REGRESSION AND SPARSE REGRESSION METHODS FOR VISCOSITY ESTIMATION OF ACID MILK FROM IT'S SLS FEATURES

Sara Sharifzadeh, Jacob L. Skytte, Otto H. A. Nielsen, Bjarne Ersbøll, Line H. Clemmensen

(sarash,jlsk,otni,be,lhc)@imm.dtu.dk

ABSTRACT

Statistical solutions find wide spread use in food and medicine quality control. We investigate the effect of different regression and sparse regression methods for a viscosity estimation problem using the spectro-temporal features from new Sub-Surface Laser Scattering (SLS) vision system. From this investigation, we propose the optimal solution for regression estimation in case of noisy and inconsistent optical measurements, which is the case in many practical measurement systems. The principal component regression (PLS), partial least squares (PCR) and least angle regression (LAR) methods are compared with sparse LAR, lasso and Elastic Net (EN) sparse regression methods. Due to the inconsistent measurement condition, Locally Weighted Scatter plot Smoothing (Loess) has been employed to alleviate the undesired variation in the estimated viscosity. The experimental results of applying different methods show that, the sparse regression lasso outperforms other methods. In addition, the use of local smoothing has improved the results considerably for all regression methods. Due to the sparsity of lasso, this result would assist to design a simpler vision system with less spectral bands.

Index Terms— Regression, Sparse Regression, Smoothing, Sub-Surface Laser Scattering (SLS)

1. INTRODUCTION

Today, statistical computations and solutions play important role in the analysis of biological, medical and various other types of measured signals. These measurements are usually performed by means of different kinds of sensors such as imaging and vision systems. A part from the type of signal and measurement system, the common requirement of all of these applications is to analyze and interpret the input data for the aim of prediction, classification or relation and correlation between different attributes, features or variables.

Online quality inspection for food control is one of the examples of such systems. For many years, human has performed this task. But, today there is more interest toward online fast non-invasive inspections. Recently, a hyperspectral imaging technique [1, 2] and subsurface laser scattering (SLS) [3] have been successfully applied on different food inspection cases. The advantages of these vision techniques over the traditional methods include minimal sample preparation, contact-less and nondestructive nature, fast acquisition times, and characterization of numerous chemical compositions as well as rheological structures simultaneously.

In this paper, statistical linear regression and sparse regression solutions are used to estimate the rheological characteristics of fermented milk from its SLS features. Acidified milk products like yoghurt are important food products and their chemical and rheological characteristics play an important role to their quality and palatability for the consumer. These characteristics changes during the acidification process. Therefore, online tracking of these changes would help to control the quality of the product.

1.1. Measurement of Rheological Properties

The viscosity of acid milk increases during the acidification process. One of the common milk acidification techniques is based on the use of $glucono - \delta - lactone$ (GDL). The traditional method for measuring the rheological properties of acid milk gels involves an oscillatory applied strain or stress [4, 5]. In the strain controlled version of this experiment, the sample is subjected to a sinusoidally oscillating strain from which the viscosity response is measured. On the other hand, it is possible to measure the chemical and structural variations of material by means of SLS vision systems [3]. One important benefit of this method is that, it is totally contact-less.

The aim of this paper is to apply linear regression as a statistical solution on the milk feature matrix obtained from the SLS system and predict the viscosity level of the product. We are also motivated to find a sparse regression solution which requires less number of wavelengths for viscosity prediction. The sparsity is important regarding to the fact that the vision systems with their spectra are costly and not feasible to implement in the industry for online food productions. Performing the prediction using the minimum number of wavelengths would enable the selection of small specific vision systems which are costly effective. Therefore, both regression and

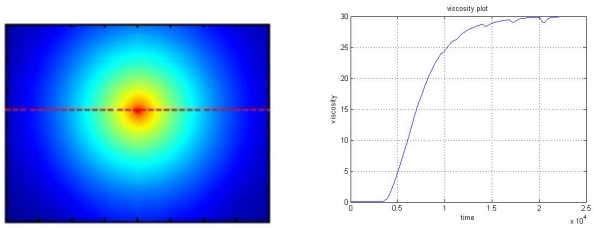
This work was (in part) financed by the Centre for Imaging Food Quality project which is funded by the Danish Council for Strategic Research (contract no 09-067039) within the Programme Commission on Health, Food and Welfare.

sparse regression methods are applied on a training set of SLS features to estimate the necessary parameters for the test set (In this paper denoted as tr and ts) and their results are compared.

