Bayes Wavelet Regression Approach to Solve Problems in Multivariable Calibration Modeling

Setiawan[1] and Sutikno[1]

***Abstract*—In the multiple regression modeling, a serious problems would arise if the independent variables are correlated among each other (the problem of ill conditioned) and the number of observations is much smaller than the number of independent variables (the problem of singularity). Bayes Regression (BR) is an approach that can be used to solve the problem of ill conditioned, but computing constraints will be experienced, so pre-processing methods will be necessary in the form of dimensional reduction of independent variables. The results of empirical studies and literature shows that the discrete wavelet transform (WT) gives estimation results of regression model which is better than the other pre-processing methods. This experiment will study a combination of BR with WT as pre-processing method to solve the problems ill conditioned and singularities. One application of calibration in the field of chemistry is relationship modeling between the concentration of active substance as measured by High Performance Liquid Chromatography (HPLC) with Fourier Transform Infrared (FTIR) absorbance spectrum. Spectrum pattern is expected to predict the value of the concentration of active substance. The exploration of Continuum Regression Wavelet Transform (CR-WT), and Partial Least Squares Regression Wavelet Transform (PLS-WT), and Bayes Regression Wavelet Transform (BR-WT) shows that the BR-WT has a good performance. BR-WT is superior than PLS-WT method, and relatively is as good as CR-WT method.**

Keywords—**Bayes, wavelet, ill conditioned, singularity**

I. Introduction

In general, the statistical model requires the number of observations greater than the number of independent variables (predictors). On some phenomena (cases) these requirements can't always be fulfilled. One of the reasons is the high cost that is used to obtain the data, so it does not allow for observation of samples in large numbers. Trooped with these problems there are also other problems on the occurrence of the independent variables of high correlation among independent variables. If there is a high correlation between independent variables *X*, there will present an classical issue called multicollinearity, hence the use of Classical Regression (Ordinary Least Squares Regression, OLS) will lead to ill-conditioned raw alleged error resulting enlarged (over estimate). In other words, multicol-linearity can cause a very low accuracy of the estimated parameters [8]. Meanwhile, if the number of independent variables is much greater than the number of observations, the structure becomes singular matrix of independent variables (the problem of singularity). This results affect in X^TX matrix has no unique inverse (typical), which is the main requirement in OLS. Consequently, a method to solve both of problems is necessarily obtained.

Ill-conditioned and the singularity problem occur in many real problems, for example in the calibration model. Calibration models are widely used in chemistry, especially Chemo metrics that is a field of science that are a combination of mathematics, statistics, and chemistry.

Several methods to solve the problem of ill conditioned and the singularity has been developed. Setiawan and Notodiputro (2007b) developed a method with the combination of Wavelet Transformation Continuum Regression (CR-WT) which resulted in a relatively satisfactory model to solve the problem of ill conditioned and the singularity on various structures of the correlation matrix independent variables. Bayes Regression method is one of alternative to solve the multi-collinearity problems. This paper will discuss about how the performance of a combination of Bayes Regression and Wavelet Transform. The focus of this research is to examine the Bayes Regression with pre-processing Wavelet Transform (BR-WT) as one alternative for dealing with the high correlation between independent variables and the number of observations is smaller than the number of independent variables. This is because both problems are common in real problems. This paper is based on the results of research that specifically aims to study and develop Bayes Regression methods with Wavelet Transform to solve the problem of high correlation between independent variables and the number of observations is smaller than the number of independent variables.

Recently, the use of medical plants is not only limited in the making of herbal medicine, but also in pharmacy, supplement products (nutraceuticals), herbal extracts, etc. the quality of the source (medical plants) must be kept in order to keep the quality of the products of the herbal medicine company and of the pharmacy and in order to fulfill the standard (Danutirto, 2001). The information of the use of the medical plants can be observed through the active compounds within them. Therefore, the study about the active compounds is necessary.

II. Theories

A. Multicollinearity

In the classical regression model requires the number of observations greater than the number of independent variables (predictors) and there is no multicolinearity. Multicolinerity is high correlation between its indepen-dent variables. One of the methode that can be used to

[1] Setiawan and Sutikno are with Department of Statistics, FMIPA, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia. E-mail: setiawan@statistika.its.ac.id.

identify multicolinearity is value of Varians Inflation Factor (VIF).

$$\text{VIF}_{x_j} = \frac{1}{1 - R_j^2}$$

Where R_j^2 is determination coefficient for Xj that constitute function from other X variable. If the VIF's value is greater than 10 shows multicolinearity.

