Electronic Journal of Applied Statistical Analysis
EJASA (2011), Electron. J. App. Stat. Anal., Vol. 4, Issue 2, 227 – 234
e-ISSN 2070-5948, DOI 10.1285/i20705948v4n2p227

USING DIFFERENTIAL GEOMETRIC LARS ALGORITHM TO STUDY THE EXPRESSION PROFILE OF A SAMPLE OF PATIENTS WITH LATEX-FRUIT SYNDROME

Luigi Augugliaro*, Angelo Marcello Mineo

Dipartimento di Scienze Statistiche e Matematiche, University of Palermo, Italy

Received 29 September 2010; Accepted 01 September 2011
Available online 14 October 2011

***Abstract**: Natural rubber latex IgE-mediated hypersensitivity is one of the most important health problems in allergy during recent years. The prevalence of individuals allergic to latex shows an associated hypersensitivity to some plant-derived foods, especially freshly consumed fruit. This association of latex allergy and allergy to plant-derived foods is called latex-fruit syndrome. The aim of this study is to use the differential geometric generalization of the LARS algorithm to identify candidate genes that may be associated with the pathogenesis of allergy to latex or vegetable.*

***Keywords**: Latex-fruit syndrome, variable selection, penalized regression, high dimensional, LARS*

1. Introduction

Natural rubber latex (NRL) is a product made primarly from the rubber tree, *Hevea brasiliensis*. NRL is used in more than 40000 medical devices, such as gloves, catheters, face masks, stethoscope, et cetera. NRL is also used in other nonmedical products, such as baloons, condoms, clothes, shoe soles, et cetera. IgE-mediated allergy to NRL is a significant health problem in industrialized contries, especially among health care workers, patients with congenital malformations, especially chidren with *spina bifida*, and children with a history of multiple surgical interventions [5]. According to some researchs, between 10% and 17% of medical personnel in Europe and US are believed to be sensitive to NRL. The prevalence of NRL allergy in patients with *spina bifida* is about 50% in industrialized countries, while is close to 0 in non

* Corresponding author. E-mail: luigi_augugliaro@unipa.it

industrialized countries. Moreover, sensitivity was found in about 34% of children with a history of 3 or more surgical interventions, that are a risk factor for the development of NRL allergy in children but not in adults [5].

Initial symptoms for people allergic to NRL are localized itching, erythema, or contact urticaria after few minutes from the NRL exposure. Progressive sensitization can also lead to generalized urticaria, angioedema, rhinitis, conjunctivitis, asthma, and anaphylactic shock minutes after dermal or mucosal contact with NRL proteins. An increasing number of individuals allergic to NRL reports severe reactions to latex, including generalized urticaria, bronchospasm, and hypotension.

Approximately 30-50% of individuals who are allergic to NRL show an associated hypersensitivity to some plant-derived foods, especially fresh fruit. This association of allergy to latex and allergy to plant-derived foods is called latex-fruit syndrome. An increasing number of plant sources, such as avocado, banana, chestnut, kiwi, peach, tomato, potato and bell pepper, has been associated with this syndrome [15]. Aim of this paper is to identify a set of genes to screen the molecular profiles of patients with allergy to latex from patients with allergy to fruit.

In literature several methods have been proposed to identify genes that can be used to discriminate between two or more groups. For this kind of problems the number of variables, say *p*, can be much larger than the sample size *n*. In this case, it is often assumed that only a small number of variables (genes) contributes to the response, which leads to assume the sparsity of the model. For sparse model we mean a generic regression model with the coefficient vector $\boldsymbol{\beta}$ sparse, i.e. with many elements equal to zero. In this field, many variable selection techniques for high dimensional statistical models are based on the penalized likelihood approach. Some examples are the least absolute shrinkage and selection operator (LASSO) estimator

$$\boldsymbol{\beta}(\lambda) = \text{argmin}\{||\mathbf{y}-\mathbf{X}\boldsymbol{\beta}||^2-\lambda||\boldsymbol{\beta}||_1\}$$

proposed by Tibshirani [14], where $||\mathbf{y}-\mathbf{X}\boldsymbol{\beta}||^2$ is the OLS loss function, $||\boldsymbol{\beta}||_1$ is the L_1-norm of the parameter vector $\boldsymbol{\beta}$ and λ is a positive tuning parameter used to select the trade-off between sparsity of the estimated parameter vector and the prediction behaviour of the model. Other important examples are the path following algorithm proposed by Park and Hastie [12] to estimate a generalized linear model with L_1-penalty function and the smoothly clipped absolute deviation (SCAD) penalized estimator proposed by Fan et al. [8], where the loss function is $||\mathbf{y}-\mathbf{X}\boldsymbol{\beta}||^2$ and the penalty function is a quadratic spline function with knots at λ and $a\lambda$, with λ the tuning parameter and *a* a given constant.

