Euphytica (2012) 183:379–387
DOI 10.1007/s10681-011-0374-5

Penalized regression techniques for modeling relationships between metabolites and tomato taste attributes

Patricia Menéndez · Paul Eilers · Yury Tikunov · Arnaud Bovy · Fred van Eeuwijk

Received: 4 March 2010 / Accepted: 25 January 2011 / Published online: 6 February 2011

Abstract The search for models which link tomato taste attributes to their metabolic profiling, is a main challenge within the breeding programs that aim to enhance tomato flavor. In this paper, we compared such models calculated by the traditional statistical approach, stepwise regression, with models obtained by the new generation of regression techniques, known as penalized regression or regularization methods. In addition, for penalized regression, different scenarios and various model selection criteria were discussed to conclude that classical crossvalidation, selects models with many superfluous variables whereas model selection criteria such as Bayesian information criterion, seem to be more suitable, when the goal is to find parsimonious models, to explain tomato taste attributes based on metabolic information. An exhaustive comparison of the discussed methodology was done for six sensory traits, showing that the most important covariates were identified by the stepwise regression as well as by some of the penalized regression methods, despite the general disagreement on the size of the regression coefficients between them. In particular, for stepwise regression the coefficients are inflated due to their high variance which is not the case with penalized regression, showing that this new methodology, can be an alternative to obtain more accurate models.

Keywords Penalized regression · Tomato taste attributes · Metabolites · Phenotype prediction · Variable selection · Stepwise regression

P. Menéndez (✉) · P. Eilers · F. van Eeuwijk
Biometris - Applied Statistics, Wageningen University,
P.O. Box 100, 6700 AC, Wageningen, The Netherlands
e-mail: patriciamenendez1@gmail.com

P. Menéndez · Y. Tikunov · A. Bovy · F. van Eeuwijk
Centre for BioSystems Genetics, P.O. Box 98,
6700 AB, Wageningen, The Netherlands

P. Eilers
Department of Biostatistics, Erasmus Medical Center,
P.O. Box 2040, 3000 CA, Rotterdam, The Netherlands

Y. Tikunov · A. Bovy
Plant Research International, Wageningen University,
6700 AA, Wageningen, The Netherlands

Introduction

A better understanding of the biochemical basis of taste attributes is a main challenge within the tomato breeding programs which aim particularly to improve tomato flavor. Tomato (*Solanum Lycopersicum* L.) belonging to the Solanacea family and originally from South America, is one of the most consumed vegetables in the world and it has a big impact upon human diet as well as on our health (Agarwal and Rao 2000). Although, it is well known that the main components contributing to the flavour in tomato fruits are a mixture of sugars, acids and amino acids

together with volatiles and minerals (Baldwin et al. 1991; Saliba-Colombani et al. 2001), identification and quantification of the constituents that account for the differences in tomato flavour is still to a large extent an open problem. In this study, various statistical approaches to provide quantitative models that explain tomato taste attributes based on metabolic measures, are compared.

Different studies have been conducted to decipher the relationship between sensory traits and metabolites, ranging from studies based on principal component analysis (Krumbein and Auerswald 1998; Krumbein et al. 2004) to some recent ones, in which networks were constructed to illustrate the correlations between sensory traits and metabolites (Ursem et al. 2008; Carli et al. 2009). Multiple linear regression seems to be one of the most appropriate platforms to provide quantitative models which link taste attributes to sensory traits. Multiple linear regression models have been proposed by Skovgaard (1995) as a general framework to model relationships between instrumental and sensory measurements. In tomato related studies, Verkerke et al. (1998) presented a model which links a set of pre-selected metabolites with certain sensory traits. More recent studies within the same scheme are those reported by Tandon et al. (2003) and Abegaz et al. (2004), in which predictive models for tomato taste were presented, based on volatile and non volatile compounds.

In the great majority of the aforementioned studies, multiple linear regression models were computed based on ordinary least squares, in combination with forward stepwise techniques for feature selection. In this paper, we compare this existing methodology with a new generation of regression techniques, called regularization or penalization methods. This new methodology enjoys fame for its ability of performing estimation and variable selection at once, handling models where the number of variables is greater than the number of observations and for producing more accurate models.