B. Bayes Regression with Normal Prior

Bayes approach in the regression is done by forming the posterior distribution of parameters. This posterior is the product of priors with the likelihood function. General multiple regression model with k independent variables (including intercept) is:

$$y = X\beta + \varepsilon \quad (1)$$

Where
y = observation of the dependent variable vector (n x 1)
X= observation matrix of independent variables (n x k)
β = regression coefficient vector (kx1)
ε = random variable vector error (nx1)
and $\underline{\varepsilon} \sim N(0, I\sigma^2)$ then y~ $N(X\beta, I\sigma^2)$.

In this paper assumed β ~ N (θ, V) where V is a variance-covariance matrix β so symmetrical, then the prior function:

$$p(\beta) \propto (2\pi)^{-k/2}|V|^{-1/2}\exp\left\{-\frac{1}{2}(\beta-\theta)^T V^{-1}(\beta-\theta)\right\} \quad (2)$$

Likelihood function of the normal multiple regression model is:

$$l(y|\beta) \propto \frac{1}{\sigma^n}(2\pi)^{-n/2}\exp\left[-\frac{1}{2\sigma^2}(y-X\beta)^T(y-X\beta)\right] \quad (3)$$

Multiplication of priors with the likelihood function produces the following joint posterior distribution:

$$h(\beta, y) \propto l(y \mid \beta).p(\theta)$$

$$\propto \frac{1}{\sigma^n}(2\pi)^{-(k+n)/2}|\mathbf{V}|^{-1/2}\exp\left[-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\theta})^T(\boldsymbol{\beta}-\boldsymbol{\theta}) - \frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\right]$$

or

$$\log h(\beta, y) \propto \left\{-\frac{k+n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2}\log|V|\right\} - \frac{1}{2}\left\{(\beta-\theta)^{T'}V^{-1}(\beta-\theta) + \frac{1}{\sigma^2}(y-X\beta)^{T'}(y-X\beta)\right\}$$

$$\propto \left\{-\frac{k+n}{2}\log(2\pi) - n\log\sigma - \frac{1}{2}\log|\mathbf{V}|\right\} - \frac{1}{2}\left\{(\boldsymbol{\beta}-\boldsymbol{\theta})'\mathbf{V}^{-1}(\boldsymbol{\beta}-\boldsymbol{\theta}) + \frac{1}{\sigma^2}(\mathbf{y}^T\mathbf{y} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta})\right\} \quad (4)$$

The maximum value of the function is obtained by finding the first derivative of β and equated to zero.

$$2V^{-1}\hat{\beta} - 2V^{-1}\theta + \frac{1}{\sigma^2}\left(-2X^Ty + 2X^TX\hat{\beta}\right) = 0$$

$$\hat{\beta} = \left(V^{-1} + \frac{1}{\sigma^2}X^TX\right)^{-1}\left(V^{-1}\theta + \frac{1}{\sigma^2}X^Ty\right) \quad (5)$$

where V is variance-covariance matrix β, θ is vector of regression parameters which are approximated by least squares regression coefficient,σ^2 = the middle of square error which is approximated by the middle of square error of the least square regression.

Statistical properties of $\hat{\beta}$ and $\hat{y}$:

1. $\hat{\beta}$ is a bias estimator for β

$$E(\hat{\beta}) = E\left(\left(V^{-1} + \frac{1}{\sigma^2}X'X\right)^{-1}\left(V^{-1}\theta + \frac{1}{\sigma^2}X'y\right)\right)$$

$$E(\hat{\boldsymbol{\beta}}) = E\left(\mathbf{V}^{-1} + \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{V}^{-1}\boldsymbol{\theta} + \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{y}\right)$$

$$= \left[\left(\mathbf{V}^{-1} + \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{V}^{-1}\boldsymbol{\theta} + \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}\right)\right]$$

2. Variance and standard deviation for $\hat{\beta}$ and $\hat{y}$

$$\text{Var}(\hat{\beta}) = \text{Var}\left[\left(V^{-1} + \frac{1}{\sigma^2}X^TX\right)^{-1}\left(V^{-1}\theta + \frac{1}{\sigma^2}X^Ty\right)\right]$$

$$= \frac{1}{\sigma^2}\left(\mathbf{V}^{-1} + \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}\left(\mathbf{V}^{-1} + \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}\right)^{-1}$$

3. Var(ŷ) is expected by the middle of the square error, so that the estimation of $\text{Var}(\hat{\beta})$ is :

$$\frac{1}{\hat{\sigma}^2}\left(V^{-1} + \frac{1}{\sigma^2}X^TX\right)^{-1}X^TX\left(V^{-1} + \frac{1}{\sigma^2}X^TX\right)^{-1}$$

Interval confidence (1-α)100% for β_j is :

$$P\left[\hat{\beta}_j - t_{\alpha/2(n-k)}s(\hat{\beta}_j) \le \beta_j \le \hat{\beta}_j + t_{\alpha/2(n-k)}s(\hat{\beta}_j)\right] = 1-\alpha$$

