Electronic Journal of Applied Statistical Analysis
Vol. 09, Issue 01, April 2016, 95-110
DOI: 10.1285/i20705948v9n1p95

Normalization of gene expression data using support vector machine approach

Sandip Shil*[a], Kishore K. Das[b], and Ananta Sarkar[c]

[a] *Central Plantation Crops Research Institute, Research Centre, Guwahati-781017, India*
[b] *Department of Statistics, Gauhati University, Guwahati- 781014, India*
[c] *Directorate of Research on Women in Agriculture, Bhubaneswar - 751003, India*

Published: 26 April 2016

Normalization of gene expression data refers the process of minimizing non-biological variation in measured probe intensity levels so that biological differences in gene expression can be appropriately detected. Several linear normalization within arrays approaches have already been proposed. Recently, use of non-linear methods has been gained quite attention. In this study, our objective is to formulate non-linear normalization methods using support vector regression (SVR) and support vector machine quantile regression (SVMQR) approaches more easier way and, assess the consistency of these methods with respect to other standard ones for further application in gene expression data. After implementation, the performances of SVR and SVMQR have been compared with respect to other standard normalization methods namely, locally weighted scatter plot smoothing and kernel regression. The results indicate that the normalized data based on proposed methods are capable of producing minimum variances within replicate groups and, also able to detect truly expressible significant genes compared to above mentioned other normalized data.

keywords: support vector machine quantile regression, support vector regression, normalization methods, microarray, intensity level.

1 Introduction

Gene expression data is often contaminated with large noise (or, high variability) and, vary from one replicate to the other. This is mainly due to non-biological variations

*Corresponding author: sandip.iasri@gmail.com

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

such as: small differences in mRNA quantities and fluctuations, array quality, dye bias, localization of the probe sequences on the microarray, guanine-cytosine (GC)-content of the probe sequences, varying sensitivity in different detection ranges due to specific feature of particular microarray technologies and so on (Stekel, 2003). Probe intensity level of such data is expressed as a sum of true biological variation and several confounding factors, resulting due to non-biological variations. Normalization is a process of adjusting or, minimizing such non-biological variations so that biological differences of actual gene expression can be undoubtedly detected. In simple words, biological variation generated from a biological process is of interest, but non-biological variation should be discarded (Workman et al., 2002). Normalized data is usually achieved by performing normalization between or, within arrays of expression data. Normalization between arrays is a process in which, intensity levels are adjusted relative to the expression of one or, more reference genes/ probes whose levels are assumed to be constant between samples. Further, these intensities (or, log-ratios) have similar distributions across a series of arrays. In contrast, intensity levels are adjusted with respect to the median so that all arrays have the same median intensities (or, log-ratios), are known as within array normalization. As later one is more vital, we have restricted our study to within arrays normalization.

In last decades, several normalization methods (for between and/ or, within microarrays) have already been proposed, namely hybridization intensity ratios, median absolute deviation, percentile, variance stabilization, global, scale, intensity-dependent, composite and linear methods (Taniguchi et al., 2001; Huber et al., 2002; Smyth and Speed, 2003; Stekel, 2003; Yang and Speed, 2003). Interestingly, these are applicable to both the situations (between and/ or, within arrays) and, most of them have been formulated to normalize two channel gene expression data. However, these methods may also be applied to single channel expression data, in similar manner. Here, we have only considered two channel based data. Recently, use of non-linear methods, which are assumed to be superior to above ones, have gained popularity (Workman et al., 2002; Park et al., 2003; Wang et al., 2004). This is because of their capabilities to provide more accurate results. Some of them are implemented using locally weighted scatter plot smoothing (LOWESS), splines, wavelets and kernel regression (KERNEL) approaches. A recent promising technique, support vector machine (SVM) (Vapnik, 1995) approach can also be used to resolve this issue. Designing an efficient and competent non-linear method using this approach has already been suggested by several authors (Fujita et al., 2006; Sohn et al., 2008). In this study, our objective is to formulate normalization methods using SVM approaches more easier way and, assess the consistency of these methods with respect to other standard ones for further application in gene expression studies.

2 Methods and materials

Here, our idea is to estimate non-biological variation within arrays using some covariate information. As fluorescence intensities readings of red and green dyes of all arrays are available, one can easily compute the average log-intensity (denoted by A) and log-

intensity ratio (indicated by M). This ratio represents the actual relative expression level for each gene within array that can be further estimated using the average log-intensity for each gene within respective array. This ratio information may be thought of our dependent variable under consideration and, average log-intensity information as covariate. Therefore, non-biological variations within an array can be estimated on the basis of this covariate using different regression approaches (especially, SVR and support vector machine quantile regression (SVMQR)) and, subsequently normalization can be achieved by subtracting the fitted value from the corresponding log-intensity ratio. Let us, define r_{ij} and $g_{ij}, i = 1, 2, \ldots, n; j = 1, 2, \ldots, p$ as fluorescence intensities readings of j^{th} probe in i^{th} microarray sample for red and green dyes, respectively. n is the total number of arrays and p is the total number of probes in i^{th} microarray. As suggested by Dudoit et al. (2002), we may use the basic strategy of MA plot concept to formulate first rule. An MA plot depicts the relation between M values, which denote the log-ratio of gene intensities and A values, which denote the average gene log-intensity (readings of red and green dyes)for a spot. Hence, we can define the log-intensity ratio by $m_{ij} = \log_2(\frac{r_{ij}}{g_{ij}})$ and average log-intensity as $a_{ij} = \frac{1}{2}\log_2(\frac{r_{ij}}{g_{ij}})$ for j^{th} probe in i^{th} array.

