under the zero-sum constraint. Here we propose an efficient, easily implemented algorithm based on an iterative modification of coordinate descent that involves combining it with the method of multipliers or the augmented Lagrangian method (Bertsekas, 1996) to deal with the constraint.

To derive the algorithm, we first form the augmented Lagrangian for problem (3),

$$L_\mu(\beta, \gamma) = \frac{1}{2n}\|y - Z\beta\|_2^2 + \lambda\|\beta\|_1 + \gamma\sum_{j=1}^p \beta_j + \frac{\mu}{2}\left(\sum_{j=1}^p \beta_j\right)^2,$$

where γ is the Lagrange multiplier and $\mu > 0$ is a penalty parameter. The method of multipliers for problem (3) consists of the iterations

$$\beta^{k+1} \leftarrow \arg\min_\beta L_\mu(\beta, \gamma^k), \qquad \gamma^{k+1} \leftarrow \gamma^k + \mu\sum_{j=1}^p \beta_j^{k+1}.$$

Define the scaled Lagrange multiplier $\alpha = \gamma/\mu$. Then the above iterations can be more conveniently expressed as

$$\beta^{k+1} \leftarrow \arg\min_\beta \left\{\frac{1}{2n}\|y - Z\beta\|_2^2 + \lambda\|\beta\|_1 + \frac{\mu}{2}\left(\sum_{j=1}^p \beta_j + \alpha^k\right)^2\right\}, \tag{4}$$

$$\alpha^{k+1} \leftarrow \alpha^k + \sum_{j=1}^p \beta_j^{k+1}. \tag{5}$$

Now the ℓ_1 terms in (4) are separable and the subproblem can be solved by coordinate descent. With the other components held fixed, the jth component of β is updated by

$$\beta_j^{k+1} \leftarrow \frac{1}{v_j + \mu} S_\lambda\left\{\frac{1}{n}z_j^{\mathrm{T}}\left(y - \sum_{i \neq j} \beta_i^{k+1} z_i\right) - \mu\left(\sum_{i \neq j} \beta_i^{k+1} + \alpha^k\right)\right\}, \tag{6}$$

where $v_j = \|z_j\|_2^2/n$ and $S_\lambda(t) = \mathrm{sgn}(t)(|t| - \lambda)_+$ is the soft thresholding operator. Combining (4)–(6) yields the following coordinate descent method of multipliers for solving problem (3).

Algorithm 1. Coordinate descent method of multipliers

Step 1. Initialize β^0 with 0 or a warm start, $\alpha^0 = 0$, $\mu > 0$ and $k = 0$.

Step 2. For $j = 1, \ldots, p, 1, \ldots, p, \ldots$, update β_j^{k+1} by (6) until convergence.

Step 3. Update α^{k+1} by (5).

Step 4. Update $k \leftarrow k+1$, and repeat Steps 2 and 3 until convergence. Output $\hat{\beta} = \beta^{k+1}$.

The minimization of subproblem (4), which is carried out in Step 2 of Algorithm 1, need not be exact; it suffices to adopt a stopping criterion such that the minimization is asymptotically exact. This results in a more efficient algorithm for which convergence is still ensured. We have the following result regarding the convergence of Algorithm 1 with inexact minimization.

Proposition 1. *Assume that Step* 2 *of Algorithm* 1 *finds at iteration k an approximate minimizer β^{k+1} such that $L_\mu(\beta^{k+1}, \gamma^k) \leqslant \min_\beta L_\mu(\beta, \gamma^k) + \delta_k$ for all k, where $\delta_k \geqslant 0$ and $\sum_{k=0}^{\infty} \surd\delta_k < \infty$. Then the sequence $\{\beta^k\}$ generated by Algorithm* 1 *is bounded. Moreover, every cluster point of $\{\beta^k\}$ is an optimal solution of problem* (3).

3·2. *Tuning parameter selection*

The regularization parameter λ can be selected by the generalized information criterion for high-dimensional penalized likelihood proposed by Fan & Tang (2013). They showed that the criterion with a uniform choice of the model complexity penalty identifies the true model with probability tending to 1 when the dimensionality p grows at most exponentially with the sample size n. For model (2) and our regularization method, we define

$$\textsc{gic}(\lambda) = \log \hat{\sigma}_\lambda^2 + (s_\lambda - 1)\,\frac{\log\log n}{n}\,\log(p \vee n),$$

where $\hat{\sigma}_\lambda^2 = \|y - Z\hat{\beta}_\lambda\|_2^2/n$, $\hat{\beta}_\lambda$ is the regularized estimator, $p \vee n = \max(p, n)$, and s_λ is the number of nonzero coefficients in $\hat{\beta}_\lambda$. Because of the zero-sum constraint, the effective number of free parameters is $s_\lambda - 1$ for $s_\lambda \geqslant 2$. We then select the optimal λ by minimizing $\textsc{gic}(\lambda)$. Alternatively, one can apply K-fold crossvalidation with $K = 5$ or 10 to choose λ, which tends to select a larger model and trades off between model selection consistency and prediction accuracy. Although crossvalidation is computationally more expensive, it is less parsimonious and can often yield more satisfactory performance in practice.

The penalty parameter μ that is needed to enforce the zero-sum constraint does not affect the convergence of Algorithm 1 as long as $\mu > 0$, and we take $\mu = 1$ in all computations.