4. Variance for $\hat{y}$ is :

$\text{Var}(\hat{y}) = \text{Var}(X\hat{\beta}) = X\,\text{Var}(\hat{\beta})X^T$

5. Interval confidence (1-α)100% for y_i is :

$$p\left[\hat{y}_i - t_{\alpha/2(n-k)}s(\hat{y}_i) \le y_i \le \hat{y}_i + t_{\alpha/2(n-k)}s(\hat{y}_i)\right] = 1-\alpha$$

The detailed algorithms Bayes approach in the following regression model are:

a. Standardization of variables
b. Least squares regression method to obtain θ and σ^2
c. Conducting a resampling and regression coefficients assumed for each sample by the least square method to obtain V.
d. Compute $\hat{\beta}$ with formula :

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{V}^{-1} + \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{V}^{-1}\boldsymbol{\theta} + \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{y}\right)$$

e. Find $\text{Var}(\hat{\beta})$, $s(\hat{\beta})$, interval confidence (1-α) 100% bagi β_i, Var(ŷ), s(ŷ), and interval confidence (1-α) 100% for y_i.

C. Discrete Wavelet Transform

Suppose there is a vector data $x = (x_0, x_1, \ldots, x_{q-i})^T$ with $q = 2^M$, M>0 is an integer. Discrete Wavelet Transform (DWT) defined as follows:

$$d_{j,k}^{(\psi)} = \sum_{t=0}^{q-1} x_t \psi_{j,k}(t) \quad (6)$$

j = 0,1,2, …, (M-1) and k = 0,1,…, (2^j-1), thus obtained (q-1) coefficients and one coefficient $c_{0,0}$ same matrix dimensions variable X. With matrix notation, DWT in the equation (6) can be written

$$d = Bx \quad (7)$$

because B is orthogonal, hence can be written:

$$x = B^T d \quad (8)$$

where $d = (c_{0,0}, d_{0,0}, d_{1,0}, d_{1,1}, d_{2,0}, d_{2,1}, d_{2,2}, d_{2,3}, \ldots)^T$ and B^T is a matrix that the elements of his column is the value of $\phi(t)$ and $\Psi_{j,k}(t)$ for various $t \in [0, 1]$. Special properties of the matrix B^T is orthogonal, the first column of the same value, and number of other elements of each column equal to zero.

Vector data x can be connected with the function f in the interval [0, 1] and defined as :

$$f(t) = \sum_{k=0}^{q-1} x_k \left\{ \frac{k}{2^M} \le t < (k+1) < \frac{(k+1)}{2^M} \right. \quad (9)$$

This function is known with a stair and included in L^2 ([0,1]) so that the wavelet decomposition of *f*(*t*) is :

$$f(t) = c_{0,0}\,\phi(t) + \sum_{j=0}^{M-1}\sum_{k=0}^{2^j-1} d_{j,k}\,\psi_{j,k}(t). \quad (10)$$

For $\phi(t) = 1$ called the scale function for Haar wavelet. The equation (10) is called discrete wavelet transform, because the value of j is only taken on the positive integers. Numbers of j in the equation (10) is called a resolution level, and f(t) can be obtained accurately, if taken by all levels of resolution for the decomposition, i.e. the resolution level 0 up to (M-1). Coefficient $c_{0,0}$ called the coefficient of smoothing or part of a function approach, whereas $d_{j,k}$ called wavelet coefficients or also referred to the detail of a function.

For example, suppose there are four observations (q=4, M=2), x = $(x_0, x_1, x_2, x_3)^T$, hence can be written as follows:

$$f(t) = c_{0,0}\,\phi(t) + \sum_{j=0}^{1}\sum_{k=0}^{2^j-1} d_{j,k}\,\psi_{j,k}(t)\cdot$$

$$f(t) = c_{0,0}\phi(t) + d_{0,0}\Psi_{0,0}(t) + d_{1,0}\Psi_{1,0}(t) + d_{1,1}\Psi_{1,1}(t)$$

$$f(t) = \begin{cases} x_0 & , 0 \le t < \frac{1}{4} \\ x_1 & , \frac{1}{4} \le t < \frac{2}{4} \\ x_2 & , \frac{2}{4} \le t < \frac{3}{4} \\ x_3 & , \frac{3}{4} \le t < 1\,. \end{cases}$$