In particular, we focused on Lasso (Tibshirani 1996) and elastic net (Zou and Hastie 2005) and different model selection strategies. We evaluate the advantages and disadvantages of this new methodology in comparison with the traditional stepwise regression, for the study of tomato sensory traits in relation to metabolic compounds.

Materials and methods

Data description

The collection of tomato germplasm analyzed in this study can be divided in three morphological types which are, beef, round and cherry, consisting of 94 cultivars, provided by six different breeding companies. This set of cultivars represents, to a large extent, the important commercial varieties in the market and have a considerable phenotypic variation between the different types (beef, round and cherry) as well as between individuals of the same type.

Sensory and metabolic measurements form the empirical data for this study. The sensory data covers the spectrum of fragance, taste, after taste and mouth feel, and was scored by a trained tasting panel of observers of taste, smell and texture. At a biochemical level, the data consisted of metabolic records that can be divided into two categories: volatiles and derivatized compounds (Table 1), analyzed from ripe tomato fruits. Of special interest are the volatile compounds (derived from different precursors including amino acids, fatty acids and carotenoids) because of their large influence on flavor perception. Volatiles were measured by using Gas Chromatography and Mass Spectrometry according to the methods formerly reported by Tikunov et al. (2005). The organic acids and sugars were profiled by the same techniques as described in the protocol employed for quantification of volatiles by Roessner-Tunali et al. (2003).

The same data set, has been studied by Ursem et al. (2008), van Berloo et al. (2008a, b), where more details about the data and their preparation can be found.

Penalized regression

In this investigation we are interested in finding the relationship between a given quantitative trait $\mathbf{Y}$, for an observed phenotype and a collection of metabolic variables $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$. A simple and yet very convenient model, to describe this type of association, is the so called linear model

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon \tag{1}$$

where p, $\beta = \{\beta_0, \ldots, \beta_p\}$ and $\epsilon = \{\epsilon_1, \ldots, \epsilon_n\}$ represent the number of variables, the regression

Table 1 Volatiles and non derivatized chemical compounds used in this study

Volatiles	Non derivatized
(1) Methylbutanal	(18) Glucose
(2) Penten3one	(19) Sucrose
(3) 3-Methylbutanol	(20) Fructose
(4) 2-Methylbutanol	(21) Myoinositol
(5) *Cis*3hexenal	(22) Malic acid
(6) Hexanal	(23) Citric acid
(7) *Trans*2hexenal	(24) Aspartic acid
(8) *Cis*3hexenol	(25) Glutamic acid
(9) *Trans*2heptenal	
(10) Methyl5hepten2one	
(11) Isobutylthiazol	
(12) Phenylacetaldehyde	
(13) Methoxyphenol	
(14) Phenylethanol	
(15) Methylsalicylate	
(16) Betadamascenone	
(17) Betaionone	

Numbers in *brackets* correspond to the encoding used in this paper

coefficients and the errors in the model. The errors are assumed to be independent and identically distributed normal random variables, with mean 0 and variance σ^2.

Ordinary least squares (OLS) estimates, are well known solutions to the multiple linear regression problem (1), obtained when minimizing the residual sum of squares. Although unbiased, OLS estimates are discredited for being unstable and overfitting of data in the presence of collinearity or in a high dimensional set up, i.e when the number of variables, p is larger than the number of observations, n. In any of the previous scenarios, OLS estimates are variance inflated and have a poor prediction accuracy. However, these problems can be partially alleviated by conducting variable selection.

Another alternative to the least squares solution drawbacks, is provided by the so called penalization or regularization techniques such as Lasso, proposed by Tibshirani (1996). Lasso, the acronym for least absolute shrinkage and selection operator, has became very popular because simultaneously performs estimation and variable selection. The main idea here to estimate the regression coefficients β, consists of minimizing the residual sum of squares plus an $\boldsymbol{L}_1$ constraint on the regression coefficients as follows

$$\hat{\beta}(\lambda) = argmin_{\beta}\left\{\|\mathbf{Y}-\mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_1\right\} \quad (2)$$

where $\|\mathbf{Y}-\mathbf{X}\beta\|_2^2 = \sum_{i=1}^{n}(Y_i - (\mathbf{X}\beta)_i)^2$ is the residual sum of squares, $\|\beta\|_1 = \sum_{i=1}^{n}|\beta_j|$ and $\lambda \geq 0$ being the penalty parameter which controls the amount of shrinkage, acting as a tuning parameter for the model. Large values of λ account for greater amount of shrinkage, drawing the model coefficients towards zero. Besides, the geometry of the $\boldsymbol{L}_1$ constraint ensures that some of them will be exactly zero, producing in that way sparse models, which depend on the choice of the penalty parameter λ.