2.1 Using SVR approach

Consider, we are given a microarray data $D = \{a_{ij}, m_{ij}\}_{j=1,2,\ldots,p}^{i=1,2,\ldots,n}$, where $a_{ij} \in R$ is a one dimensional scalar input for j^{th} probe in i^{th} array, $m_{ij} \in R$ is a one dimensional scalar output corresponds to the same, n is the total number of arrays and p is the total number of observed probes. In a linear case, m_{ij} is assumed to be linearly related as well as independent and identically distributed (*i.i.d.*) scalar corresponds to a_{ij} for i^{th} array. Then, a linear regression function using SVR approach can be expressed as follows:

$$\hat{m_{ij}} = f(a_{ij}) = w\ a_{ij} + b;\ j = 1, 2, \ldots, p;\ i = 1, 2, \ldots, n \tag{1}$$

where, w is a weight parameter and $b \in R$ is a scalar bias (or, model error). Now, we need to find out an optimal fitting of D. Depending upon chosen kernel, b may be implicitly part of a kernel function (Kecman, 2001). In a non linear case, a non linear regression function can be achieved by introducing the kernel trick principle of SVM (Scholkopf and Smola, 2001). A non linear SVR has the property to transform a complex regression function (in input space) into the comparatively simpler one (in high dimensional feature space). This is done by a non linear mapping $\phi(a_{ij}) : R \to R^q$, where q is usually equal to one or, higher dimensional space. Then, this regression function takes a general form suitable for both linear and non linear cases as given in 2:

$$\hat{m_{ij}} = f(a_{ij}) = \mathbf{w}_{(1\times q)}^t \phi_{(q\times 1)} + b;\ j = 1, 2, \ldots, p;\ i = 1, 2, \ldots, n \tag{2}$$

where, $a_{ij} \to \phi_{(q\times 1)}$ and $\phi_{(q\times 1)} = [\phi_1(a_{ij})\phi_2(a_{ij}) \ldots \phi_q(a_{ij})]^t$ be the weight parameter and the bias, respectively. Now, one can easily solve equation 2 using structural risk minimization principle of SVM (Vapnik, 1995). This may be achieved by employing different types of loss functions (Vapnik, 1998). Here, we have used the linear ϵ-insensitive

loss function (Vapnik, 1998) and obtained the result as follows:

$$L(m_{ij}, f(a_{ij})) = \mid m_{ij} - f(a_{ij}) \mid_{\epsilon} = \begin{cases} 0, \text{if } \mid m_{ij} - f(a_{ij}) \mid \leq \epsilon \\ \mid m_{ij} - f(a_{ij}) \mid -\epsilon, \text{otherwise} \end{cases} \tag{3}$$

where, $j = 1, 2, \ldots, p,\ \ i = 1, 2, \ldots, n$ and ϵ is a measure of the permissible deviation (see Figure 1). Our aim is now to find unknown parameters $\mathbf{w}_{(q\times 1)}$ and b, which minimize

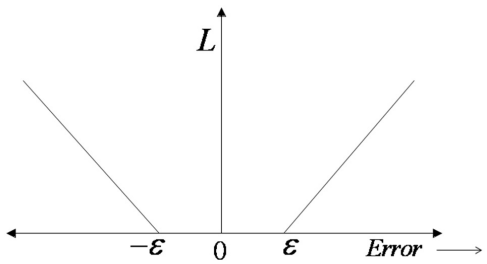


Figure 1: This shows the linear ϵ -insensitive loss function along with introducing slack vectors ξ and ξ^*

the empirical risk as follows:

$$R_{emp}(\mathbf{w}, b) = \frac{1}{p}\sum_{j=1}^{p} \mid m_{ij} - \mathbf{w}^t_{(1\times q)}\phi_{(q\times 1)} - b \mid_{\epsilon} \text{ such that } \min \mid\mid \mathbf{w}_{(1\times q)} \mid\mid^2 \tag{4}$$

This optimization is equivalent to a problem of finding $\mathbf{w}_{(q\times 1)}$ and b that minimizes the risk quantity in a primal space, i.e., in space of $\mathbf{w}_{(q\times 1)}$ and b. By introducing (non-negative) slack variables ξ_{ij} and ξ^*_{ij}, an equivalent form of equation 4 can be obtained as follows:

$$\begin{cases} L_p(\mathbf{w}, b, \xi, \xi^*) = \min\limits_{\mathbf{w}, b, \xi, \xi^*} \{\frac{1}{2}\mathbf{w}^t_{(1\times q)}\mathbf{w}_{(q\times 1)} + C\sum_{j=1}^{p}(\xi_{ij} + \xi^*_{ij})\} \\ \text{such that } \{m_{ij} - \mathbf{w}^t_{(1\times q)}\phi_{(q\times 1)} - b\} \leq (\epsilon + \xi_{ij}) \\ \{\mathbf{w}^t_{(1\times q)}\phi_{(q\times 1)} + b - m_{ij}\} \leq (\epsilon + \xi^*_{ij}) \\ \xi_{ij}, \xi^*_{ij} \geq 0; j = 1, 2, \ldots, p;\ i = 1, 2, \ldots, n \end{cases} \tag{5}$$

where, ξ_{ij} and ξ^*_{ij} have been used to measure the upper and lower deviation of training samples outside ϵ-insensitive zone, respectively (see Figure 1), ϵ is an error scalar and C is a positive scalar (regularization parameter), which determines a trade-off between degree of flatness of $f(a_{ij})$ and an amount, upto which a deviation larger than ϵ can be tolerated. From equation 5, a Lagrangian can be formed. By taking partial derivatives of the Lagrangian with respect to $\mathbf{w}, b, \xi, \xi^*, \alpha, \alpha^*, \beta, \beta^*$ ($\alpha, \alpha^*, \beta, \beta^*$: positive Lagrangian multipliers), subsequent computations have been performed (see for further details Burges (1998)). Finally, a dual formulation of the primal Lagrangian problem