With DWT obtained:

$$x_0 = f(t \in [0, \tfrac{1}{4})) = c_{00}\,\phi(t \in [0, \tfrac{1}{4})) + d_{00}\,\psi_{00}(t \in [0, \tfrac{1}{4})) + d_{10}\,\psi_{10}(t \in [0, \tfrac{1}{4})) + d_{11}\,\psi_{11}(t \in [0, \tfrac{1}{4}))$$

$$x_0 = f(t \in [0, \tfrac{1}{4})) = c_{00}\,\phi(t \in [0, \tfrac{1}{4})) + d_{00}\,\psi_{00}(t \in [0, \tfrac{1}{4})) + d_{10}\,\psi_{10}(t \in [0, \tfrac{1}{4})) + d_{11}\,\psi_{11}(t \in [0, \tfrac{1}{4}))$$

$$x_1 = f(t \in [\tfrac{1}{4}, \tfrac{2}{4})) = c_{00}\,\phi(t \in [\tfrac{1}{4}, \tfrac{2}{4})) + d_{00}\,\psi_{00}(t \in [\tfrac{1}{4}, \tfrac{2}{4})) + d_{10}\,\psi_{10}(t \in [\tfrac{1}{4}, \tfrac{2}{4})) + d_{11}\,\psi_{11}(t \in [\tfrac{1}{4}, \tfrac{2}{4}))$$

$$x_2 = f(t \in [\tfrac{2}{4}, \tfrac{3}{4})) = c_{00}\,\phi(t \in [\tfrac{2}{4}, \tfrac{3}{4})) + d_{00}\,\psi_{00}(t \in [\tfrac{2}{4}, \tfrac{3}{4})) + d_{10}\,\psi_{10}(t \in [\tfrac{2}{4}, \tfrac{3}{4})) + d_{11}\,\psi_{11}(t \in [\tfrac{2}{4}, \tfrac{3}{4}))$$

$$x_3 = f(t \in [\tfrac{3}{4}, 1)) = c_{00}\,\phi(t \in [\tfrac{3}{4}, 1)) + d_{00}\,\psi_{00}(t \in [\tfrac{3}{4}, 1)) + d_{10}\,\psi_{10}(t \in [\tfrac{3}{4}, 1)) + d_{11}\,\psi_{11}(t \in [\tfrac{3}{4}, 1))\,.$$

In matrix notation can be written:

$$\begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} \phi(t \in [0, \frac{1}{4})) & \psi_{00}(t \in [0, \frac{1}{4})) & \psi_{10}(t \in [0, \frac{1}{4})) & \psi_{11}(t \in [0, \frac{1}{4})) \\ \phi(t \in [\frac{1}{4}, \frac{2}{4})) & \psi_{00}(t \in [\frac{1}{4}, \frac{2}{4})) & \psi_{10}(t \in [\frac{1}{4}, \frac{2}{4})) & \psi_{11}(t \in [\frac{1}{4}, \frac{2}{4})) \\ \phi(t \in [\frac{2}{4}, \frac{3}{4})) & \psi_{00}(t \in [\frac{2}{4}, \frac{3}{4})) & \psi_{10}(t \in [\frac{2}{4}, \frac{3}{4})) & \psi_{11}(t \in [\frac{2}{4}, \frac{3}{4})) \\ \phi(t \in [\frac{3}{4}, 1)) & \psi_{00}(t \in [\frac{3}{4}, 1)) & \psi_{10}(t \in [\frac{3}{4}, 1)) & \psi_{11}(t \in [\frac{3}{4}, 1)) \end{bmatrix} \begin{bmatrix} c_{00} \\ d_{00} \\ d_{10} \\ d_{11} \end{bmatrix}$$

or :

$x = B^T d$

B is called a wavelet transform matrix. In order to obtain matrix B^Twhich is ortogonal, then selected $\phi(t)$ and $\Psi_{j,k}(t)$ such that :

1. $\int_0^1 \phi(t)\,dt = 1$
2. $\int_0^1 \psi_{jk}(t)\,dt = 0$
3. $\int_0^1 \psi_{jk}^2(t)\,dt = 1$
4. $\int_0^1 \psi_{jk}(t)\;\psi_{lm}(t)\,dt = 0$ for j= l and k=m not occur simultaneously.
5. $\int_0^1 \phi(t)\;\psi_{jk}(t)\;dt = 0$.

Then matrix B^T obtained by multiplying all components $\phi(t)$ and $\Psi_{j,k}(t)$ with $\frac{1}{\sqrt{p}}$ for $t \in [0, 1]$.

Suppose for the Haar wavelet:

$$\phi(t) = \begin{cases} 1 & , \quad 0 \le t < 1 \\ 0 & , \text{otherwise} \end{cases}$$

$$\psi_{jk}(t) = \begin{cases} 2^{j/2}, & \frac{k}{2^j} \le t < \frac{2k+1}{2^{j+1}} \\ -2^{j/2}, & \frac{2k+1}{2^{j+1}} \le t < \frac{k+1}{2^j} \\ 0 \quad , & \text{etc .} \end{cases}$$

In the case *p* = 4 then for the Haar *wavelet* is :

$$\mathbf{B}^T = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{\sqrt{2}} & 0 \\ \frac{1}{2} & -\frac{1}{2} & 0 & \frac{1}{\sqrt{2}} \\ \frac{1}{2} & -\frac{1}{2} & 0 & -\frac{1}{\sqrt{2}} \end{bmatrix}.$$

That fulfill $B^TB = BB^T = I$.