Similarly to Lasso, elastic net is a shrinkage and variable selection method for linear regression, proposed by Zou and Hastie (2005). Elastic net tries to combine the good properties of Lasso together with the ones from Ridge regression (Hoerl and Kennard 1970), to obtain sparse models with reduced standard error estimates. It solves problem (1) by minimizing the residual sum of squares, adding a convex constraint for the regression coefficients to find $\hat{\beta}$ as the following minimizer

$$\hat{\beta}(\lambda) = argmin_{\beta} \left\{\|\mathbf{Y}-\mathbf{X}\beta\|_2^2 + \lambda\Big((1-\alpha)\|\beta\|_1 + \alpha\|\beta\|_2^2\Big)\right\} \quad (3)$$

with $\|\beta\|_2^2 = \sum_{i=1}^{n}\beta_j^2$ and $\lambda \geq 0$ being the penalty parameter, behaving as in the Lasso regression. Furthermore, the value $\alpha \in [0,1]$ decides on the type of constraint applied, being a compromise between the ones in Lasso and Ridge regression. The first part of the constraint, equivalent to $\|\beta\|_1 \leq s$ (Lasso) generates a sparse model. The second one $\|\beta\|_2^2 \leq s$ (Ridge), encourages a grouping effect, removes the limitation on the number of selected variables and, stabilizes the Lasso regularization path. In addition, the convex constraint $(1-\alpha)\|\beta\|_1 + \alpha\|\beta\|_2^2$, depends on the value $\alpha \in [0,1]$ which allows the construction of a broad range of possible models that enjoy different properties. Finally, it is worth noting that here, the optimal models will depend on the choice of the two parameters λ and α.

The elastic net, being a combination of Lasso and ridge regression was expected to be suitable for the modeling of sensory traits on a set of metabolites,

where these metabolites are assumed to come from a small set of metabolic pathways, with metabolites within pathways showing some correlation. The Lasso property of the elastic net then should select the pathways that enter the regression model for a particular sensory trait. The ridge property subsequently shrinks the metabolites within pathways in about the same amount.

These regularization techniques rely on fast and efficient computing algorithms, to calculate the set of possible Lasso or elastic net solutions $\{\hat{\beta}(\lambda), \lambda \in [0, \infty)\}$, that depend on the parameter λ, and are known as solution paths or traces. So far, different algorithms have been proposed to compute the whole path of Lasso solutions. One of the most popular, was a path following algorithm, called the least angle regression algorithm (LARS) and proposed by Efron et al. (2004). This algorithm has the same order of computation as a least square fit (Hastie et al. 2009). An alternative algorithm for computing Lasso as well as elastic net path solutions, is the coordinate descendant algorithm by Friedman et al. (2007). In addition to being faster for resolving large problems, this algorithm can be applied to a non convex penalty functions.

Model selection

Variable selection is a common problem in modern statistical analysis, arising from the necessity of identifying the set of important variables among all the superfluous ones. Noisy variables add complexity to the models and do not lead to great improvements in prediction power. The usual variable selection procedure is based on the residual sum of squares and a penalty which take into account the number of parameters in the candidate model. In analogy to Lasso and elastic net, stepwise regression finds the candidate model as the minimizer of

$$\hat{\beta}(\lambda) = argmin_{\beta}\left\{ \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_0 \right\} \quad (4)$$

where the $\boldsymbol{L}_0$ norm penalty is $\|\beta\|_0 = \sum_{i=1}^{p} I(\beta_j \neq 0)$, that is equivalent to the number of variables included in the model. For stepwise regression, first the variables that belong into the model are identified and then once the model has been identified, the coefficients are estimated. Regularization techniques offer an alternative to the traditional variable selection methods such as forward or stepwise regression (Efroymson 1960), which are known for being unstable under certain situations (Breiman 1996). Under the Lasso or elastic net framework, feature selection is equivalent to model choice. The penalty parameters there, account for the amount of shrinkage in the regression coefficients, and therefore for the number of variables appearing in the model. When a high penalty is chosen, few variables are included, whereas when a very low penalty is selected, most of them will be present. Because complex models are not necessarily performing better than the simpler ones, the main challenge here is to find a trade off between sparsity and prediction accuracy.