In a recent paper, Efron et al. [7] introduced a new method to select important variables in a linear regression model called least angle regression method (LARS). LARS algorithm can be described as follows. Starting with all coefficients equal to 0, the LARS algorithm finds the covariate that is most correlated with the response variable and proceeds on this direction. Then, the algorithm takes the largest step possible in the direction of this covariate until some other covariate has as much correlation with the current residual. LARS algorithm proceeds in a direction equiangular between the two covariates until a new covariate earns its way into the most correlated set (*A*) and then proceeds in the direction that has an equal angle with the three covariates until a new covariate is included in *A*, and so on. In recent years, there has been an enormous amount of research activity devoted to automatic model-building algorithms: Yuan and Lin [16] have extended LARS algorithm to linear regression models with grouped variables; Park and

Hastie [12] have proposed a path following algorithm for generalized linear models with L_1-penalty function; Rosset and Zhu [13] studied the generic regularized optimization problem for almost quadratic loss functions with L_1-penalty. These algorithms are based on the use of the L_1-penalty function to define a path in the parameter space. Aim of this paper is to identify a set of differentially expressed genes using the differential geometric LARS (dgLARS)[3]. As we shall see in the following of this paper, the geometrical theory underlying the dgLARS can be used to extend the sure independence screening method [9] used when we have an ultra-high dimensional feature space ($p>>n$), as suggested by Augugliaro and Mineo [2].

2. Differential Geometric LARS (dgLARS) algorithm

Let $\mathbf{Y}=(Y_1, Y_2, \ldots, Y_n)^T$ be a random variable vector having a probability density function

$$p_Y(\mathbf{y}; \boldsymbol{\theta}, \lambda) = a(\mathbf{y}; \lambda) \exp\{\lambda (\mathbf{y}^T\boldsymbol{\theta} - k(\boldsymbol{\theta}))\}, \qquad \mathbf{y}\in Y\subseteq \mathrm{R}^n, \tag{1}$$

with respect to a σ-finite measure ν on R^n, where $a(\cdot)$ and $k(\cdot)$ are specific given functions, the canonical parameter $\boldsymbol{\theta}$ varies in the subset $\Theta\subseteq\mathrm{R}^n$ and λ varies in a subset Λ of R^+. The model (1) is called *exponential dispersion model* [10]. We have chosen to use this model for its high flexibility and for its very general probabilistic assumptions. We denote the mean value of $\mathbf{Y}$ by $\boldsymbol{\mu}_Y=\mu(\boldsymbol{\theta})$. Standard theory on the exponential dispersion model tells us that the mean value and the canonical parameter are related to the gradient of the function $k(\cdot)$, namely, $\mu(\boldsymbol{\theta})=\partial k(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ (see [10] for more details), which is called *mean value mapping*. Since $\mu(.)$ is a one-to-one function from $int\Theta$ onto $\Omega=\mu(int\Theta)$, the exponential dispersion model may be parameterized by $(\boldsymbol{\mu};\sigma^2)$, where $\sigma^2=\lambda^{-1}$ is called *dispersion parameter*. In what follows we shall assume that the dispersion parameter is fixed. Under this assumption, following Amari [1], the parameter space can be treated as a n-dimensional Riemannian manifold where $\boldsymbol{\mu}_Y$ plays the role of coordinate system and the Fisher information matrix is a Riemannian metric. A generalized linear model is completely specified by the following assumptions:

a) $\mathbf{y}=(y_1, y_2, \ldots, y_n)^T$ is a set of n independent observations taken from (1);
b) for each random variable Y_i we have a column of covariates $\mathbf{x}_i=(x_{i1}, x_{i2},\ldots, x_{ip})^T\in X\subseteq\mathrm{R}^p$, with $p<n$. These covariates are related to the mean value of $\mathbf{Y}$ by a known function such that $\mu_i=f(\mathbf{x}_i^T\boldsymbol{\beta})$, where $\boldsymbol{\beta}\in B\subseteq\mathrm{R}^p$; in order to simplify our notation, we denote $\boldsymbol{\mu}(\boldsymbol{\beta})=(f(\mathbf{x}_1^T\boldsymbol{\beta}), f(\mathbf{x}_2^T\boldsymbol{\beta}),\ldots, f(\mathbf{x}_n^T\boldsymbol{\beta}))^T$. We assume that $\boldsymbol{\mu}(\cdot)$ is an embedding with domain B;
c) the dispersion parameter σ^2 does not depend on the vector of covariates.