Because of B is orthogonal, the wavelet coefficients can be calculated by

$d = Bx$.

If x has a large dimension, computation of wavelet coefficients matrix inefficient manner, so that the pyramid algorithm is used. From the simple example shown the number of elements of each column of the matrix B^T equal to zero except the first column associated with the scale function (father wavelet). This is because the elements other than the first column of the matrix B^T obtained from the mother wavelet function has the general properties $\int_{-\infty}^{\infty} \psi(t)\,dt = 0$.

If matrix $X_{(nxp)} = [x_1, x_2, \ldots, x_p]^T$ where $p = q = 2^M$ (M= positive integers) the DWT can be written:

$D_{(n\,x\,p)} = X_{(n\,x\,p)}\,B^T_{(n\,x\,p)}$

Because the dimensions of B very large (p>>), then choose a particular resolution levels such that the number of wavelet coefficients selected for p' with p'< (n-1) < p and the result will be obtained

$$D^*_{(n\,x\,p)} = X_{(n\,x\,p)}\,B^{*T}_{(n\,x\,p)} \qquad (11)$$

which reduces the observation of p vertices of each sample into p' wavelet coefficients selected point.

D. Bayes Regression with Pre-processing Discrete Wavelet Transformation

Bayes regression with pre-processing Discrete Wavelet Transformation (BR-DWT), is essentially Bayes regression between the response variable Y with independent variable D from DWT results. Matrix D from Wavelet Transformation not guaranteed independent. Sunaryo (2005) show that the new variable result from DWT still have a fairly high correlation, although its value is smaller than the correlation between variables of origin. Therefore, the regression model between the response variable (Y) with variable results from DWT cannot using OLS method because there are still multicollinearity problems.

Data analysis using the BB-WT divided into two phases that are to build a model calibration and validation model.

1. Phase 1

   Build a calibration model with the following steps :

   a. Search for a wavelet coefficient matrix

      i. Matrix $X_{(nxp)}$ is the results of discretizing of the infrared spectrum of gingerol is the independent variable, whereas the response variable $y_{(nx1)}$ is the concentration of active compounds of ginger powder from HPLC measurements. Data were divided into two groups: the first group is n_1 data to build models, and the second group is n_2 data for validation.

      ii. Get DWT matrix B with the involvement of 1024 points (M = 10), which is determined based on the mother wavelet. Wavelet matrix calculations in this study using the software Wavetresh (Nason, 1998).

      iii. Get the matrix of wavelet coefficients D with the formula $D_{(n1xp)} = X_{(n1xp)}B^{T}_{(pxp)}$ then select a specific resolution levels such that the number of wavelet coefficients selected for p'with p' < (n_1xp) < p. Subsequently obtained $D^{*}_{(n1xp)} = X_{(n1xp')}B^{*T}_{(pxp')}$ which reduces the observation of p vertices of each sample into p' point wavelet coefficients are selected. In software Wavetresh 3 (Nason, 1998 There are 10 levels of mother wavelet Daubechies (D-1 to D-10) so the matrix will be obtained 10 D^{*}.

   b. Making a regression model between the respon variables $\underline{y}_{(n_1x1)}$ and the matrix of independent variable $D^{*}_{(n1xp')}$ (D-1 to D-10) using Bayes regression method.

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{V}^{-1} + \frac{1}{\sigma^2}\mathbf{D}^{*T}\mathbf{D}^{*} \right)^{-1} \left(\mathbf{V}^{-1}\boldsymbol{\theta} + \frac{1}{\sigma^2}\mathbf{D}^{*T}\mathbf{y} \right)$$

$$\mathrm{Rag}(\hat{\beta}) = \mathrm{Rag}\left[\left(V^{-1} + \frac{1}{\sigma^2}D^{*T}D^{*} \right)^{-1} \left(V^{-1}\theta + \frac{1}{\sigma^2}D^{*T}y \right) \right]$$

2. Phase 2

   Validation of the model with the following steps :

   a. Perform predictive value of y in the group of validation data using the model produced in phase 1, with the following steps:

      i. Multiplication of the i-th observation vector $x_{(i)}^{T}$ have a measurement of (1xp) with the wavelet transform matrix B^{*} thus obtained vector $d_{(i)}^{*T}$ with formula $d_{(i)}^{*T} = x_{(i)}^{T}.\, B_{(pxp')}^{*T}$

      ii. Predict the value *y* by the formula:

$$\hat{y}_{(i)} = y_{predict(i)} = \underline{d}_{(i)}^{T}\hat{\beta} \;;\; i = 1,2,\ldots,n_2 .$$

   b. Further validating the model with RMSEP criterion, coefficient of determination of prediction results ($R_{predict}^{2}$).