Many model selection techniques have been developed during the last years and crossvalidation (Stone 1974), based on the performance of the estimated model into a new data set (generalization error), is one of the most widely used among them. Crossvalidation, based on generalization performance, describes the model performance in a new data set, by selecting those that have the best prediction performance. Other very popular model selection criteria are those of the form

$$\Phi(\gamma) = -2\ln(L) + |\gamma|D(n) \quad (5)$$

where L corresponds to the maximized value of the likelihood function for the estimated model γ, $|\gamma|$ is the effective model dimension and $D(n)$ is a function of the sample size (Broman and Speed 2002). Very well known examples of them, are the Akaike Criteria, (Akaike 1974) in which $D(n) = 2$, or BIC (Schwarz 1978) when $D(n) = \ln(n)$.

For regression models computed via Lasso regression the effective model dimension is equal to the number of variables included in the model (Zou et al. 2007); for the elastic net it is equal to $\sum_{j=1}^{size(A)} \frac{d_j}{d_j+2}$, were A denotes the set of variables in the model and d_j is the jth eigenvalue of the matrix $X_A^t X_A$ (van der Kooij 2007).

Results and discussion

Models selected by crossvalidation

Six tomato sensory traits were analyzed by different regression techniques, to find their underlying

metabolic models.Multiple linear regression models with 3 particular elastic net penalties, namely $\alpha = 0.25$, $\alpha = 0.50$, $\alpha = 0.75$, as defined in Eq. 3, and from Lasso were computed. For those fits, the number of predictor variables selected in each model is rather large, as can be seen in Table 2, together with their corresponding goodness of fit R^2. The model selection criterion in all the cases was crossvalidation. Crossvalidation, tunes models to achieve the best prediction accuracy (P.A.)

$$P.A.(\lambda) = 1 - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - X_i\hat{\beta}_{\lambda}}{1 - \mathcal{H}_{ii}}\right)^2 \quad (6)$$

where X_i is the ith row of X and $\mathcal{H}_{ii}$ is the ith diagonal element of the "hat" matrix $\mathcal{H}$ (such that $\hat{Y} = \mathcal{H}Y$). The prediction accuracy measures the predictive power of a given model on new sample data. Nevertheless, crossvalidation does not take into account the model complexity, as a consequence, correlated predictors may be included in the model leading to a decrease in prediction accuracy.

The relative minor influence of the different three elastic net and Lasso penalties in the final models, is because the prediction error curves for the four different models reached their minimum almost at the same location, as shown in Fig. 1. Crossvalidation selects optimal models to be the minimizers of those curves, and in this study the minimum of the four prediction error curves falls very close together, producing models which have almost identical number of regressors. However, in all the cases the most parsimonious models were those given by the Lasso since this method applies the strongest constraint to the regression coefficients.

Models selected by BIC and stepwise regression

Models selected by BIC (Bayesian information criterion), particular case of (5), for elastic net penalties, contained a large number of regressors (Table 3). The selected variables did not show clear grouping structures which could be interpretable in terms of chemical pathways. Therefore, we decided to focus on Lasso and stepwise regression and further compare the performance between these methods.

Lasso regression models, selected by the BIC criterion, were superior in terms of the coefficient of determination R^2, from those selected by crossvalidation, and achieved similar predictive power as those from stepwise regression (Table 4). For stepwise regression, the criteria used to decide whether a variable entered or left the model, was BIC. It is also important to notice that the number of variables selected by stepwise regression is in general smaller than those selected by Lasso (Table 3). Stepwise regression coefficients are of larger size than those from Lasso having an influence on the number of variables entering into the model. That is shown in Fig. 2, in particular for the sensory trait *taste spicy*, although it was the case for all the traits. In general, the regression coefficients signs, obtained from Lasso and stepwise regression coincide in all the cases. Models calculated by Lasso which contained five variables, were studied to assess the order in which predictor variables were selected along the traces. In addition, it is also of interest to compare the predictor variables selected by those models, with the ones obtained by stepwise regression and Lasso when BIC was used as a selection criterion.