(equation 5) has been obtained as follows :

$$L_D(\alpha_{ij}, \alpha^*_{ij}) = \max_{\alpha,\alpha^*}\{\frac{1}{2}(\alpha_{ij} - \alpha^*_{ij})(\alpha_{ij} - \alpha^*_{ij})\ \phi^t_{(1\times q)}\phi_{(q\times 1)} + \sum_{j=1}^{p} m_{ij}(\alpha_{ij} - \alpha^*_{ij}) - \sum_{j=1}^{p} \epsilon\ (\alpha_{ij} - \alpha^*_{ij})\} \quad (6)$$

such that $\sum_{j=1}^{p}(\alpha_{ij} - \alpha^*_{ij}) = 0$ and $\alpha_{ij}, \alpha^*_{ij} \in [0, C];\ j = 1, 2, \ldots, p;\ i = 1, 2, \ldots, n.$

A solution of $\alpha_{ij}, \alpha^*_{ij}$ in (equation 6) can be obtained using any standard quadratic algorithm solver. Further, defining $\mathbf{K}_\sigma(a_{ij}, a^t_{ij}) = \phi^t_{(1\times q)}\phi_{(q\times 1)}$, the non linear regression function (equation 2) can be further rewritten as:

$$\begin{cases} \hat{m_{ij}} & = f(a_{ij})_{(1\times 1)} = \mathbf{w}^t_{(1\times q)}\phi_{(q\times 1)} + b \\ & = \sum_{j=1}^{p}(\alpha_{ij} - \alpha^*_{ij})\mathbf{K}_\sigma(a_{ij}, a^t_{ij}) + b \end{cases} \quad (7)$$

where, $\sum_{j=1}^{p}(\alpha_{ij} - \alpha^*_{ij})\ \phi_{(q\times 1)}$ and $\mathbf{K}_\sigma(a_{ij}, a^t_{ij})$ is a kernel function satisfying the Mercer conditions (Smola et al., 1996). For given kernel parameters (C, ϵ), b can be calculated using Karush-Kuhn-Tucker (KKT) conditions as follows:

$$b = \frac{1}{p}\sum_{j=1}^{p} \mid m_{ij} - f(a_{ij}) \mid \quad (8)$$

where $f(a_{ij})$ is defined in (equation 2). In this study, we have considered RBF kernel with width σ. As RBF function implicitly defines bias, there of b can be removed directly from equation 1. Moreover, one can also avoid estimation of b using RBF kernel, whose functional form $\mathbf{K}_\sigma$ is as follows:

$$\mathbf{K}_\sigma(a_{ij}, a^t_{ij}) = \exp(-\frac{1}{2} \mid\mid a_{ij} - a^t_{ij} \mid\mid^2) \quad (9)$$

2.2 Using SVMQR approach

Quantile regression is a robust approach for estimating the conditional quantile of a distribution based on a covariate information. The basic idea behind quantile regression arises from observations that minimizing the l_1 loss function for a location estimator, which yields the median. l_1 loss function is basically minimizing the sum of the absolute differences between the target value m_{ij} and estimated values $f(a_{ij})$. It is also known as least absolute error (or, deviation). However, this idea can also be generalized to obtain regression estimates for any quantile (Koenker and Park, 1996). Incorporating the principle of this regression into SVR, SVMQR has been implemented (Takeuchi et al., 2005). SVMQR may be thought as extended version of SVR that involves the estimation of conditional quantile rather than mean. The modification in SVR is done by defining a quantile based loss function, instead of taking linear ϵ-insensitive loss function

in equation 3. The conditional estimates of τ -quantile is derived using the following loss function:

$$\rho_\tau(\xi_{ij}) = \begin{cases} \tau\ \xi_{ij};\ \text{if } \xi_{ij} \geq \xi_{ij} \\ (\tau - 1)\ \xi_{ij};\ \xi_{ij} < 0 \end{cases} \quad \text{where,} \tau \in (0,1) \tag{10}$$

An alternative way (Koenker and Park, 1996) is to use a check function, $\rho_\tau(\xi_{ij})$, which is expressed as below:

$$\rho_\tau(\xi_{ij}) = \tau\ I(\xi_{ij} \geq \xi_{ij}) + (\tau - 1)\ I(\xi_{ij} < 0) \tag{11}$$

where, $I(.)$ is the indicator function. If $\xi_{ij} \geq \xi_{ij}$ is true, $I(.) = 1$ is taken; otherwise $I(.) = 0$. Onwards, we need to find unknown parameters $\mathbf{w}_{(q\times 1)}$ and b of equation 12 that minimize the empirical risk defined in equation 4:

$$\begin{cases} R_{emp}(\mathbf{w}, b) = \frac{1}{p} \sum_{j=1}^{p} \mid m_{ij} - \mathbf{w}_{(1\times q)}^t \phi_{(q\times 1)} - b \mid_\epsilon, \tau \in (0,1) \\ \text{subject to } \min \mid\mid \mathbf{w}_{(1\times q)} \mid\mid^2 \end{cases} \tag{12}$$