Given the assumption (b), $\boldsymbol{\mu}(B)=\Omega_B\subset\Omega$ is a Riemannian submanifold of Ω, then we can generalize the notion of angle between two given vectors. Let $\boldsymbol{\beta}(\gamma)$ be a differentiable curve. Following Kass and Vos [11], we have that the derivative of the log-likelihood function $L(\boldsymbol{\beta}(\gamma))$ with respect to β_i is given by

$$\partial_{\beta_i} L(\boldsymbol{\beta}(\gamma)) = \left\langle \partial_{\beta_i}\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma)); \mathbf{r}(\boldsymbol{\beta}(\gamma))\right\rangle_{\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))} \tag{2}$$

where $\langle .;. \rangle_{\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))}$ is the inner product between the current residual vector $\mathbf{r}(\boldsymbol{\beta}(\gamma))$ and the i-th base of the tangent space of Ω_B at $\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))$. Using expression (2) we have the following differential geometric identity

$$\begin{aligned} &\partial_{\beta_i} L(\boldsymbol{\beta}(\gamma)) = \cos(\rho_i(\boldsymbol{\beta}(\gamma))) \cdot \left\|\mathbf{r}(\boldsymbol{\beta}(\gamma))\right\|_{\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))} \cdot \left\|\partial_{\beta_i} \boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))\right\|_{\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))} = \\ &= \cos(\rho_i(\boldsymbol{\beta}(\gamma))) \cdot \left\|\mathbf{r}(\boldsymbol{\beta}(\gamma))\right\|_{\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))} \cdot (I_i(\boldsymbol{\beta}(\gamma)))^{1/2} \end{aligned} \tag{3}$$

where $\rho_i(\boldsymbol{\beta}(\gamma))$ is the local angle between $\mathbf{r}(\boldsymbol{\beta}(\gamma))$ and $\partial_{\beta_i} \boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))$, $I_i(\boldsymbol{\beta}(\gamma))$ is the expected Fisher information for $\boldsymbol{\beta}(\gamma)$ and $\|\cdot\|_{\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))}$ is the norm defined on the tangent space of Ω_B at $\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))$. Condition (3) shows that the gradient of the log-likelihood function does not generalize the notion of equiangular condition, since we are not considering the variation related to $(I_i(\boldsymbol{\beta}(\gamma)))^{1/2}$. To overcome this problem, Augugliaro and Wit [3] propose a generalization of the LARS algorithm based on the following condition:

$$\left|r_i^u(\boldsymbol{\beta}(\gamma))\right| = \left|(I_i(\boldsymbol{\beta}(\gamma)))^{-1/2} \cdot \partial L(\boldsymbol{\beta}(\gamma))\right| = \cos(\rho_i(\boldsymbol{\beta}(\gamma))) \cdot \left\|\mathbf{r}(\boldsymbol{\beta}(\gamma))\right\|_{\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))} \quad \forall i \in A \tag{4}$$

where $\left|r_i^u(\boldsymbol{\beta}(\gamma))\right|$ is the i-th Rao score statistic evaluated in $\boldsymbol{\beta}(\gamma)$ and A is the active set, namely, the set of indices of covariates that are included in the actual model. Condition (4) is called generalized equiangularity condition and it is used in [3] to define a method that generalizes LARS to generalized linear models. This method is called dgLARS and the interested reader can refer to [3] for more details about it.