4. Theoretical properties

We establish model selection consistency and ℓ_∞ bounds for the proposed estimator under deterministic designs. We first introduce some notation. Let Z^r denote the log-ratio matrix with the rth component taken as the reference component, and let $C^r = n^{-1}(Z^r)^{\mathrm{T}} Z^r$ be the corresponding sample log-ratio covariance matrix. Let $S = \{j : \beta_j^* \neq 0\}$ denote the support of β^* and $s = |S|$ the cardinality of S. For any subset $J \subset \{1, \ldots, p\}$ and $j \in J$, denote by J^{c} the complement of J and define $J_{\setminus j} = J \setminus \{j\}$. We will use subsets to index a vector or matrix; for example, $C^r_{S^{\mathrm{c}} S_{\setminus r}}$ is the submatrix formed by the (i, j)th entries of C^r with $i \in S^{\mathrm{c}}$ and $j \in S_{\setminus r}$. Define $\beta_{\min} = \min_{j \in S} |\beta_j^*|$, the minimum signal. Let $\|\cdot\|_\infty$ denote the ℓ_∞ or matrix ∞-norm, i.e., $\|A\|_\infty = \max_i \sum_j |a_{ij}|$ for a matrix $A = (a_{ij})$.

We assume without loss of generality that $p \in S$. Central to guaranteed support recovery through our ℓ_1 regularization method is the following condition.

Condition 1. There exists some $\xi \in (0, 1]$ such that

$$\left\| C^{p}_{S^{c} S_{\backslash p}} (C^{p}_{S_{\backslash p} S_{\backslash p}})^{-1} \{\mathrm{sgn}(\beta^{*}_{S_{\backslash p}}) - \mathrm{sgn}(\beta^{*}_{p}) 1_{s-1}\} + \mathrm{sgn}(\beta^{*}_{p}) 1_{p-s} \right\|_{\infty} \leqslant 1 - \xi. \tag{7}$$

Also, our assumption for the minimum signal threshold involves the quantity φ defined by

$$\varphi = \| D_{S S_{\backslash p}} (C^{p}_{S_{\backslash p} S_{\backslash p}})^{-1} (D_{S S_{\backslash p}})^{\mathrm{T}} \|_{\infty}.$$

Although the definitions of ξ and φ seem to depend on the choice of the reference component, we show that this is not the case. Let D^{r} denote the matrix formed by interchanging the rth and pth rows of D. The following proposition asserts the permutation invariance of ξ and φ.

Proposition 2. *For every $r \in S_{\backslash p}$, we have*

$$\begin{aligned} C^{r}_{S^{c} S_{\backslash r}} (C^{r}_{S_{\backslash r} S_{\backslash r}})^{-1} \{\mathrm{sgn}(\beta^{*}_{S_{\backslash r}}) - \mathrm{sgn}(\beta^{*}_{r}) 1_{s-1}\} + \mathrm{sgn}(\beta^{*}_{r}) 1_{p-s} \\ = C^{p}_{S^{c} S_{\backslash p}} (C^{p}_{S_{\backslash p} S_{\backslash p}})^{-1} \{\mathrm{sgn}(\beta^{*}_{S_{\backslash p}}) - \mathrm{sgn}(\beta^{*}_{p}) 1_{s-1}\} + \mathrm{sgn}(\beta^{*}_{p}) 1_{p-s} \end{aligned} \tag{8}$$

and

$$D^{r}_{S S_{\backslash r}} (C^{r}_{S_{\backslash r} S_{\backslash r}})^{-1} (D^{r}_{S S_{\backslash p}})^{\mathrm{T}} = D_{S S_{\backslash p}} (C^{p}_{S_{\backslash p} S_{\backslash p}})^{-1} (D_{S S_{\backslash p}})^{\mathrm{T}}. \tag{9}$$

Condition 1 is in the spirit of the irrepresentability condition for linear regression in Zhao & Yu (2006), though important differences exist. It is worthwhile to compare Condition 1 with its counterparts for two usual lasso estimators:

(i) the condition

$$\| C^{p}_{S^{c} S_{\backslash p}} (C^{p}_{S_{\backslash p} S_{\backslash p}})^{-1} \mathrm{sgn}(\beta^{*}_{S_{\backslash p}}) \|_{\infty} \leqslant 1 - \xi \tag{10}$$

for the lasso problem

$$\hat{\beta}^{(\mathrm{i})}_{\backslash p} = \arg\min_{\beta_{\backslash p}} \left(\frac{1}{2n} \| y - Z^{p} \beta_{\backslash p} \|_{2}^{2} + \lambda \| \beta_{\backslash p} \|_{1} \right), \tag{11}$$

which is a direct application of lasso to model (1);

(ii) the condition

$$\| C_{S^{c} S} (C_{SS})^{-1} \mathrm{sgn}(\beta^{*}_{S}) \|_{\infty} \leqslant 1 - \xi, \tag{12}$$

where $C = n^{-1} Z^{\mathrm{T}} Z$, for the lasso problem

$$\hat{\beta}^{(\mathrm{ii})} = \arg\min_{\beta} \left(\frac{1}{2n} \| y - Z\beta \|_{2}^{2} + \lambda \| \beta \|_{1} \right), \tag{13}$$

which simply ignores the zero-sum constraint in problem (3).