The rest of this paper is organized as follows; Section 2 describes the data set. Different regression and local smoothing methods used for viscosity prediction are described in section 3. In section 4, we present the experimental results and finally there is a conclusion in section 5.

2. DATA DESCRIPTION

In the applied SLS technique, a hyper-spectral laser beam (450-1020nm with steps of 10nm) is exposed to the surface of a sample, while a vision system acquires an image for each wavelength [3]. From these images, scattering profiles are extracted from the scattering center and outwards as shown in Fig.1a. Each of these profiles yields an intensity curve, which is analyzed by applying the double logarithm to intensities. This approximates a linear behavior, which is modeled by a linear model. This model consists of a slope and an intercept, which makes up the aforementioned SLS features. The feature matrix is $X_{N\times P}$, where N denotes the number of time points and P is the number of wavelengths. The viscosity level is $Y_{N\times 1}$, Which is shown in Fig.1b.



(a) SLS profile (b) Milk viscosity

Fig. 1: The SLS scattering profile (a) The viscosity of acid milk during acidification (b).

3. VISCOSITY ESTIMATION FROM SLS FEATURES

The general linear regression problem $\hat{Y} = X\hat{\beta}_{ols} + \varepsilon$, results to a minimum biased output [6]. However, in practical experiments, usually there are variations and inconsistency in measurements due to different sources of noise. On the other hand, it is always probable that some inputs have higher correlation to the output, which necessitates weighing and shrinking strategies. In order to decrease the variance of the output, regularization strategies would be employed to the regression problem.

3.1. Ridge Regression

In ridge regression, the regression coefficients are shrunk in size by a $L2$ norm penalty term:

$$\hat{\beta}_{ridge} = argmin_{\beta}\left\{\sum_{i=1}^{N}(y_i - \beta_0 - \sum_{j=1}^{P} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{P}\beta_j^2\right\} \quad (1)$$

The second term in Equ.1, increases the bias of the model and, also controls the shrinkage of the coefficients to decrease the variance. Thus, the choice of λ is a trade off between bias and variance.

3.2. Least Angle Regression (LAR)

In Least angle regression, the coefficients are sequentially added to the model. At the first step, the variable with highest correlation with the response is moved to an active set, and its corresponding coefficient will move toward its least square value. Thus, its correlation with the residual would decrease Then, the search for the next variable with the same correlation with the residual starts and when found, it would be moved also to the active set and the coefficients of all active variables would move together in a way that keeps their correlations tied and decreasing. This is continued until all the variables join the model, and ends at the full least-squares fit. But, the minimum error is not necessarily obtained from the full coefficients set. The training error reduces as the steps increase. But, for test error, the error starts to increase in a critical point again due to the ovefitting of the model. Therefore, a sparse LAR solution with less number of non-zero regression coefficients works better for a test set.

3.3. Principle Component Regression (PCR)

Principal component regression uses the principal components of the input vector ($z_m = Xv_m$: v_m is the m^{th} eigen vector of X), and then regresses Y on $z_1, z_2, ..., z_M$ for some $M \leq P$. Since the z_m are orthogonal, this regression is just a sum of uni-variate regressions:

$$\hat{y}_{pcr}(M) = \bar{y}1 + \sum_{m=1}^{M}\hat{\theta}_m z_m, \quad (2)$$

where $\bar{y}1$ is the mean vector and $\hat{\theta}_m = \langle z_m, y\rangle / \langle z_m, z_m\rangle$. Usually M is the number of components including more than 90 or 95 percent of variance of the input. Therefore, PCR discards the P–M smallest eigenvalue components. However, PCR is not a sparse solution.

3.4. Partial Least Squares (PLS)

PLS uses Y, in addition to X, for construction of a set of linear combinations of the inputs for regression. Therefore, its solution path is a nonlinear function of Y. PLS seeks directions that have high variance and have high correlation with

the response, in contrast to PCR which is only based on high variance. However, if there is a lot of variation in X that has no connection to the variation of Y , PCR would have problem. Because, it finds those latent variables that describe as much as possible of the variation in X even correlated or not to Y. If the output be only sensitive to only small variations in X, and if the interferences vary a lot, then the latent variables found by PCR may not describe Y in best way. PLS is capable of finding better solutions in this situation, since it forms variables that are relevant for describing Y.

The PLS algorithm is iterative. First some parameters should be trained:

1. Standardize each x_{jtr} column of training $X_{tr} = \{x_1, ..., x_P\}$ and Set $\hat{y}^{(0)} = \bar{Y}_{tr}1$, and $x_j^{(0)} = x_{jtr}; j = 1, ..., P$.