When we work in a ultra-high dimensional feature space, namely $p >> n$, the following expression

$$\frac{\left(r_i^u(\boldsymbol{\beta}(\gamma))\right)^2}{\left\|\mathbf{r}(\boldsymbol{\beta}(\gamma))\right\|^2_{\boldsymbol{\mu}(\boldsymbol{\beta}(\gamma))}} = \cos^2(\rho_i(\boldsymbol{\beta}(\gamma))) \tag{5}$$

can be used to define a genuine generalization of the sure independence screening method [9] for generalized linear models. Using expression (5), Augugliaro and Mineo [2] have proposed the following method to identify the relevant variables in a generalized linear model defined in a ultra-high dimensional feature space. Following Fan and Lv [9], we consider

$$\boldsymbol{\omega} = (\cos^2(\rho_1(\hat{\beta}_0)), \cos^2(\rho_2(\hat{\beta}_0)), \ldots, \cos^2(\rho_p(\hat{\beta}_0)))^{\mathrm{T}},$$

where $\hat{\beta}_0$ is the maximum likelihood estimate of the parameter of a generalized linear model with only the intercept. For a given value $d<n$, we sort the p component-wise magnitudes of the vector $\boldsymbol{\omega}$ in decreasing order and define the submodel:

$M_d = \{i \in \{1, 2, \ldots, p\} : \cos^2(\rho_i(\hat{\beta}_0))$ is among the first d largest of all the corresponding values$\}$.

This is a straightforward way to reduce the dimensionality from p to d. Then, submodel M_d is studied by using the dgLARS algorithm.

3. Application of dgLARS algorithm to a latex-fruit allergy data set

In oder to identify a set of genes to screen the molecular profiles of patients with allergy to latex from patients with allergy to fruit, a logistic regression model estimated by using the dgLARS algorithm was applied to a DNA microarray data set. From a sample of 6 patients allergic to latex and from a sample of 5 patients allergic to vegetable food, peripheral blood mononuclear cells were isolated on ficoll gradient. Each sample was hybridized on Affymetrix human focus array and data were processed with Affymetrix MAS 5.0 software. These data sets are available from the internet site of the National Center for Biotechnology Information (NCBI), in the Gene Expression Omnibus (GEO) archives (URL: http://www.ncbi.nlm.nih.gov/geo/). The data sets reference is GSE13619. The two sample sizes are very low, but usually this is the dimension of the sample size of studies conducted in this field, since in the World the people affected from the latex-fruit syndrome are not so many and it is very difficult to lead a medical study with a big number of patients. Anyway, in spite of the sample size, we think that the following analysis is interesting in its own right.
The expression profile of 8746 genes was used for the logistic regression model estimated by using the dgLARS algorithm. As suggested by Fan and Lv [9] a value of $d = [n/\log(n)] = 4$ was used for the sure independence screening step.

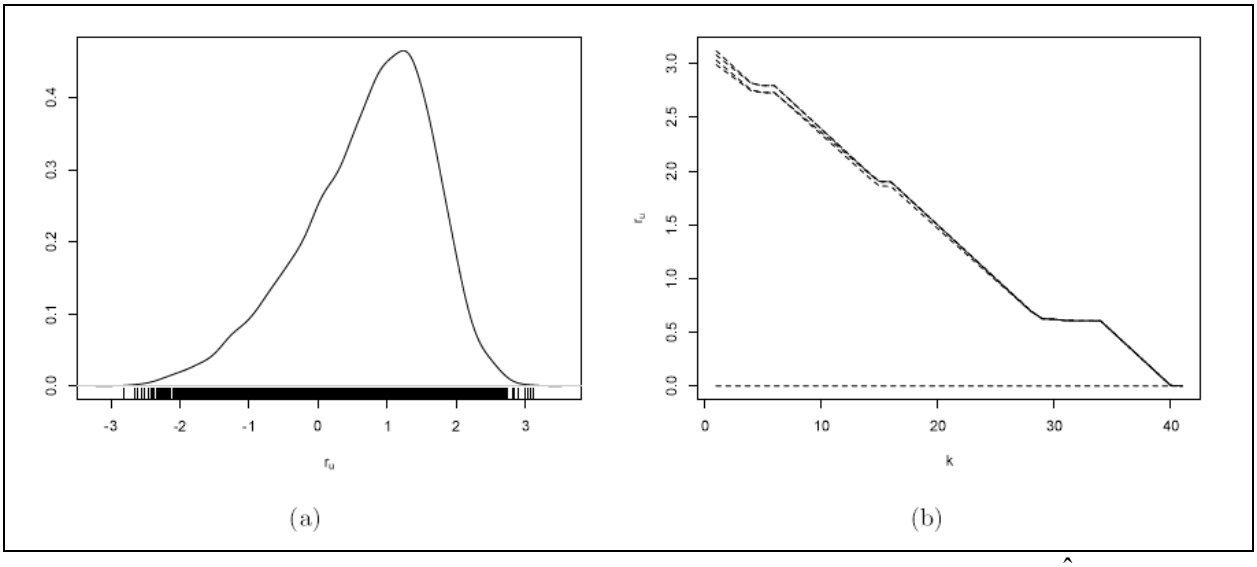

Figure 1. Panel (a) shows the density of the Rao score test statistics evaluated at the point $\hat{\boldsymbol{\beta}}(\gamma_0)$. Panel (b) shows the path of the Rao score test statistics as function of the number (k) of the algorithm