Expression (10) lacks the permutation invariance of Condition 1, reflecting the fact that the pth component is not regularized in problem (11) and hence no recovery guarantees can be provided. Condition (12) is ideally suited to nearly orthogonal designs but would be problematic for designs with generally negative correlations, such as those common in compositional data analysis. In contrast, the extra term $\mathrm{sgn}(\beta^{*}_{p}) 1_{p-s}$ in Condition 1 enables it to adapt well to the negative correlations resulting from the compositional constraint.

To develop further intuition with Condition 1, we consider an illustrative example where the covariate matrix X is generated from an orthogonal design $W = (w_{ij})$ with $W^{\mathrm{T}}W = nI$ via the transformation $x_{ij} = \exp(w_{ij}) / \sum_{k=1}^{p} \exp(w_{ik})$. This represents an extreme case where the dependence among the components is purely due to the unit-sum constraint. In this example, we have $C^p = n^{-1}(Z^p)^{\mathrm{T}} Z^p = n^{-1} D^{\mathrm{T}} W^{\mathrm{T}} W D = D^{\mathrm{T}} D = I_{p-1} + 1_{p-1} 1_{p-1}^{\mathrm{T}}$, and then

$$C^p_{S^{\mathrm{c}} S_{\setminus p}} = 1_{p-s} 1_{s-1}^{\mathrm{T}}, \quad (C^p_{S_{\setminus p} S_{\setminus p}})^{-1} = (I_{s-1} + 1_{s-1} 1_{s-1}^{\mathrm{T}})^{-1} = I_{s-1} - s^{-1} 1_{s-1} 1_{s-1}^{\mathrm{T}}.$$

Some straightforward calculation yields that the left-hand side of (7) equals

$$s^{-1} \left| 1_s^{\mathrm{T}} \operatorname{sgn}(\beta_S^*) \right| \leqslant (s-2)/s < 1,$$

where the first inequality is due to the constraint $1_s^{\mathrm{T}} \beta_S^* = 0$. This implies that Condition 1 holds, and ξ can be taken close to 1 provided that the signals are nearly evenly divided between positive and negative signs.

We are now ready to state our main result concerning the model selection consistency of the proposed estimator. We assume without loss of generality that the columns of Z are normalized such that $\max_j \|z_j\|_2 \leqslant \surd n$.

Theorem 1. *Assume that Condition* 1 *holds, the regularization parameter* λ *satisfies* $\lambda \geqslant c_1 \sigma \{(\log p)/n\}^{1/2} / \xi$ *for some constant* $c_1 > 2\surd 2$, *and the minimum signal satisfies* $\beta_{\min} > 3\varphi\lambda/2$. *Then, with probability at least* $1 - p^{-c_2}$ *for some constant* $c_2 > 0$, *problem* (3) *has an optimal solution* $\hat\beta$ *that satisfies the following properties:*

(i) *sign consistency, i.e.,* $\operatorname{sgn}(\hat\beta) = \operatorname{sgn}(\beta^*)$;
(ii) ℓ_∞ *loss, i.e.,* $\|\hat\beta_S - \beta_S^*\|_\infty \leqslant 3\varphi\lambda/2$.

To understand the asymptotic implications of Theorem 1, assume for simplicity that ξ and φ are constants. Then Theorem 1 implies that the proposed estimator is model selection consistent and uniformly estimation consistent as long as $\log p = o(n)$. Taking the smallest possible λ, we have the convergence rate $\|\hat\beta_S - \beta_S\|_\infty = O_{\mathrm{p}}[\{(\log p)/n\}^{1/2}]$. The rates we have derived here parallel those for the usual lasso estimator (Wainwright, 2009) but are established under a different form of the irrepresentability condition, which explicitly takes the zero-sum constraint into account.

5. Numerical studies

5·1. *Simulations*

We conducted simulation studies to compare the numerical performance of the proposed method with the two usual lasso estimators defined in (11) and (13), which we refer to as lasso (i) and lasso (ii), respectively. In lasso (i), the reference component is chosen at random from the p components, and after $\hat\beta^{(\mathrm{i})}_{\setminus p}$ is obtained we let $\hat\beta^{(\mathrm{i})}_p = -1^{\mathrm{T}} \hat\beta^{(\mathrm{i})}_{\setminus p}$. Lasso (i) and the proposed estimator satisfy the zero-sum constraint, whereas lasso (ii) does not.

We generated the covariate data in the following way. We first generated an $n \times p$ data matrix $W = (w_{ij})$ from a multivariate normal distribution $N_p(\theta, \Sigma)$, and then obtained the covariate matrix $X = (x_{ij})$ by the transformation $x_{ij} = \exp(w_{ij}) / \sum_{k=1}^{p} \exp(w_{ik})$. The covariates generated thus follow a logistic normal distribution (Aitchison & Shen, 1980). To reflect the fact that the components of a composition in metagenomic data often differ by orders of magnitude, we let $\theta = (\theta_j)$ with $\theta_j = \log(0{\cdot}5p)$ for $j = 1, \ldots, 5$ and $\theta_j = 0$ otherwise. To describe different levels

Table 1. *Means and standard errors* (*in parentheses*) *of various performance measures for three methods based on* 100 *simulations*