For simplicity of the formulation, we have considered the ϵ-insensitive SVR ($\epsilon = 0$). Therefore, equation 4 can be restructured as follows:

$$\begin{cases} L_p(\mathbf{w}, b, \xi, \xi^*) = \min\limits_{\mathbf{w}, b, \xi, \xi^*} \{\frac{1}{2} \mathbf{w}_{(1\times q)}^t \mathbf{w}_{(q\times 1)} + C \sum_{j=1}^{p} (\xi_{ij} + \xi_{ij}^*)\};\ \tau \in (0,1) \\ \text{subject to } \{m_{ij} - \mathbf{w}_{(1\times q)}^t \phi_{(q\times 1)} - b\} \leq \xi_{ij} \\ \{\mathbf{w}_{(1\times q)}^t \phi_{(q\times 1)} + b - m_j\} \leq \xi_{ij}^* \\ \xi_{ij}, \xi_{ij}^* \geq 0; j = 1, 2, \ldots, p,\ i = 1, 2, \ldots, n. \end{cases} \tag{13}$$

Similar to equation 6, a dual formulation of this problem can easily be obtained by introducing α, α^* positive Lagrangian multipliers into equation 13 (Takeuchi et al., 2005). Now, taking partial derivatives with respect to $\mathbf{w}, b, \xi, \xi^*, \alpha, \alpha^*$, a new solution can be derived as follows:

$$\begin{aligned} L_D(\alpha_{ij}, \alpha_{ij}^*) = \max_{\alpha, \alpha^*} \{ & \frac{1}{2} (\alpha_{ij} - \alpha_{ij}^*)(\alpha_{ij} - \alpha_{ij}^*) \phi_{(1\times q)}^t \phi_{(q\times 1)} \\ & + \sum_{j=1}^{p} m_{ij} (\alpha_{ij} - \alpha_{ij}^*) \end{aligned} \tag{14}$$

$$\text{subject to } \sum_{j=1}^{p} (\alpha_{ij} - \alpha_{ij}^*) = 0 \text{ and } \alpha_{ij}, \alpha_{ij}^* \in [(1-\tau)C, \tau C]$$

$$\text{where } j = 1, 2, \ldots, p;\ i = 1, 2, \ldots, n \text{ and } \tau \in (0,1).$$

This regression solution takes similar form as equation 7, the only difference is in the ranges of Lagrangian multiplier of the dual formulation in equation 14. Similar result has been obtained by Sohn et al. (2008). Further, equation 14 can also be solved using any standard quadratic programming. In addition, this equation satisfies the quantile property of having a fraction of $\tau = 0.5$ points on either side of the regression as well

as capability of tracking the observations more or less closely. Also, here we have used RBF kernel. We have implemented both these methods, namely SVR and SVMQR, in R environment using 'kernlab' package (Karatzoglou et al., 2004). Moreover, the practical estimates, defined in equation 6 and 14, requires a procedure for setting the regularization parameters, which has been discussed in the following section.

2.3 Regularization parameters (σ, C, ϵ) selection approach

The user defined parameters (σ, C, ϵ) need to be tuned by an analyst (Vapnik, 1998). However, we have applied the analytical selection method to choose those parameters automatically along with 10-cross validation techniques (see for more details (Cherkassky and Ma, 2004)). This parameter selection is quite powerful (Cherkassky and Ma, 2004; Zhaoa et al., 2013). Our R-script is also capable of generating such parameter values.

2.4 Using LOWESS regression approach

One of the most commonly used normalization techniques is LOWESS approach that utilizes principle of least squares regression by fitting simple models to localized data-subsets, and also yields flexibility of a nonlinear regression (Cleveland et al., 1992). The basic idea of this approach is that the regression function $f(a_{ij})$ (as defined in equation 1) can be locally approximated by fitting a linear least square regression surface to the data points within chosen neighborhood points. However, weighed least squares may also be used to fit linear or quadratic functions.

2.5 Using KERNEL regression approach

Another most widely used procedure in nonparametric curve estimator can be applied in the form of NadarayaWatson kernel regression (Wand and Jones, 1995). The strategy is to fit the function $f(a_{ij})$ (as defined in equation 1) within chosen neighborhood points by locally fitting a certain degree polynomial function via weighted least squares principle. This prediction completely depends on past observations. Here, we have used the Gaussian Kernel, because of its centered mean characteristic and symmetric property.

2.6 Implementation of normalization methods

In this section, we have mainly formulated rules of normalization based on the above mentioned approaches. The key idea is that each m_{ij} needs to be normalized by subtracting the corresponding regression fitted value a_{ij}. Therefore, the normalized log-intensity ratios n_{ij} are basically estimated residuals from average of their respective regression fits. Consequently, we may define rules using the following approaches:

I. SVR normalization:

$$n_{ij} = m_{ij} - \mathbb{SVR}(a_{ij}) \tag{15}$$

where, $\mathbb{SVR}(a_{ij})$ are fitted values corresponds to a_{ij} for j^{th} probe of i^{th} array using SVR.

II. SVMQR normalization:

$$n_{ij} = m_{ij} - \frac{1}{2}\{\mathbb{SVR}_{\tau}(a_{ij}) + \mathbb{SVR}_{(1-\tau)}(a_{ij})\} \tag{16}$$

where, $\mathbb{SVR}_{\tau}(a_{ij})$ and $\mathbb{SVR}_{(1-\tau)}(a_{ij})$ are τ^{th} and $(1-\tau)^{th}$ quantile SVMQR fitted values corresponds to a_{ij} for j^{th} probe of i^{th} array.