There is a general agreement on the selected variables by the BIC and step wise regression for all

Table 2 Number of predictor variables included in the models selected by crossvalidation for Lasso and elastic net with penalties: $\alpha = 0.25$, $\alpha = 0.50$, $\alpha = 0.75$

Sensory trait	$\alpha = 0.25$	R^2	P.A.	$\alpha = 0.50$	R^2	P.A.	$\alpha = 0.75$	R^2	P.A.	Lasso	R^2	P.A.
Taste spicy	18	0.860	0.836	17	0.861	0.836	17	0.861	0.836	17	0.861	0.836
Taste watery	16	0.825	0.799	17	0.826	0.799	17	0.827	0.799	16	0.827	0.799
Scent smoky	14	0.655	0.595	13	0.657	0.599	13	0.657	0.601	13	0.655	0.602
Taste sour	18	0.464	0.536	17	0.459	0.541	17	0.457	0.543	17	0.455	0.545
After taste bitter	18	0.370	0.266	17	0.378	0.274	17	0.386	0.278	16	0.389	0.280
Scent tomato	18	0.343	0.229	17	0.354	0.241	16	0.355	0.246	16	0.359	0.249

The quality of the different models is measured by the coefficient of determination R^2 and prediction accuracy values (P.A.)

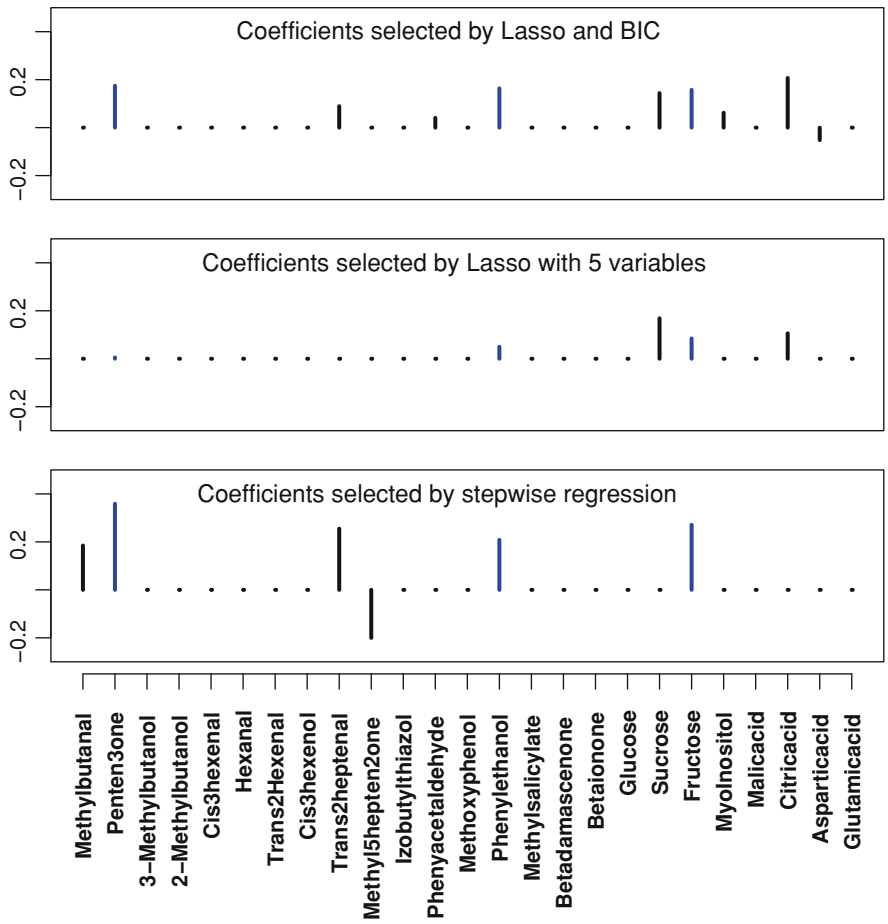

Fig. 1 Regression coefficients for sensory trait *taste spicy* computed by different models, common variables are shown in *blue*. *Upper panel*: Coefficients computed by Lasso with BIC as model selection criteria. *Middle panel*: Coefficients computed by Lasso stopping the algorithm when the model contains five variables. *Lower panel*: Coefficients obtained by stepwise regression