(n, p)	Method	PE	ℓ_1 loss	ℓ_2 loss	ℓ_∞ loss	FP	FN
				$\rho = 0{\cdot}2$			
(50, 30)	Lasso (i)	0·43 (0·01)	1·16 (0·03)	0·19 (0·01)	0·25 (0·01)	5·44 (0·29)	0·00 (0·00)
	Lasso (ii)	0·42 (0·01)	1·10 (0·03)	0·19 (0·01)	0·25 (0·01)	4·15 (0·28)	0·00 (0·00)
	Proposed	0·42 (0·01)	1·05 (0·03)	0·18 (0·01)	0·24 (0·01)	3·57 (0·23)	0·00 (0·00)
(100, 200)	Lasso (i)	0·45 (0·01)	1·25 (0·03)	0·24 (0·01)	0·27 (0·01)	4·94 (0·28)	0·00 (0·00)
	Lasso (ii)	0·42 (0·01)	1·12 (0·02)	0·21 (0·01)	0·26 (0·01)	2·96 (0·23)	0·00 (0·00)
	Proposed	0·41 (0·01)	1·07 (0·02)	0·19 (0·01)	0·24 (0·01)	3·03 (0·24)	0·00 (0·00)
(100, 1000)	Lasso (i)	0·82 (0·05)	2·01 (0·07)	0·69 (0·06)	0·42 (0·02)	5·18 (0·27)	0·28 (0·08)
	Lasso (ii)	0·66 (0·03)	1·68 (0·05)	0·52 (0·03)	0·38 (0·01)	2·84 (0·21)	0·13 (0·04)
	Proposed	0·61 (0·02)	1·57 (0·04)	0·43 (0·03)	0·34 (0·01)	3·10 (0·22)	0·04 (0·02)
				$\rho = 0{\cdot}5$			
(50, 30)	Lasso (i)	0·46 (0·01)	1·55 (0·05)	0·35 (0·03)	0·33 (0·01)	6·70 (0·32)	0·08 (0·04)
	Lasso (ii)	0·43 (0·01)	1·40 (0·04)	0·30 (0·02)	0·31 (0·01)	5·00 (0·30)	0·02 (0·01)
	Proposed	0·42 (0·01)	1·32 (0·04)	0·28 (0·02)	0·30 (0·01)	4·81 (0·27)	0·02 (0·01)
(100, 200)	Lasso (i)	0·62 (0·03)	2·11 (0·07)	0·75 (0·06)	0·48 (0·02)	7·16 (0·39)	0·33 (0·08)
	Lasso (ii)	0·47 (0·01)	1·60 (0·04)	0·45 (0·03)	0·37 (0·01)	4·61 (0·27)	0·09 (0·05)
	Proposed	0·45 (0·01)	1·54 (0·03)	0·40 (0·02)	0·36 (0·01)	4·60 (0·29)	0·01 (0·01)
(100, 1000)	Lasso (i)	1·51 (0·08)	3·70 (0·08)	2·32 (0·11)	0·81 (0·02)	3·39 (0·23)	2·55 (0·12)
	Lasso (ii)	0·94 (0·05)	2·72 (0·08)	1·40 (0·08)	0·62 (0·02)	2·44 (0·20)	1·29 (0·13)
	Proposed	0·91 (0·07)	2·59 (0·08)	1·25 (0·09)	0·59 (0·02)	3·73 (0·29)	0·99 (0·13)

PE, prediction error; FP, number of false positives; FN, number of false negatives.

of correlations among the components, we let $\Sigma = (\rho^{|i-j|})$ with $\rho = 0{\cdot}2$ or $0{\cdot}5$. We generated the responses according to model (2) with $\beta^* = (1, -0{\cdot}8, 0{\cdot}6, 0, 0, -1{\cdot}5, -0{\cdot}5, 1{\cdot}2, 0, \ldots, 0)^{\mathrm{T}}$ and $\sigma = 0{\cdot}5$, so that three of the six nonzero coefficients were among the five major components and the rest were among the minor components.

We set $(n, p) = (50, 30)$, $(100, 200)$ and $(100, 1000)$, and repeated 100 simulations for each setting. The tuning parameter λ was selected by the generalized information criterion as described in § 3·2. We used six performance measures for our comparisons. The prediction error $\|y - Z\hat{\beta}\|_2^2/n$ was computed from an independent test sample of size n. The estimation accuracy was assessed by the ℓ_q losses $\|\hat{\beta} - \beta^*\|_q$ with $q = 1, 2$ and ∞. Two variable selection measures were the number of false positives and the number of false negatives, where positives and negatives refer to nonzero and zero coefficients, respectively. The means and standard errors of these performance measures for the three methods are reported in Table 1.