Panel (a) in figure 1 shows the density of the Rao score test statistics evaluated at the starting point $\boldsymbol{\beta}(\gamma_0)$, while panel (b) shows the path of the Rao score test statistics. The number of variables with non-zero coefficients was selected using the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC).

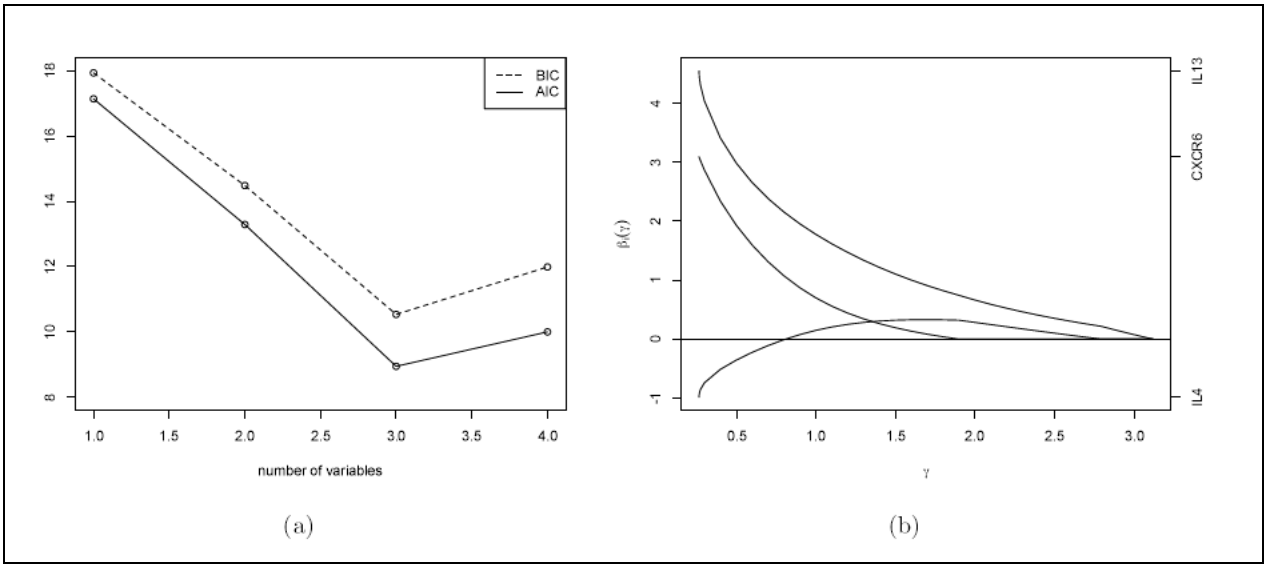

Figure 2. Panel (a) shows AIC and BIC used to select the number of predictors. Panel (b) shows the coefficients path of the selected variables

In this setting, the Akaike Information Criterion (AIC) is defined as

$$\text{AIC} = -2\, L(\boldsymbol{\beta}_A(\gamma)) + 2|A(\gamma)|$$

where $|A(\gamma)|$ is the cardinality of the active set, while the Bayesian Information Criterion (BIC) is defined as

$$\text{BIC} = -2\, L(\boldsymbol{\beta}_A(\gamma)) + \log(n)\, |A(\gamma)|.$$

Panel (a) in figure 2 shows AIC and BIC values as function of the number of variables with nonzero coefficients. Using these measures of goodness of fit only 3 variables were used for further analysis. Panel (b) shows the coefficient paths of the selected variables. This set of genes was also identified using the logistic regression model with L_1-penalty function. The identified genes are the interleukin 13 (IL-13), the interleukin 4 (IL-4) and chemokine (C-X-C motif) receptor 6 (CXCR6). In particular, the interleukin 13 (IL-13) and the interleukin 4 (IL-4) are closely related with the latex-fruit syndrome, since the products of these genes play a critical role in allergic reaction regarding antigen presentation, IgE synthesis, and activation of mast cell, respectively (see [6] and [4] for further details). The majority of food allergen-specific $CD4^+$ T lymphocytes isolated from food-allergic individuals was found to synthesize high levels of IL-4 and IL-13.