Table 3 Number of variables selected in the optimal models computed by BIC for elastic net

Sensory	$\alpha = 0.25$ Nr. var.	R^2	P.A.	$\alpha = 0.50$ Nr. var.	R^2	P.A.	$\alpha = 0.75$ Nr. var.	R^2	P.A.
Taste spicy	18	0.865	0.834	10	0.830	0.819	10	0.836	0.164
Taste watery	17	0.829	0.798	15	0.827	0.799	15	0.828	0.799
Scent smoky	6	0.516	0.502	5	0.544	0.529	5	0.554	0.540
Taste sour	10	0.508	0.484	3	0.346	0.343	3	0.383	0.378
After taste bitter	10	0.155	0.148	7	0.175	0.165	1	0.021	0.029
Scent tomato	3	0.059	0.063	3	0.074	0.076	1	0.037	0.043

Goodness of fit is expressed by the coefficient of determination, R^2, and prediction accuracy values (P.A.)

the traits except *taste spicy* (Table 3). Models which contained exactly five variables succeeded in selecting those which are more important although failed on the estimation of the coefficient value as is clear from Table 3. Furthermore, we are studying different criteria based on (5), to compare the performance of stepwise regression, Lasso and elastic net.

As it was shown by the above comparison, there are traits such as *taste spicy* and *taste watery*, that were predicted reasonable well by Lasso as well as by stepwise regression, whereas some others like *after-taste bitter* or *scent tomato*, could hardly be predicted by the considered set of metabolites, with any of the studied methodology. For the traits that failed to be predicted by any of the proposed methods, we can

Table 4 Sensory traits linear models based on volatiles and non derivatized tomato chemical compounds, computed by Lasso

Sensory	Lasso	Nr. var.	R^2	P.A.	Lasso for 5 predictors	Nr. var.	R^2	P.A.	Stepwise regr.	Nr. var.	R^2
Taste spicy	23, 2, 14, 20, 19, 9, 21, 24, 12	9	0.836	0.823	19, 23, 20, 14, 2	5	0.510	0.507	2, 20, 9, 14, 10, 1	6	0.836
Taste watery	23, 14, 20, 9, 21, 3, 17, 12, 7	9	0.773	0.764	23, 14, 20, 21, 9	5	0.619	0.616	23, 14, 20, 17	4	0.774
Scent smoky	15, 13, 23	3	0.527	0.518	15, 13, 23, 19, 16	5	0.539	0.528	15, 23	2	0.447
Taste sour	3, 23, 17	3	0.392	0.387	3, 23, 17, 11, 25	5	0.420	0.413	3, 23	2	0.586
After taste bitter	14	1	0.057	0.062	14, 4, 19, 13, 6	5	0.104	0.104	19	1	0.111
Scent tomato	11	1	0.067	0.070	11, 7, 1, 16, 4	5	0.123	0.118	11	1	0.128

Shown are the optimal models selected by BIC, models containing five variables, calculated from the Lasso traces, and models obtained by stepwise regression. The models goodness of fit is illustrated by the coefficient of determination R^2 and prediction accuracy values (P.A.)