As seen from Table 1, the lasso (i) estimator has inferior performance in almost all settings, since the reference component is not regularized and is always included in the selected model. The lasso (ii) estimator performs better than lasso (i), but always violates the zero-sum constraint in finite samples. The proposed estimator performs slightly better than lasso (ii) in terms of prediction and estimation. The variable selection performance of the proposed estimator is comparable to that of lasso (ii) with low to moderate dimensionality, but it tends to select fewer false negatives at the cost of slightly increased false positives in high dimensions. This is reasonable because the omission of important variables is more influential than the inclusion of unimportant variables with shrunk coefficients. A potential remedy for the violation of the zero-sum constraint of the lasso (ii) estimator would be to refit the unpenalized linear log-contrast model with the constraint using the selected variables; this approach would also be useful for reducing the bias caused by

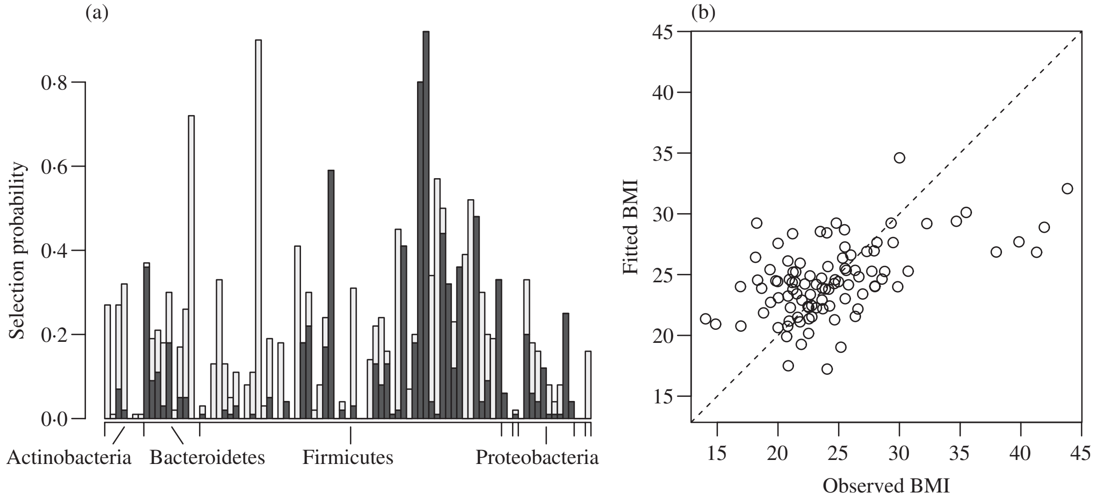

Fig. 1. Analysis of gut microbiome data. (a) Selection probabilities with bootstrapped crossvalidation for 87 genera belonging to eight phyla: selections with a positive sign are shown by dark grey blocks and those with a negative sign are represented by light grey blocks; only four major phyla are indicated. (b) Fitted versus observed values of BMI.

the ℓ_1 penalty. In the Supplementary Material, we compare the performance of the two-step procedures formed by adding a refitting step to lasso (ii) and to the proposed method, confirming the advantages of our method in the more challenging settings.

5·2. *Application to gut microbiome data*

Gut microbiome composition is considered an important factor that affects energy extraction from the diet and contributes to human health or diseases such as obesity. We illustrate the usefulness of our proposed method by applying it to the dataset reported in Wu et al. (2011), which comes from a cross-sectional study of 98 healthy volunteers carried out at the University of Pennsylvania to investigate the connections between long-term dietary patterns and gut microbiome composition. Stool samples were collected from the subjects, and DNA samples were analysed by 454/Roche pyrosequencing of 16S rRNA gene segments from the V1–V2 region. The pyrosequences were denoised to yield an average of 9265 reads per sample, with a standard deviation of 3864. After taxonomic assignment of the denoised sequences, 3068 operational taxonomic units were combined into 87 genera that appeared in at least one sample. Since the number of sequencing reads varied greatly across samples, these count data should not be used directly in a standard regression analysis, and we transformed them into compositional data after replacing zero counts by the maximum rounding error 0·5 (Aitchison, 2003, § 11.5). Demographic information on the subjects, including body mass index, BMI, was also collected. We are interested in identifying a subset of important genera whose subcomposition is associated with BMI.

We applied the proposed method to this dataset with BMI as the response, and used a refitted version of ten-fold crossvalidation to choose the tuning parameter, where the prediction error for each sample split was computed with the refitted coefficients obtained after model selection and without penalization. To obtain stable selection results, we generated 100 bootstrap samples and used the same crossvalidation procedure to select the genera. The selection probabilities of 87 genera with bootstrapped crossvalidation are shown in Fig. 1(a). Four genera were

Table 2. *Selection probabilities and refitted coefficients of four selected genera in the gut microbiome data*

		Selection probability		Refitted
Phylum	Genus	Boot. CV	Stab. sel.	coefficient
Bacteroidetes	*Alistipes*	0·72	0·89	−0·76
Firmicutes	*Clostridium*	0·90	0·96	−1·35
Firmicutes	*Acidaminococcus*	0·80	0·92	0·61
Firmicutes	*Allisonella*	0·92	0·87	1·50

Boot. CV, bootstrapped crossvalidation; Stab. sel., stability selection.

selected over 70 times out of the 100 bootstrap replicates. We also followed the stability selection approach of Meinshausen & Bühlmann (2010) to assess the stability of the selected genera, where 100 subsamples of size $n/2$ were taken to compute the selection probabilities. All four genera had a selection probability greater than 0·85, indicating that the selection results are quite stable. These four genera along with their selection probabilities and refitted coefficients are presented in Table 2. A plot of the fitted versus observed values of BMI in Fig. 1(b) shows that, except for five obese subjects, the model with four selected genera fits the data reasonably well.