III. LOWESS normalization:

$$n_{ij} = m_{ij} - \mathbb{LOWESS}(a_{ij}) \tag{17}$$

where, $\mathbb{LOWESS}(a_{ij})$ are fitted values corresponds to a_{ij} for j^{th} probe of i^{th} array using LOWESS.

IV. KERNEL normalization:

$$n_{ij} = m_{ij} - \mathbb{KERNEL}(a_{ij}) \tag{18}$$

where, $\mathbb{KERNEL}(a_{ij})$ are fitted values corresponds to a_{ij} for j^{th} probe of i^{th} array using KERNEL.

2.7 Methods to assess proposed normalizations

2.7.1 Computation of variance within replicate groups

To assess the efficiency of each normalization method, we have considered the variance within the replicate groups, as defined in Workman et al. (2002). The estimator is as follows:

$$\hat{\sigma}^2 = \frac{1}{n(p-1)} \sum_{i=1}^{n} \sum_{j=1}^{p} \{\log_2(a_{ij}) - \log_2(a_{i.})\}^2 \tag{19}$$

where, i is an index over n probe intensities, j is an index over p replicates and $a_{i.} = \sum_{j=1}^{p} a_{ij}$. However, smaller variance within replicates, a normalization method provide better solution.

2.7.2 Identification of differential genes

Identification of differential genes from these normalized data can also be used as an alternative method to assess their efficiencies. In this study, we have chosen non-parametric empirical Bayes (*eBayes*) approach to identify genes, those are differentially expressed (Smyth et al., 2004).

3 Results and Discussions

In this section, we have assessed all these normalization methods using three simulated model based expression datasets and apoAI-K mouse model real dataset (Dudoit et al., 2002). Further, we have also evaluated SVR as well as SVMQR methods with respect to other standard normalizations, namely LOWESS and KERNEL methods.

3.1 Performance evaluation of SVM based normalization methods with simulated data

Three simulated microarray datasets have been generated using a simulation model, proposed by Balagurunathan et al. (2002); each of which contains six thousands spots. All the true expression intensity readings for red and green dyes have been generated using exponential distribution with parameter $\lambda = \frac{1}{3000}$, and 5% of total spots have further been simulated as outliers using a Beta distribution with parameters $\beta(1.7, 4.8)$. These outliers are differential genes. In addition, we have applied a non linear transformation function to intensity values, and also added random noises those follow Gaussian distribution. The parameters for red and green dyes of three simulated models used are as follows: $(a_0^1 = 0, a_1^1 = 100^{\frac{1}{0.9}}, a_2^1 = -0.9, a_3^1 = 1)$ and $(a_0^2 = 0, a_1^2 = 100^{\frac{1}{0.7}}, a_2^2 = -0.7, a_3^2 = 1)$ for sinusoid shape, $(a_0^1 = 0, a_1^1 = 500, a_2^1 = -1, a_3^1 = 1)$ and $(a_0^2 = 0, a_1^2 = 10, a_2^2 = -1, a_3^2 = 1)$ for banana shape and, $(a_0^1 = 0, a_1^1 = 10, a_2^1 = -1, a_3^1 = 1)$ and $(a_0^2 = 0, a_1^2 = 100^{\frac{1}{0.7}}, a_2^2 = -0.7, a_3^2 = 1)$ for mixed shape, respectively. Each simulation model has been generated for 100 times. Finally, we have obtained three simulated microarray models, each having 30 slides along with six thousands spots. We have applied SVR and SVMQR to those three models, which might be thought of representatives to true gene expression data, and observed their performances using the above mentioned evaluation criteria. We have used analytical selection technique for setting up regular-

Table 1: This table represents variance/ mean-variance estimates within the replicate groups for simulated as well as real dataset

Normalization	apoAI-K mouse model	dataset for banana shape	dataset for sinusoid shape	dataset for mixed shape
No normalization	0.1615	0.1251	0.1123	0.122
SVR normalization	0.1071	**0.0385**	**0.0352**	**0.0364**
SVMQR normalization	0.1069*	0.0388*	0.0356*	0.0367*
LOWESS normalization	0.1069	0.0668	0.0502	0.07
KERNEL normalization	0.1228	0.0566	0.0445	0.0579

$*$ *Best fitted results obtained from SVMQR normalization is at* $\mu = 0.05$.

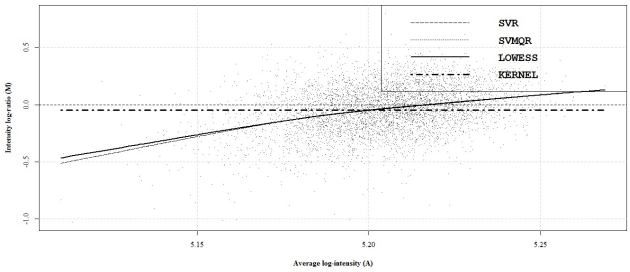
ization parameters of both methods, namely SVR and SVMQR, and 10-cross validation technique has been applied to generalize these estimates. The tolerance levels has been fixed at 0.001, and lowest variance within replicates at $\mu = 0.05$ has been achieved for all

these models. We have found that both of them have successfully reduced non-biological variabilities from these model datasets by more than 50% (see Table 1). The estimated mean-variance within replicates ($\hat{\sigma}^2$) are computed as 0.1251 (for banana shape), 0.1123 (for sinusoid shape) and 0.1220 (for mixed shape), respectively. Fitting with SVR, $\hat{\sigma}^2$ has been reduced by 52.93% (for banana shape), 52.27% (for sinusoid shape) and 54.04% (for mixed shape), respectively. Whereas, fitting with SVMQR at ($\mu = 0.05$), 52.65% (for banana shape), 51.86% (for sinusoid shape) and 53.75% (for mixed shape) reduction in such estimates have been observed. It is desirable that a good normalization approach should yield unbiased corrections and smaller variance estimates. This implies both SVR and SVMQR are efficiently capable of reducing non-biological variabilities from a raw gene expression data. Afterwards, we have proceed to assess whether the SVM based