2. For $m = 1, 2, ..., P$

- (a) $z_m = \sum_{j=1}^{P} \hat{\varphi}_{mj} x_{jtr}^{(m-1)}$, Where z_m is the m^{th} partial least squares direction, and $\hat{\varphi}_{mj} = \left\langle x_{jtr}^{(m-1)}, Y_{tr} \right\rangle$, are the weights that describe the uni-variate effects of x_{jtr} on Y_{tr}
- (b) $\hat{y}^{(m)} = \hat{y}^{(m-1)} + \hat{\theta}_m z_m$, where $\hat{\theta}_m = \langle z_m, Y_{tr} \rangle / \langle z_m, z_m \rangle$ shows the regression of Y_{tr} on partial directions z_m
- (c) $x_j^{(m)} = x_j^{(m-1)} - \hat{p}_{mj} z_m$, $j = 1, 2, ..., P$, where $\hat{p}_{mj} = \left\langle z_m, x_j^{(m-1)} \right\rangle / \langle z_m, z_m \rangle$ shows the regression of x_{jtr} on partial directions z_m

As the number of partial directions increases, the average error decreases asymptotically to a minimum level. As shown in Fig.2a, before all partial directions be added to the model, the error reaches to a minimum point, which could be used to choose the proper number of partial directions. The prediction or test algorithm is also iterative and uses the trained parameters $\hat{\theta}_m, \hat{\varphi}_m, \hat{p}_m$. The following steps should be repeated for the selected number of partial direction z_m:

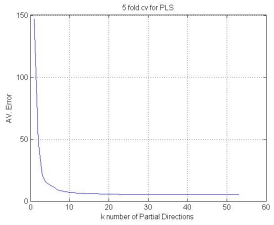
- (a) $z_m = \sum_{j=1}^{P} \hat{\varphi}_{mj} x_{jts}^{(m-1)}$
- (b) $x_j^{(m+1)} = x_j^m - \hat{p}_{mj} z_m$
- (c)$\hat{y}^{(m+1)} = \hat{y}^{(m)} + \hat{\theta}_m z_m$

3.5. Lasso

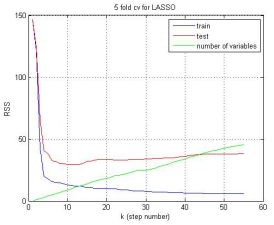
The lasso is a sparse shrinkage method. The lasso penalty is $L1$ norm:

$$\hat{\beta}_{lasso} = argmin_\beta \left\{ \frac{1}{2} \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{P} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{P} |\beta_j| \right\} \quad (3)$$

The $\hat{\beta}_{lasso}$ could be calculated using the LAR method just by a simple modification. In lasso path, if a non-zero coefficient hits zero, its variable should be dropped out from the active set of variables and be treated like other zero coefficients. The plot of average training and test errors versus step numbers are shown in Fig.2b. In addition, the average number of non-zero coefficients in each step is plotted. Like in LAR, the training error reduces as the steps increase, while after some steps, the error increases in case of the test set.



(a) The errors in PLS



(b) The errors in Lasso

Fig. 2: The average error versus number of partial directions for PLS over 5 folds CV. of milk SLS data (a) The Lasso training and test errors versus step numbers in 5 fold CV. of milk SLS data (b).

3.6. Elastic-Net

Elastic-net is in fact a compromise between lasso and ridge. Each regression coefficient is calculated as a weighted combination of ridge and lasso. The elastic-net selects variables like lasso, and shrinks together the coefficients of correlated predictors like ridge. This will reduce the variance, but, at the same time the bias is not as much as it is in lasso.

$$\hat{\beta}_{en} = \lambda \sum_{j=1}^{p} ((1 - \alpha) |\beta_j| + \alpha \beta_j^2) \quad (4)$$

3.7. Local Smoothing of the Regression Output

For smoothing the output, further post processing techniques could be employed to reduce the variation of the output and improve the results. We propose the use of local regression function "Loess" for the estimated viscosity.

Loess, applies a local smoothing procedure which means that each smoothed value is determined as a weighted combination of its neighboring data points defined within a span. In each span, a regression weight shown in Equ.5 is calculated for each point inside the span.

$$w_i = \left(1 - \left|\frac{x - x_i}{d(x)}\right|^3\right)^3 \quad (5)$$

where x is the data point and x_i are the neighbors in the span. $d(x)$ is the distance to the most distant point inside the span. Then a weighted linear regression using a quadratic polynomial is performed to estimate the new value for smooth curve.