Since our simulations have demonstrated that the lasso (i) estimator is inferior in all respects, we compare our method only with the lasso (ii) estimator. With selection probabilities above the cut-off value of 0·7, three genera were selected by lasso (ii) with bootstrapped crossvalidation, which coincide with three of the four previously selected genera, the exception being *Alistipes*. To compare the prediction performance of the two methods, we randomly divided the data into a training set of 70 subjects and a test set of 28 subjects, and used the fitted model chosen by crossvalidation based on the training set to evaluate the prediction error on the test set. The prediction error averaged over 100 replicates was 30·30 for the proposed method and 30·55 for lasso (ii), with standard errors of 0·97 and 1·04, respectively, suggesting that the prediction performance of the proposed method is similar to or better than that of lasso (ii).

It is interesting to contrast the variable selection results at the phylum level: the proposed method selected both Bacteroidetes and Firmicutes as being associated with BMI, whereas lasso (ii) selected only Firmicutes. Thus, our method seems more consistent with the previous finding that the relative proportion of Bacteroidetes to Firmicutes is lower in obese mice and humans than in lean subjects (Ley et al., 2005, 2006). One biological explanation for the finding, as suggested by metagenomic and biochemical analyses, is that the Firmicutes-enriched microbiome holds a greater metabolic potential than the Bacteroidetes-enriched microbiome for more efficient energy harvest from the diet, which in turn contributes to changes in energy balance and subsequent weight gain (Turnbaugh et al., 2006). Furthermore, our selection results at the genus level indicate that obesity may be associated with changes in gut microbiome composition at a finer taxonomic level than previously thought.

6. Discussion

The linear log-contrast model assumes that the absolute amounts of the covariate components have no effect on the response. We have adopted this modelling approach in our microbiome data analysis because the total amount of the microbiome cannot be reliably measured in experiments. Nevertheless, if such measurements were available, it would be worthwhile to assume a more flexible model in which the total amount also plays a role in affecting the response. To this end, one could consider the semiparametric varying-coefficient log-contrast model

$$y_i = \beta_0(a_i) + \sum_{j=1}^{p} \beta_j(a_i) \log x_{ij} + \varepsilon_i, \qquad \sum_{j=1}^{p} \beta_j(a_i) = 0,$$

with a_i being the total amounts. This reduces to model (2) when all the coefficients $\beta_0, \ldots, \beta_p$ are constants. A regularized estimation procedure for this model could be developed by combining the ideas of our approach and the kernel lasso method of Wang & Xia (2009).

Another possible extension of our method for microbiome data analysis would be to take into account the phylogenetic relationships among the bacterial taxa. Under the biologically plausible assumption that phylogenetically close taxa tend to have similar effects on the clinical trait, one can combine the ℓ_1 penalty in our regularization problem with a Laplacian penalty that encourages smoothness among the regression coefficients of closely related taxa on the phylogenetic tree (Chen et al., 2013). Such an extension is likely to increase the power of identifying important taxa that are relatively rare but phylogenetically close.

Acknowledgement

This research was supported in part by the U.S. National Institutes of Health. We thank the editor, an associate editor and two referees for constructive comments that have helped to improve the paper. Wei Lin is now affiliated with Peking University.

Supplementary material

Supplementary material available at *Biometrika* online includes additional simulation results and the proofs of Propositions 1 and 2.

Appendix

Proof of Theorem 1. Let $\hat{S} = \{j : \hat{\beta}_j \dotplus 0\}$ denote the support of $\hat{\beta}$. If $p \in \hat{S}$, the optimality conditions for problem (3) can be written as

$$-n^{-1}(Z^p_{\hat{S}_{\backslash p}})^{\mathrm{T}}(y - Z^p \hat{\beta}_{\backslash p}) + \lambda\big\{\mathrm{sgn}(\hat{\beta}_{\hat{S}_{\backslash p}}) - \mathrm{sgn}(\hat{\beta}_p) 1_{s-1}\big\} = 0, \tag{A1}$$

$$\big\|n^{-1}(Z^p_{\hat{S}^{\mathrm{c}}})^{\mathrm{T}}(y - Z^p \hat{\beta}_{\backslash p}) + \lambda\, \mathrm{sgn}(\hat{\beta}_p) 1_{p-s}\big\|_\infty \leqslant \lambda, \tag{A2}$$

where $\hat{\beta}_p = -1^{\mathrm{T}}_{p-1}\hat{\beta}_{\backslash p}$. The idea of the proof is to define an event that occurs with high probability and, conditioning on that event, find some $\hat{\beta}$ with the desired properties such that (A1) and (A2) hold.

For $J \subset \{1, \ldots, p\}$, let Z_J denote the submatrix formed by the jth columns of Z with $j \in J$. By the union bound and the classical Gaussian tail bound, we have

$$\mathrm{pr}\{\|n^{-1}(Z_S)^{\mathrm{T}}\varepsilon\|_\infty \geqslant \lambda/2\} \leqslant \sum_{j \in S} \mathrm{pr}\{|n^{-1} z_j^{\mathrm{T}} \varepsilon| \geqslant \lambda/2\} \leqslant s \exp\{-n\lambda^2/(8\sigma^2)\}$$

and, since $\Pi \equiv I - n^{-1} Z^p_{S_{\backslash p}} (C^p_{S_{\backslash p} S_{\backslash p}})^{-1} (Z^p_{S_{\backslash p}})^{\mathrm{T}}$ is a projection matrix and has spectral norm at most 1,