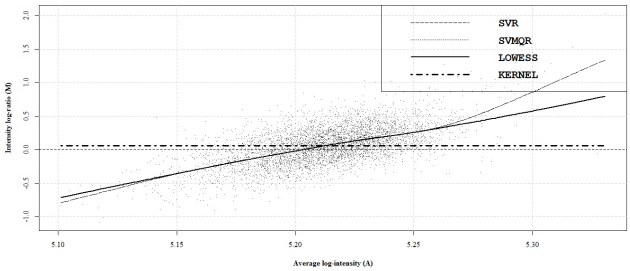
Table 2: This table contains percentage of identified genes, out of 300 truly expressible differential genes, from different normalized data using *eBayes* method at 1% and 5% significance level

Normalization	**Model for banana shape**		**Model for sinusoid shape**		**Model for mixed shape**	
	at 1% level	*at 5% level*	*at 1% level*	*at 5% level*	*at 1% level*	*at 5% level*
No normalization	27	35	20	30	36.33	37.33
SVR normalization	48	53	45.67	50.33	47.33	48.33
SVMQR normalization	**51.33**	**55.4**	**46.33**	**52.67**	**49.3**	**51.6**
LOWESS normalization	44.67	50.67	42.67	45.67	47.33	51
KERNEL normalization	38.33	39.33	37.33	37.33	38	38.67

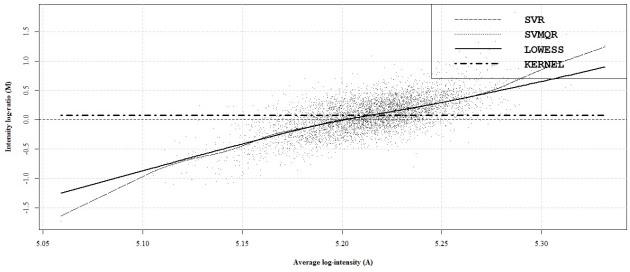
normalized model data have the capability to identify significant numbers of differential genes. Therefore, *eBayes* gene selection method has been applied to these data. Each simulated model dataset contains 300 differential genes, which has been generated using beta distribution, and may also be thought of truly expressible differential genes. It has been found that about 144 (for banana shape), 137 (for sinusoid shape) and 142 (for mixed shape) highly significant genes, out of those 300 differential genes, could be identified at 1% significance level, and 159 (for banana shape), 151 (for sinusoid shape) and 145 (for mixed shape) differential genes could be identified at 5% significance level from SVR based normalized datasets. In contrast, SVMQR based such datasets (i.e. replace such by normalized) have identified 154, 139, 148 genes (at 1% level), and 166, 158, 149 genes (at 5% level) for banana, sinusoid and mixed shape, respectively. These mean that two proposed methods have found to be superior, in comparison to LOWESS and KERNEL approaches (see Table 1 and 2). We have also drawn MA plot as a diagnostic tool, which can be used to assess different normalization methods, visually or/and how well data have been normalized. Figure 2 represents MA plot of log-intensities of raw data for first slide along with fitted lines based on SVR and SVMQR, LOWESS and KERNEL methods for simulated model dataset (2(a)) banana shape, (2(b)) sinusoid shape, and (2(c)) mixed shape. Figure 2 also depicts that the fitted values could be populated on MA plots using all these above approaches.



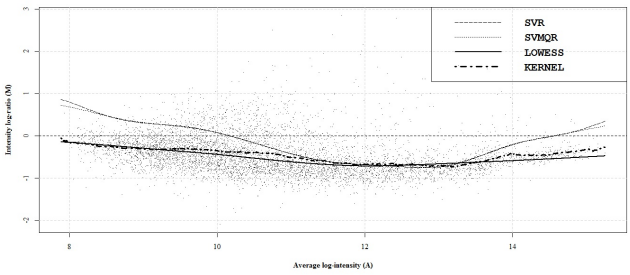
(a) banana shape



(b) sinusoid shape



(c) mixed shape



(d) apoAI-K dataset

Figure 2: This figure represents MA plot representing log-intensities of raw data for first slide along with fitted lines based on SVR and SVMQR, LOWESS and KERNEL methods for simulated model dataset (a) banana shape, (b) sinusoid shape, (c) mixed shape, and (d) apoAI-K dataset, respectively.