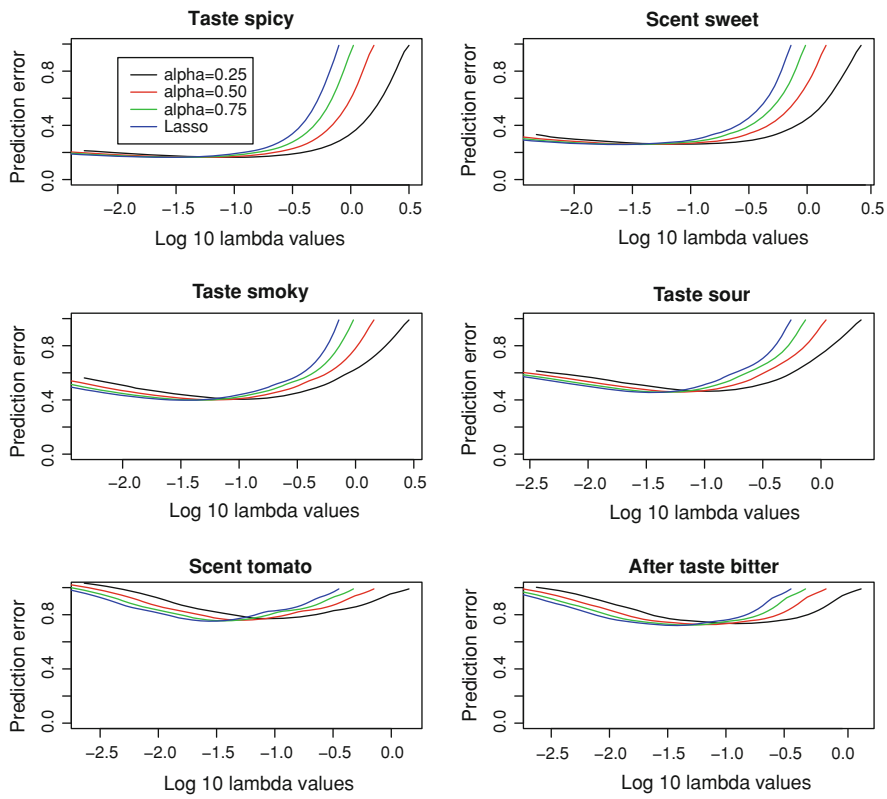

Fig. 2 Prediction error curves for the studied sensory traits. The *x*-axis represents the grid of 70 $log_{10}\lambda$ values for which models were computed, *y*-axis presents the corresponding prediction error values. Prediction error values on the *right side* of the *x*-axis correspond to models for which few variables were included while moving towards the *left* along the *x*-axis lead to models which contain more variables

conclude that the chemical basis behind them was not contained in the set of studied metabolites. For the sensory traits that were predicted well, the relevance of some variables is clear since they appeared in all the models regardless of the technique used.

Concluding remarks

In this study we have compared existing regression methodology for linear models, namely stepwise regression, with a new generation of regression

procedures known as elastic net and Lasso. The aim was to analyze the different approaches to find optimal biochemical models based on metabolic information to predict a group of sensory traits. In the set up of this investigation, that is, when the number of variables in the model is smaller than the number of observations, Lasso models, selected by BIC, achieved a comparable fit to those from stepwise regression in terms of R^2, not being very clear which method was superior. However, looking at the size of the regression coefficients, it was clear that those estimated by stepwise regression, had larger size than the ones calculated by the Lasso approach. The Lasso models contained more correlated predictors than the stepwise regression models which may have induced the smaller estimates for the coefficients.

Based on our analysis, stepwise regression provided a very good platform to find satisfactory prediction models a as we have seen in this study.

Elastic net and Lasso models selected by crossvalidation failed finding the set of most important variables. That result, agreed with the conclusions in Leng et al (2006), where they proofed that regularization models selected by techniques based on prediction accuracy, as is the case with crossvalidation, are not consistent in terms of variable selection. In other words, variable selection and model prediction are different issues which need to be simultaneously taken into account suggesting that model selection criteria such as BIC lead to more appropriate models.

To further improve the prediction accuracy of the sensory traits that were not well predicted by none of the discussed methods a further analysis with a more extensive set of metabolites will be carried out. Finally, we aim to obtain more accurate sensory-metabolic models by including genetical considerations.

Acknowledgments This study was financed by the Centre for BioSystems Genetics, project BB12.