$$\begin{aligned}
\mathrm{pr}\{&\|n^{-1}(Z^p_{S^{\mathrm{c}}})^{\mathrm{T}}\varepsilon - C^p_{S^{\mathrm{c}} S_{\backslash p}} (C^p_{S_{\backslash p} S_{\backslash p}})^{-1} n^{-1} (Z^p_{S_{\backslash p}})^{\mathrm{T}} \varepsilon\|_\infty \geqslant \lambda\xi\} \\
&= \mathrm{pr}\{\|n^{-1}(Z^p_{S^{\mathrm{c}}})^{\mathrm{T}}\, \Pi \varepsilon\|_\infty \geqslant \lambda\xi\} \\
&\leqslant \sum_{j \in S^{\mathrm{c}}} \mathrm{pr}\{|n^{-1}(z_j - z_p)^{\mathrm{T}}\, \Pi\varepsilon| \geqslant \lambda\xi\} \leqslant (p - s) \exp\{-n\lambda^2\xi^2/(8\sigma^2)\},
\end{aligned}$$

where we have used the normalization assumption $\max_j \|z_j\|_2 \leqslant \surd n$. Thus, with probability at least $1 - p\exp\{-n\lambda^2\xi^2/(8\sigma^2)\}$, the following inequalities hold:

$$\|n^{-1}(Z_S)^{\mathrm{T}}\varepsilon\|_\infty \leqslant \lambda/2, \quad \|n^{-1}(Z^p_{S^{\mathrm{c}}})^{\mathrm{T}}\varepsilon - C^p_{S^{\mathrm{c}}S_{\backslash p}}(C^p_{S_{\backslash p}S_{\backslash p}})^{-1}n^{-1}(Z^p_{S_{\backslash p}})^{\mathrm{T}}\varepsilon\|_\infty \leqslant \lambda\xi. \tag{A3}$$

In what follows, we condition on the event that (A3) holds and analyse the optimality conditions (A1) and (A2) using deterministic arguments.

First, we take $\hat\beta_{S^{\mathrm{c}}} = 0$. Substituting $y = Z^p_{S_{\backslash p}}\beta^*_{S_{\backslash p}} + \varepsilon$ and replacing $\hat S$ by S, we write (A1) as

$$\hat\beta_{S_{\backslash p}} - \beta^*_{S_{\backslash p}} = (C^p_{S_{\backslash p}S_{\backslash p}})^{-1}\big[n^{-1}(Z^p_{S_{\backslash p}})^{\mathrm{T}}\varepsilon - \lambda\{\mathrm{sgn}(\hat\beta_{S_{\backslash p}}) - \mathrm{sgn}(\hat\beta_p)1_{s-1}\}\big]. \tag{A4}$$

Now define $\hat\beta_{S_{\backslash p}}$ by (A4) with $\mathrm{sgn}(\hat\beta_{S_{\backslash p}})$ and $\mathrm{sgn}(\hat\beta_p)$ replaced by $\mathrm{sgn}(\beta^*_{S_{\backslash p}})$ and $\mathrm{sgn}(\beta^*_p)$, respectively. By (A3), (A4) and the triangle inequality, we have

$$\begin{aligned}\big\|\hat\beta_S - \beta^*_S\big\|_\infty &= \big\|D_{SS_{\backslash p}}(C^p_{S_{\backslash p}S_{\backslash p}})^{-1}\big[n^{-1}(Z^p_{S_{\backslash p}})^{\mathrm{T}}\varepsilon - \lambda\{\mathrm{sgn}(\hat\beta_{S_{\backslash p}}) - \mathrm{sgn}(\hat\beta_p)1_{s-1}\}\big]\big\|_\infty\\ &\leqslant \big\|D_{SS_{\backslash p}}(C^p_{S_{\backslash p}S_{\backslash p}})^{-1}(D_{SS_{\backslash p}})^{\mathrm{T}}\big\|_\infty\big\|n^{-1}(Z_S)^{\mathrm{T}}\varepsilon\big\|_\infty + \lambda\,\big\|D_{SS_{\backslash p}}(C^p_{S_{\backslash p}S_{\backslash p}})^{-1}(D_{SS_{\backslash p}})^{\mathrm{T}}\big\|_\infty\\ &\leqslant \varphi\lambda/2 + \varphi\lambda = 3\varphi\lambda/2 < \beta_{\min}\end{aligned}$$

by assumption. This implies that $\mathrm{sgn}(\hat\beta_S) = \mathrm{sgn}(\beta^*_S)$, and hence we have found a $\hat\beta$ such that the desired properties and (A1) hold.

It remains to verify that $\hat\beta$ also satisfies (A2). Substituting $y = Z^p_{S_{\backslash p}}\beta^*_{S_{\backslash p}} + \varepsilon$ and using (A4), we write

$$\begin{aligned}n^{-1}(Z^p_{S^{\mathrm{c}}})^{\mathrm{T}}(y - Z^p\hat\beta_{\backslash p}) + \lambda\,\mathrm{sgn}(\hat\beta_p)1_{p-s} &= n^{-1}(Z^p_{S^{\mathrm{c}}})^{\mathrm{T}}\varepsilon - C^p_{S^{\mathrm{c}}S_{\backslash p}}(\hat\beta_{S_{\backslash p}} - \beta^*_{S_{\backslash p}}) + \lambda\,\mathrm{sgn}(\beta^*_p)1_{p-s}\\ &= n^{-1}(Z^p_{S^{\mathrm{c}}})^{\mathrm{T}}\varepsilon - C^p_{S^{\mathrm{c}}S_{\backslash p}}(C^p_{S_{\backslash p}S_{\backslash p}})^{-1}n^{-1}(Z^p_{S_{\backslash p}})^{\mathrm{T}}\varepsilon\\ &\quad + C^p_{S^{\mathrm{c}}S_{\backslash p}}(C^p_{S_{\backslash p}S_{\backslash p}})^{-1}\lambda\big\{\mathrm{sgn}(\beta^*_{S_{\backslash p}}) - \mathrm{sgn}(\beta^*_p)1_{s-1}\big\} + \lambda\,\mathrm{sgn}(\beta^*_p)1_{p-s}.\end{aligned}$$