III. METHOD

A. Data

The research is a step in continuous research about tatistics modelling. The aim of the long-term research is to solve the issues of classical assumptions in statistics modelling. The assumption studied here is the issue of multicollinearity, the number of observations which is much less than the number of independent variables, and the presence of outliers. In order to get a suitable model for solving these issues, a method combination is done.

The first step is to study the combination of Bayes Regression and Wavelet Transform (BR-WT) theoretically, which is expected to be an alternative for solving ill-conditioned and singularity problems.

The second step is to study the use of BR-WT method by using simulation data. And the last step is to use BR-WT method in the calibration modeling.

The calibration model in this research is calibration model of the level of gingerol compounds. The sample is gingerol from (a) medical plants farmers in Kulonprogo, Centre Java and Karanganyar, DIY, (b) the experiments in Biofarmaka experimental field IPB Bogor, and (c) Balitro, Bogor, Majalengka, and Sukabumi. The gingerol will be chemical-analyzed in three laboratories in Laboratorium Kimia Analitik Jurusan Kimia IPB, Laboratoriun Terpadu IPB, a Laboratorium Pusat Studi Biofarmaka LPPM-IPB.

The research sample is 20 gingerols. The composition of gingerol will be analysed through HPLC method, which will be used as the dependent variable. FTIR method gives the spectrum of infrared. The discretization process will gives the percentage of transmittance, which is observed in 1866 points for wavelength 4000 - 400 cm^{-1}. This value represents the level of gingerol, which is used as the independent variable.

DWT must have 2^{M} (M is positive integers), so that the observations are only taken 1024 points from the actual 1866 points, by considering the region of infrared spectrum.

The chosen 1024 points of observations will be transformed by using Discrete Wavelet Transform (DWT), by considering the possible resolution which gives wavelet coefficients size less than the sample size. Daubechies wavelet is used as the mother wavelet because it has been used in most applications and brought a better result [18].

B. Data

The analysis is using BR-WT method, which is divided into 2 steps, i.e. building the calibration model and the model validation. The criterions used to evaluate the model performance are R^2, RMSE, $R^2_{predict}$, RMSEP, and plot between the actual data and the prediction. The best model is a model which gives much R^2 and $R^2_{predict}$, less RMSE and RMSEP, and the fittest prediction to the actual data. BR-WT performance will be compared with CR-WT (Continuum Regression–Wavelet Transform) and PLSR-WT (Partial Least Square Regression–Wavelet Transform).

IV. RESULT

C. Multicolinearity Identification

According to VIF's value, shows that all of the independent variables are identified multicollinearity, cause of VIF's value is greater than 20.

D. Model Calibration Levels Gingerol

The measurement results obtained by FTIR data percent transmitans ginger powder to 20 samples at 1866 points as shown in Fig. 1. Furthermore, 1024 was chosen for each observation point. This is to meet the requirements of DWT that need 2^M (M=10) and the spectrum presented in Fig. 2.

Fig. 1 and Fig. 2 show that the spectrum is in tune (almost parallel). As a result, the process of discrediting the infrared spectrum that produces 1024 points and used to obtain independent variables X_1 to X_{1024} as a compiler matrix X, multicollinearity problems will arise because of the high correlation between X_1 and X_{1024}. In addition to the large number of points produced (1024) results in the independent variables is much greater than the number of observations. Therefore it needs to be done using data compression in order to obtain TWD wavelet transform coefficient matrix D^*.

Calibration model is built using 16 data. From the results of data analysis concluded that the best model for predicting levels of gingerol is built using 11 wavelet coefficients (for the mother wavelet Daubechies-10) at a resolution of 0, 1 and 3 and a coefficient function of scale discrete wavelet transforms. This is because in these conditions result whose behavior is relatively better than others, that can capture good measures of such models are relatively better R^2. The results proved the vectors that form the matrix D^* still has a high correlation, so if regressed between the response variable Y with D^* arise the multicollinearity problems. To solve these problems, Bayes Regression method is used. Summary results of the processing method with BR-WT MINITAB macro programs are presented in Table 1. Fig. 3 presents the scatter diagram of observed data ($y_{insample}$) with the prediction results using BR-WT, CR-WT, and PLS-WT. While Fig. 4 presents the scatter diagram data observations ($y_{outsample}$) with predicted results using BR-WT, CR-WT, and PLS-WT.

From Table 1 we can see that the results of BR-WT method obtained good value model that is better than PLS-WT method, and relatively as good as CR-WT methods. Furthermore, estimated by the method derived BR-WT R^2 = 97.9 %, $R_{predict}{}^2$ = 99.0 %, an RMSEP = 0,0254.