References

Agarwal S, Rao AV (2000) Tomato lycopene and its role in human health and chronic diseases. CMAJ 163:739–744

Abegaz EG, Tandon KS, Scott JW, Baldwin EA, Shewfelt RL (2004) Partitioning taste from aromatic flavor notes of fresh tomato (*Lycopersicon esculentum, Mill*) to develop predictive models as a function of volatile and nonvolatile component. Postharvest Biol Technol 34: 227–235

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Automat Control 19:716–723

Baldwin EA, Nisperos-Carriedo MO, Moshonas MG (1991) Quantitative analysis of flavour and other volatiles and for certain constituents of two tomato cultivars during ripening. J Am Soc Hortic Sci 116:265–269

Breiman L (1996) Bagging predictors. Mach Learn 26:123–140

Broman KW, Speed TP (2002) A model selection approach for the identification of quantitative trait loci in experimental crosses. J R Stat Soc B 64:641–656

Carli P, Arima S, Fogliano V, Tardella L, Frusciante L, Ercolano MR (2009) Use of network analysis to capture key traits affecting tomato organoleptic quality. J Exp Bot 60:3379–3386

Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression. Ann Statist 32:407–499

Efroymson MA (1960) Multiple regression analysis. Mathematical methods for digital computers. John Wiley & Sons, New York, pp 191–203

Friedman J, Hastie T, Hoefling H, Tibshirani T (2007) Pathwise coordinate optimization. Ann Appl Statist 1:302–332

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction. Springer, New York

Hoerl AE, Kennard RW (1970) Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12:55–67

Krumbein A, Auerswald H (1998) Characterization of aroma volatiles in tomatoes by sensory analyses. Nahrung 42:395–399

Krumbein A, Peters P, Bruckner B (2004) Flavour compounds and quantitative descriptive analysis of tomatoes (*Lycopersicon esculentum* Mill.) of different cultivars in short-term storage. Postharvest Biol Technol 32:15–28

Leng C, Linand Y, Wahba G (2006) A note on the Lasso and related procedures in model selection. Statistica Sinica 16:1273–1284

Roessner-Tunali U, Hegemann B, Lytovchenko A, Carrari F, Bruedigam C, Granot D, Fernie AR (2003) Metabolic profiling of transgenic tomato plants overexpressing hexokinase reveals that the influence of hexose phosphorylation diminishes during fruit development. Plant Physiol 133:84–99

Saliba-Colombani V, Causse M, Langlois D, Philouze J, Buret M (2001) Genetic analysis of organoleptic quality in fresh market tomato. 1. Mapping QTLs for physical and chemical traits. Theor Appl Genet 102:259–272

Skovgaard IM (1995) Modelling relations between instrumental and sensory measurements in factorial experiments. Food Qual Pref 6:239–244

Schwarz GE (1978) Estimating the dimension of a model. Ann Statist 6:461–464

Stone M (1974) Cross-Validatory Choice and Assessment of Statistical Predictions. J R Statist Soc B 36:111–147

Tandon KS, Baldwin JW, Scott JW, Shewfelt RL (2003) Linking sensory descriptors to volatile and nonvolatile components of fresh tomato flavor. J Food Sci 68:2366–2371

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J Royal Statist Soc B 58:267–288

Tikunov Y, Lommen A, De Vos CHR, Verhoeven HA, Bino RJ, Hall RD, Bovy AG (2005) A novel approach for nontargeted data analysis for metabolomics. Large-scale profiling of tomato fruit volatiles. Plant Physiol 139:1125–1137

Ursem R, Tikunov Y, Bovy A, van Berloo R, van Eeuwijk FA (2008) A correlation network approach to metabolite data analysis for tomato fruits. Euphytica 161:181–193

van Berloo R, Zhu A, Ursem R, Verbakel H, Gort G, van Eeuwijk FA (2008a) Diversity and linkage disequilibrium analysis within a selected set of cultivated tomatoes. Theor Appl Genet 117:89–101

van Berloo R, van Heusden AW, Bovy AG, Meijer-Dekens RG, Lindhout P, Eeuwijk van FA (2008b) Genetic research in a public-private research consortium: prospects for indirect use of Elige breeding germplasm in academic research. Euphytica 161:293–300

van der Kooij AJ (2007) Prediction accuracy and stability of regression with optimal scaling transformations. Child and Family Studies and Data Theory (AGP-D), Department of Education and Child Studies, Faculty of Social and Behavioural Sciences, Leiden University

Verkerke W, Janse J, Kersten M (1998) Instrumental measurement and modeling of tomato fruit taste. Acta Hort 456:199–205

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Statist Soc B 67:301–320

Zou H, Hastie T, Tibshirani R (2007) On the degrees of freedom of the lasso. Annals Statist 35:2173–2192