4. Conclusions

In this paper we have used the differential geometric generalization of the LARS algorithm proposed by Augugliaro and Wit [3] and linked to the sure independence screening method by Augugliaro and Mineo [2] to select a set of genes to discriminate between patients with latex allergy and with fruit allergy. Latex-fruit syndrome is a well-defined disorder affecting from 20% to 60%

of patients with latex allergy. A number of environmental and genetic factors seems to contribuite to the latex-sensitivity phenotype. Although exposure to natural rubber latex products is necessary for sensitization, it is not sufficient. A number of other environmental and genetic factors seems to contribute to the latex-sensitive phenotype. Known risk factors for latex allergy include an atopic history, concomitant food allergies, and delayed skin reactions to NRL-containing products. Although there is overwhelming support for a genetic component for allergic disease, the multigenic nature of the phenotype has made the identification of susceptibility genes a difficult task. A large number of studies have explored general risk factors for allergic disease by means of the candidate gene approach or genome-wide analyses which are based on the penalization methods, such as the generalized linear model with the L_1-penalty function or the ridge regression. In this paper, we have seen that the dgLARS method can also be used with good results.
In spite of the sample size very low, our analysis identifies two of the well known genes that are related with the latex allergy, namely the interleukin 4 (IL-4) and the interleukin 13 (IL-13), that several studies have identified. Products of these genes play a critical role in allergic reaction regarding antigen presentation, IgE synthesis, and activation of mast cell, respectively.

Acknowledgement

We want to thank the University of Palermo for partially supporting this research. We want also to thank an anonymous referee for his/her valuable comments.

References

[1]. Amari, S.I. (1985). *Differential-Geometrical Methods in Statistics*. New York: Springer-Verlag.
[2]. Augugliaro, L., Mineo, A.M. (2009). Applying Differential Geometric LARS Algorithm to Ultrahigh Dimensional Feature Space. In *Actes de XVIemes Rencontres de la Société Francophone de Classification*, Grenoble, 2-4 September 2009, 201–204.
[3]. Augugliaro, L., Wit, E.C. (2009). Generalizing LARS Algorithm Using Differential Geometry. In *Book of Short Papers of the 7° Meeting of the Classification and Data Analysis Group of the Italian Statistical Society*, Catania, 9-11 September 2009, 189-192.
[4]. Blanco, C., Sanchez-Garcia, F., Torres-Galvan, M.J., Dumpierrez, A.G., Almeida, L., Figueroa, J., Ortega, N., Castillo, R., Gallego, M.D., Carrillo, T. (2004). Genetic basis of the latex-fruit syndrome: Association with HLA class II alleles in a Spanish population. *Journal of allergy and clinical immunology*, 114, 1070–1076.
[5]. Brehler, R., Kutting, B. (2001). Natural rubber latex allergy: a problem of interdisciplinary concern in medicine. *Archives of internal medicine*, 161, 1057–1064.
[6]. Brown, R.H., Hamilton, R.G., Mintz, M., Jedlicka, A.E., Scott, A.L., Kleeberger, S.R. (2005). Genetic predisposition to latex allergy: role of interleukin 13 and interleukin 18. *Anesthesiology*, 102, 496–502.
[7]. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32, 407–451.

[8]. Fan, J., Li, R. (2001). SCAD, Penalized likelihood, Oracle estimator. *Journal of the American Statistical Association*, 96, 1348–1360.
[9]. Fan, J., Lv, J. (2008). Sure independent screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society series B*, 70, 849–911.
[10]. Jørgensen, B. (1987). Exponential dispersion models. *Journal of the Royal Statistical Society series B*, 49, 127–162.
[11]. Kass, R.E., Vos, P.W. (1997). *Geometrical Foundation of Asymptotic Inference*. New York: John Wiley.
[12]. Park, M.Y., Hastie, T. (2007). L_1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society series B*, 69, 659–677.
[13]. Rosset, S., Zhu, J. (2007). Piecewise linear regularizated solution paths. *The Annals of Statistics*, 35, 1012–1030.
[14]. Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society series B*, 58, 267-288.
[15]. Wagnerand, S., Breiteneder, H. (2002). The latex-fruit syndrome. *Biochemical Society Transactions*, 30, 935–940.
[16]. Yuan, M., Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society series B*, 68, 49–67

.