Then, by (A3), Condition 1 and the triangle inequality, we have

$$\begin{aligned}\big\|n^{-1}(Z^p_{S^{\mathrm{c}}})^{\mathrm{T}}(y - Z^p\hat\beta_{\backslash p}) + \lambda\,\mathrm{sgn}(\hat\beta_p)1_{p-s}\big\|_\infty &\leqslant \big\|n^{-1}(Z^p_{S^{\mathrm{c}}})^{\mathrm{T}}\varepsilon - C^p_{S^{\mathrm{c}}S_{\backslash p}}(C^p_{S_{\backslash p}S_{\backslash p}})^{-1}n^{-1}(Z^p_{S_{\backslash p}})^{\mathrm{T}}\varepsilon\big\|_\infty\\ &\quad + \lambda\big\|C^p_{S^{\mathrm{c}}S_{\backslash p}}(C^p_{S_{\backslash p}S_{\backslash p}})^{-1}\{\mathrm{sgn}(\beta^*_{S_{\backslash p}}) - \mathrm{sgn}(\beta^*_p)1_{s-1}\} + \mathrm{sgn}(\beta^*_p)1_{p-s}\big\|_\infty\\ &\leqslant \lambda\xi + \lambda(1-\xi) = \lambda,\end{aligned}$$

which verifies (A2) and completes the proof. □

References

Aitchison, J. (1982). The statistical analysis of compositional data (with Discussion). *J. R. Statist. Soc.* B **44**, 139–77.

Aitchison, J. (2003). *The Statistical Analysis of Compositional Data*. Caldwell, New Jersey: Blackburn Press.

Aitchison, J. & Bacon-Shone, J. (1984). Log contrast models for experiments with mixtures. *Biometrika* **71**, 323–30.

Aitchison, J. & Shen, S. M. (1980). Logistic-normal distributions: Some properties and uses. *Biometrika* **67**, 261–72.

Bertsekas, D. P. (1996). *Constrained Optimization and Lagrange Multiplier Methods*. Belmont, Massachusetts: Athena Scientific.

Bühlmann, P. & van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Berlin: Springer.

Chen, J., Bushman, F. D., Lewis, J. D., Wu, G. D. & Li, H. (2013). Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostatistics* **14**, 244–58.

Chen, J. & Li, H. (2013). Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *Ann. Appl. Statist.* **7**, 418–42.

Fan, Y. & Tang, C. Y. (2013). Tuning parameter selection in high dimensional penalized likelihood. *J. R. Statist. Soc.* B **75**, 531–52.

Friedman, J. H., Hastie, T. J., Höfling, H. & Tibshirani, R. J. (2007). Pathwise coordinate optimization. *Ann. Appl. Statist.* **1**, 302–32.

Ley, R. E., Bäckhed, F., Turnbaugh, P., Lozupone, C. A., Knight, R. D. & Gordon, J. I. (2005). Obesity alters gut microbial ecology. *Proc. Nat. Acad. Sci.* **102**, 11070–5.

LEY, R. E., TURNBAUGH, P. J., KLEIN, S. & GORDON, J. I. (2006). Human gut microbes associated with obesity. *Nature* **444**, 1022–3.

MEINSHAUSEN, N. & BÜHLMANN, P. (2010). Stability selection (with Discussion). *J. R. Statist. Soc.* B **72**, 417–73.

TIBSHIRANI, R. J. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B **58**, 267–88.

TIBSHIRANI, R. J. & TAYLOR, J. (2011). The solution path of the generalized lasso. *Ann. Statist.* **39**, 1335–71.

TURNBAUGH, P. J., LEY, R. E., MAHOWALD, M. A., MAGRINI, V., MARDIS, E. R. & GORDON, J. I. (2006). An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444**, 1027–31.

WAINWRIGHT, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Info. Theory* **55**, 2183–202.

WANG, H. & XIA, Y. (2009). Shrinkage estimation of the varying coefficient model. *J. Am. Statist. Assoc.* **104**, 747–57.

WU, G. D., CHEN, J., HOFFMANN, C., BITTINGER, K., CHEN, Y.-Y., KEILBAUGH, S. A., BEWTRA, M., KNIGHTS, D., WALTERS, W. A., KNIGHT, R., SINHA, R., GILROY, E., GUPTA, K., BALDASSANO, R., NESSEL, L., LI, H., BUSHMAN, F. D. & LEWIS, J. D. (2011). Linking long-term dietary patterns with gut microbial enterotypes. *Science* **334**, 105–8.

ZHAO, P. & YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.* **7**, 2541–63.

[*Received June* 2013. *Revised April* 2014]