3.2 Performance evaluation with real data

The apoAI knockout mouse model (apoAI-K) dataset has been retrieved from `http://www.stat.berkeley.edu/users/terry/zarray/Html/matt.html` (Dudoit et al., 2002). The apoAI-K dataset contains the Cy3 and Cy5 intensities of two channels for 16 mice; out of which eight mice have the apoAI knocked-out gene (considered as treatment group) and rest are normal C57B l/6 mice (as control group). Each array consists of 6384 spots, including 257 genes (spots) thought to be related to lipid metabolism. This dataset has been already analyzed in several previous studies (Callow et al., 2000; Dudoit et al., 2002). So, those results have been helped to assess their consistencies, and also validated our results. We have applied both SVR and SVMQR to the apoAI-K mouse

Table 3: This table contains the names of the genes identified as differentially genes by Dudoit et al. (2002) in the apoAI-K mouse model dataset along with the row numbers of gene expression matrix

Row Number	Gene Name
540	EST, Highly similar to APOLIPOPROTEIN A-I PRECURSOR [Mus musculus], lipid-UG
1496	EST
1739	Apo CIII, lipid-Img
2149	Apo AI, lipid-Img
2537	ESTs, Highly similar to APOLIPOPROTEIN C-III PRECURSOR [Mus musculus], lipid-UG
4139	EST, Weakly similar to C-5 STEROL DESATURASE [Saccharomyces cerevisiae], lipid-UG
4941	Similar to yeast sterol desaturase, lipid-Img
5356	CATECHOL O-METHYLTRANSFERASE, MEMBRANE-BOUND FORM, Brain-Img

model. Unlike earlier case, we have followed analytical selection technique for setting up regularization parameters, and rest of those steps. SVMQR has achieved lowest variance ($\hat{\sigma}^2 = 0.1069$) within replicates at $\mu = 0.05$. In addition, we have chosen box plot as a diagnostic plot, which has been employed to compare different approaches, visually or/and check whether artifacts have been removed from data or, not. The box plot (in figure 3) depicts that artifacts have been removed using SVR and SVMQR methods, nicely. However, we also have alternative options to choose diagnostic tools such as MA plot, density plot, heat map, relative log expression (RLE) plot, histogram, spatial plot and dendrogram. Consequently, results have showed that SVR and SVMQR have performed, efficiently (see Table 1). The estimated variance within the replicate groups for raw data has been 0.1615. But, fitting with SVR method, $\hat{\sigma}^2$ has been reduced by 20.25%, while fitting with SVMQR method ($\mu = 0.05$), has resulted 20.34% reduction in such estimate. Using second evaluation criterion, we have checked the capability of those methods identifying 8 differential genes, whose clone sequences have been biologically verified and validated (Dudoit et al., 2002). Since the underlying biology of the mouse model experiment has already been well understood, we can use this information to assess their normalized qualities. Therefore, we have applied the *eBayes* gene selection method. The names of those genes identified as differentially genes in the apoAI-K mouse model dataset along with the row numbers of gene expression matrix are listed

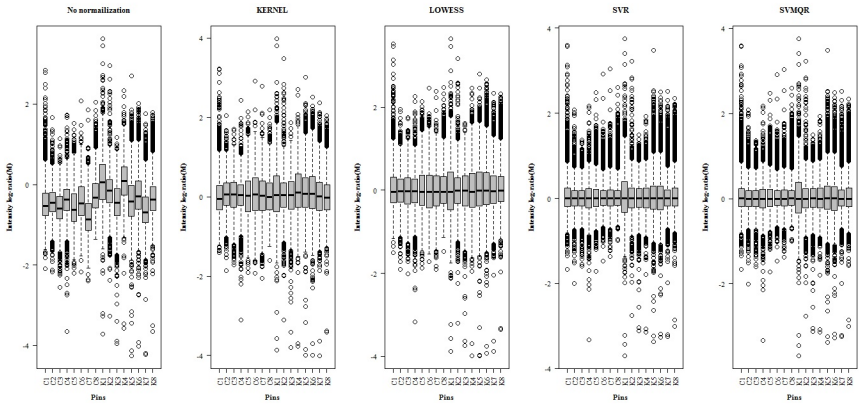


Figure 3: In this figure, the box plots represents log-intensities of raw data, normalized data using SVR, SVMQR, LOWESS and KERNEL methods, respectively for apoAI-K dataset. The first box plot shows that raw dataset contains so much variation within arrays. But, after normalization using any methods, such variation can be removed. Both SVR and SVMQR performed as our expectation.

in Table 3. From Table 4, it has been clarified that both SVR and SVMQR have the ability to identify those genes, and further yielded highly significant adjusted p-values at 1% level of significance. Moreover, Table 4 reflects that other standard approaches, namely LOWESS and KERNEL, are effective to identify the eight benchmark gene-list and have achieved highly significant adjusted p-values. Figure (2(d)) represents MA plot of log-intensities of raw data for first slide along with fitted lines based on appearances, namely SVR, SVMQR, LOWESS and KERNEL, respectively. Also, it assures the successful fitting of those values on the MA plot using all approaches.

4 Conclusion

The basic aim of normalization is to remove the non-biological variations as much as possible from expression data. In this study, the performance of SVR and SVMQR methods with respected to other standard ones, namely LOWESS and KERNEL, have been compared. It has been observed that minimum estimated mean variance within replications and identification of maximum number of differential genes could be possible using our proposed approaches. Box plot of mouse model data also depicts that artifacts could successfully be removed using those approaches. Further, to assess the efficiency of each method, the variance within the replicate groups and identification of differential genes were considered. In identification process, the *eBayes* approach was chosen due to the *eBayes* statistic provides more stable inference than others, namely analysis of variance, Welch t- statistic, significance analysis of microarrays modified t-statistic, and permutation t- statistic. Furthermore, this approach works well, even if the number of replicates is small and has also the flexibility to compare multiple groups simultaneously in arbitrary complicated designed experiments. The present study clearly indicates

Table 4: This table illustrates a comparison of the differentially expressed genes in the apoAI-K mouse model identified from different normalized methods. The p-values are computed using *eBayes* static and FDR correction are made to adjust the p-values. All genes are confirmed by biological methods.