This is reinforced by Fig. 3 which shows that the dots prediction results using RB-TW obtained the points closer to the observational data points. So also in Fig. 4 shows that the points outcome prediction BR-WT method closer to the observational data points. So it can be said that the model for prediction of external data (data that is not involved in the modeling) is quite satisfactory. BR-WT method gives better results than the PLS-WT method. Meanwhile, when compared with CR-WT method is relatively equally well. Thus it can be said that the BR-WT method is one alternative that can be used to overcome the multicollinearity and singularity.

Thus we can conclude that the CR-DWT has very good potential for modeling calibration. Gingerol content of the calibration model using the CR-DWT approach the results of this study can be used to predict the content of the active compound (gingerol) in all kinds of ginger, if known FTIR spectrum. This is because all kinds of ginger have gingerol content of active substances that have the same chemical formula and the same spectral pattern. The difference is in the high and low percent of transmittance. Therefore, if the pattern resembles the pattern spectrum of FTIR spectra in Fig. 3, then this model can be used to predict the gingerol content of ginger powder.

Gingerol content of the calibration model can be used to predict the concentration of active compound gingerol in ginger with a high degree of accuracy, but cannot be used to predict the levels of active compounds in the rhizomes of plants of other drugs (e.g. turmeric). This is because the type of active compounds in each different medical plant means that its chemical formula is also different, resulting FTIR spectrum is also different with gingerol. Infrared spectrum of organic compounds has characteristic physical properties, which means the possibility of the two compounds has the same spectrum is very small (Nur and Adijuwana 1989). These differences affect the spectrum pattern of each active compound has a different calibration models. However, BR-WT method can be used to obtain calibration models for all levels of the active compounds of medicinal plant species.

V. CONCLUSION

The conclusion from the theoretical and empirical studies is the combination of Bayes Regression and preprocessing Discrete Wavelet Transform, which is called BR-WT, has a good effect in solving ill-conditioned and singularity. The application of CR-WT in the gingerol calibration brings R^2 = 97.9%, $R^2_{predict}$= 99%, and RMSEP 0.0254. Therefore, gingerol calibration model can be used predict the level of gingerol compounds, but it can not be used to predict other medical plants.

CR-TW method is the best method for solving ill conditioned and singularity issues, but it is not good for data having outliers. Therefore, a robust method is preferred and the combination with CR-TW, which will be called Robust Continuum Regression–Wavelet Transform (RCR-WT), is an interesting solution.

ACKNOWLEDGMENT

The research has been supported by DP2M DIKTI DEPDIKNAS through Hibah Penelitian Fundamental 2009.

REFERENCES

[1] J.O. Berger, 1985,"Statistical decision theory and bayesian analysis (2nd ed.)", *Springer-Verlag*, New York.

[2] G.E.P. Box, and G.C. Tiao, 1973, "Bayesian inference in statistical analysis", *Reading Mass.*, Addison-Wesley Publishing Co, London.

[3] Brown PJ, Fearn T, Vanucci M., 2001, "Bayesian wavelet regression on curves with application to a spectroscopic calibration problem",*J Amer Statist Assoc*, Vol. 96, pp. 398-408.

[4] Danutirto H., 2001,"Pengembangan fitofarmaka di Indonesia", *Lokakarya dan Pameran Pengembangan Agribisnis Berbasis Biofarmaka*, Kerjasama Departemen Pertanian dengan Institut Pertanian Bogor,Jakarta, Tgl. 13-16 November.

[5] McNulty SC, Mauze G., 1998, "Application of wavelet analysis determining glucose concentration of aqueous solution using nir spectroscopy",*Hewlett-Packard Comp.*

[6] Nason GP, Silverman BW., 1994,"The discrete wavelet transform in S.", *Journal Comp Graph. Stat.*, Vol. 3, pp. 163-191.

[7] Nason GP. 1998, "Wavethresh 3 software", Department of Mathematics, University of Bristol, UK.