Table 1: Comparison of the results for different regression methods

	R-Square		NonZeroCoeff.
	Original	Smoothed	
OLS	71.6970	77.4572	55
Ridge	79.3328	84.5143	55
LAR	72.0458	74.2769	55
Sparse-LAR	80.2842	80.9357	12
PCR	76.0871	77.0668	55
PLS	79.5066	84.5439	55
LASSO	89.0863	92.2666	12
EN	83.7776	84.8672	35

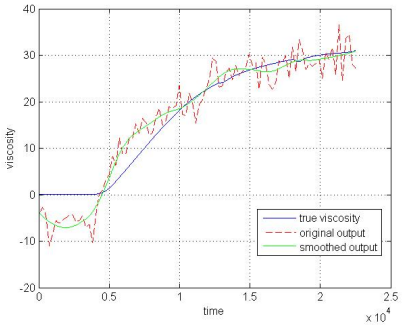


Fig. 3: Comparison of the regression and smooth regression Results for lasso

4. EXPERIMENTAL RESULTS

At DTU food laboratories, the SLS imaging experiments have been performed during milk acidification process. In these experiments, the GDL acidification technique has been used. In each experiment, the feature matrix includes $N = 22501$ observations in $P = 54$ different wavelength from the slope of the SLS profile. These features cover the whole acidification process. Also, the viscosity vector is measured and has the same length of each feature vector 22501×1. We have used two sets as training and one as test. Although the regression plots of these 3 sets show similar behavior, they are not numerically consistent. This means that we should expect error in our estimations.

The different regression methods have been implemented and applied on the training data, using the leave-one-out 5 fold cross validation. We have used some solutions from [7]. Cross validation is used to estimate the expected prediction error, from which, it could be possible to choose the best parameters for the regression model from the minimum error point. The Loess smoothing method with a span of 30%, has been applied on the output results from the test data.

R-square measure has been used for evaluation of the results. It is a statistical measure of how well a regression line approximates real data points; an R-squared of 1.0 (100%) indicates a perfect fit. It is calculated based on the Residual Sum of Squares (RSS) and Total Sum of Squares (TSS):

$$RSS = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2, TSS = \sum_{i=1}^{N}(y_i - \bar{Y})^2 \quad (6)$$

$$R^2 = (1 - \frac{RSS}{TSS}) \times 100 \quad (7)$$

Table 1 shows the R-Square results for the test sets before and after smoothing. Fig.3 shows the results for the lasso method which yields the best R-square. Therefore, the vision system could have just 12 spectral bands instead of 54 which is more efficient.

5. CONCLUSION

In this paper, different regression and sparse regression methods have been applied on the hyper-spectral SLS features from milk acidification process to estimate the viscosity level of the acid milk. In addition, the local smoothing method, Loess has been used for the purpose of smoothing and variation reduction of the estimated output. Comparison of the experimental results of different methods, show that the sparse lasso regression yields the highest R-square values. Because of the sparse nature of lasso, we can conclude that, the vision system could have only 12 spectral bands, instead of 54 which is more efficient.

6. REFERENCES

[1] D.W. Sun, *Hyperspectral Imaging for Food Quality Analysis and Control*, Elsevier, 30 Corporate Drive, Suite 400, Burlington, MA 01803, USA, 2010.

[2] H. F. Grahan and P. Geladi, *Techniques and Applications of Hyper Spectral Image Analysis*, John Wiley and Sons Ltd, West Sussex PO19 8SQ, England, 2007.

[3] R. Larsen F. Mller F. Donbk Nielsen L. Th. Carsten H. Aans H. O. H. A. Nielsen, A. Lindbjerg . Dahl and M.J. Carstensen, "Super-continuum light sources for hyperspectral subsurface laser scattering applica-tions for food inspection," *In Proc. SCIA2011 Sweden*, 2011.

[4] A. J. Lucey and H. Singh, "Formation and physical properties of acid milk gels: a review," *Food Research International*, vol. 30-No. 7, pp. 529–542, August 1997.

[5] A. Kamenan K. A. FranÃ§oise, T. Kablan and A. Lagaude, "Rheological and biochemical properties of acidified milk / pectin co-gels," *European Journal of Scientific Research*, vol. 25 No.4, pp. 584–596, January 2009.

[6] R. Tiboshirani T. Hastie and J. Friedman, *The Elements of Statistical Learning*, Springer, 2001.

[7] K. Sjöstrand, *PhD thesis: Regularized Statistical Analysis of Anatomy, Informatics and Mathematical Modelling*, Technical University of Denmark, 2007.