Row Number	No normalization	SVR	SVMQR	LOWESS	KERNEL
540	1.12e-06	1.22e-07	1.19e-07	2.03e-07	3.05e-07
1496	**2.43e-02***	5.90e-04	5.83e-04	2.42e-05	1.21e-05
1739	1.81e-04	4.64e-05	4.01e-05	9.97e-08	3.05e-07
2149	6.70e-11	8.02e-11	7.77e-11	2.91e-12	1.80e-11
2537	6.15e-04	3.78e-06	3.65e-06	2.42e-05	1.20e-04
4139	6.15e-04	2.88e-06	2.82e-06	4.37e-06	4.77e-06
4941	6.31e-03	8.02e-04	7.95e-04	2.42e-05	1.34e-04
5356	2.92e-06	6.54e-07	6.34e-07	4.84e-08	1.08e-07

* *Genes are not identified as differential genes from the respective normalized dataset at 1% level of significance and e is the Euler's Number*

that SVM based normalization approaches are competent with other standard methods. These may efficiently be employed to gene expression studies.

Acknowledgment

We thank the anonymous reviewers for their valuable comments and constructive suggestions, which helped to improve this paper.

References

Balagurunathan, Y., Dougherty, E. R., Chen, Y., Bittner, M. L., and Trent, J. M. (2002). Simulation of cdna microarrays via a parameterized random signal model. *Journal of Biomedical Optics*, 7(3):507–523.

Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167.

Callow, M. J., Dudoit, S., Gong, E. L., Speed, T. P., and Rubin, E. M. (2000). Microarray expression profiling identifies genes with altered expression in HDL deficient mice. *Genome Research*, 10:2022–2029.

Cherkassky, V. and Ma, Y. (2004). Practical selection of svm parameters and noise estimation for svm regression. *Neural networks*, 17(1):113–126.

Cleveland, W. S., Grosse, E., and Shyu, W. M. (1992). Local regression models. *Statistical models in S*, pages 309–376.

Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. *Statistica sinica*, 12(1):111–140.

Fujita, A., Sato, J. R., Rodrigues, L. O., Ferreira, C. E., and Sogayar, M. C. (2006). Evaluating different methods of microarray data normalization. *BMC bioinformatics*, 7(1):469.

Huber, W., Von Heydebreck, A., Sültmann, H., Poustka, A., and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18(suppl 1):S96–S104.

Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. (2004). kernlab- an s4 package for kernel methods in R. *Journal of Statistical Software*, 11(9):1–20.

Kecman, V. (2001). *Learning and soft computing: support vector machines, neural networks, and fuzzy logic models*. MIT press.

Koenker, R. and Park, B. J. (1996). An interior point algorithm for nonlinear quantile regression. *Journal of Econometrics*, 71(1):265–283.

Park, T., Yi, S.-G., Kang, S.-H., Lee, S., Lee, Y.-S., and Simon, R. (2003). Evaluation of normalization methods for microarray data. *BMC bioinformatics*, 4(1):33.

Scholkopf, B. and Smola, A. J. (2001). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Smola, A. J. et al. (1996). Regression estimation with support vector learning machines. *Master's thesis, Technische Universit at M unchen*.

Smyth, G. K. et al. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1):3.

Smyth, G. K. and Speed, T. (2003). Normalization of cdna microarray data. *Methods*, 31(4):265–273.

Sohn, I., Kim, S., Hwang, C., and Lee, J. W. (2008). New normalization methods using support vector machine quantile regression approach in microarray analysis. *Computational Statistics & Data Analysis*, 52(8):4104–4115.

Stekel, D. (2003). *Microarray bioinformatics*. Cambridge University Press.

Takeuchi, I., Le, Q. V., Sears, T., and Smola, A. J. (2005). Nonparametric quantile regression estimation. *Journal of Machine Learning Research*, 7:1001–1032.

Taniguchi, M., Miura, K., Iwao, H., and Yamanaka, S. (2001). Quantitative assessment of dna microarrays comparison with northern blot analyses. *Genomics*, 71(1):34–39.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag, New York.

Vapnik, V. (1998). *Statistical Learning Theory*. John Wiley and Sons, Inc., New York.

Wand, M. P. and Jones, M. C. (1995). *Kernel smoothing*. Hapman and Hall, first edition edition. 114-141.

Wang, J., Ma, J. Z., and Li, M. D. (2004). Normalization of cdna microarray data using wavelet regressions. *Combinatorial Chemistry & High Throughput Screening*, 7(8):783–791.

Workman, C., Jensen, L. J., Jarmer, H., Berka, R., Gautier, L., Nielser, H. B., Saxild, H.-H., Nielsen, C., Brunak, S., and Knudsen, S. (2002). A new non-linear normalization

method for reducing variability in dna microarray experiments. *Genome biol*, 3(9):1–16.

Yang, Y. H. and Speed, T. P. (2003). *Statistical Analysis of Gene Expression Microarray Data: Chapter-Design and analysis of comparative microarray experiments*. Chapman and Hall/CRC Press. 35-91.

Zhaoa, W., Taoa, T., and Ziob, E. (2013). Parameters tuning in support vector regression for reliability forecasting. *Chemical Engineering*, 33:523–528.