(http://www.stats. bris.ac.uk/~wavethresh), 20 juni 2003.
[8] Notodiputro KA, 2003,Pendekatan statistika dalam kalibrasi", *Conference on Statistical and Mathematical Sciences of Islamic Society in South East Asia Region*, Bandung, 25-26 April 2003.
[9] Percival DB., 2005,*Wavelets: Data Analysis, Algorithms and Theory*, University Washington. http://www.ms.washing-ton.edu/~s530/18 April 2005.
[10] Setiawan, Notodiputro KA., 2003,Pendekatanbayes dengan prior normal dalam kalibrasi", *Prosiding Seminar Nasional Statistika VI*, Jurusan Statistika FMIPA ITS, Surabaya tanggal 11 Oktober 2003.
[11] Setiawan, Notodiputro KA., 2005a,"Regresi kontinum sebagai bentuk umum dari RKT, RKU, serta RKTP",*Prosiding Seminar Nasional Statistika VII*, Jurusan Statistika FMIPA ITS, Surabaya tanggal 26 Nopember 2005.
[12] Setiawan, Notodiputro KA., 2005b,"Regresi kontinum dengan prapemrosesan transformasi wavelet dalam model kalibrasi", *Prosiding Seminar Nasional MIPA*, FMIPA UNESA, Surabaya, 17 Desember.
[13] Setiawan, Notodiputro KA, 2006a,"Penerapan regresi kontinum pada model kalibrasi untuk menentukan kadar senyawa aktif pada impang jahe", *Prosiding Seminar Nasional Basic Science*. FMIPA-UNIBRAW, Malang, 25 Februari.
[14] Setiawan, Notodiputro KA, 2006b,"Sifat-sifat statistik dari regresi kontinum",*Makalah Seminar Nasional Matematika, Statistika dan Pendidikan Matematika*, Jurusan Matematika FMIPA UNPAD Bandung, 22 April.
[15] Setiawan, Notodiputro KA, 2007a,"Pengembangan model kalibrasi untuk menentukan kadar senyawa aktif pada rimpang temulawak dengan metode regresi kontinumwavelet", *J Tropika*.
[16] Setiawan, Notodiputro KA, 2007b,"Regresi kontinum dengan prapemrosesan transformasi wavelet diskret", *Jurnal Ilmu Dasar*.
[17] Shao X, Yadong Zhuang, 2004, Determining of chlorogenic acid in plant samples by using near-infrared spectrum with wavelet transform preprocessing, *Analytical Sciences* 20.
[18] Sunaryo S., 2005,*Model Kalibrasi dengan Transformasi Wavelet sebagai Metode Pra-pemrosesan [Disertation]*, Sekolah Pascasarjana, Institut Pertanian Bogor,Bogor.
[19] Yi-yu Cheng, Chen min-jun, 2000,"A new computing multivariate spectral analysis method based on wavelet transform,"*Journal of Zhejiang University Science*, Vol. 1, pp. 15-19.

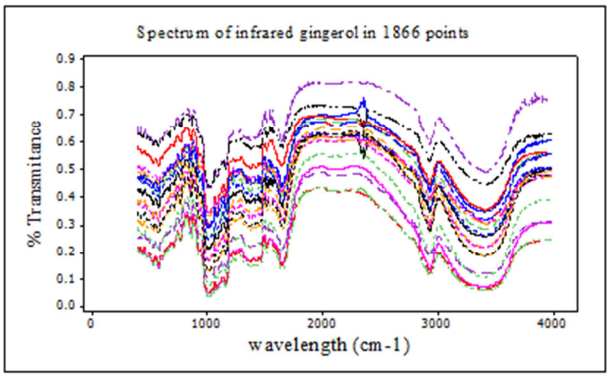

Fig.1. Infrared spectra of gingerol for 20 samples of powder Ginger on 1866 points

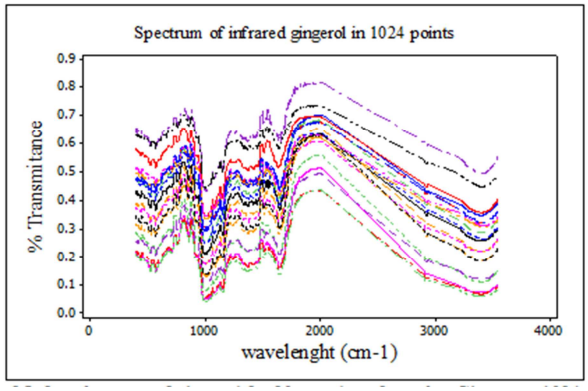

Fig. 2. Infrared spectra of gingerol for 20 samples of powder Ginger on 1024 points

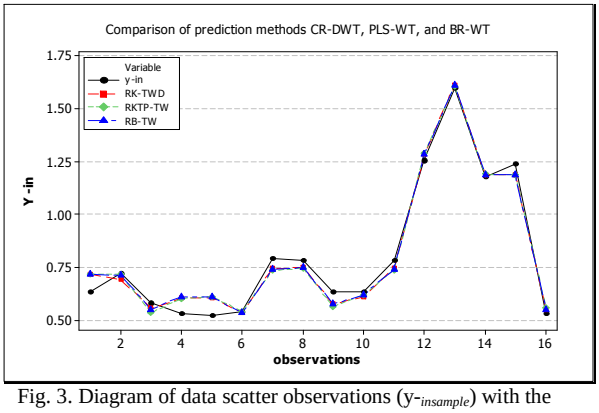

Fig. 3. Diagram of data scatter observations (y-$_{insample}$) with the prediction results using BR-WT, CR-WT, and PLS-WT.

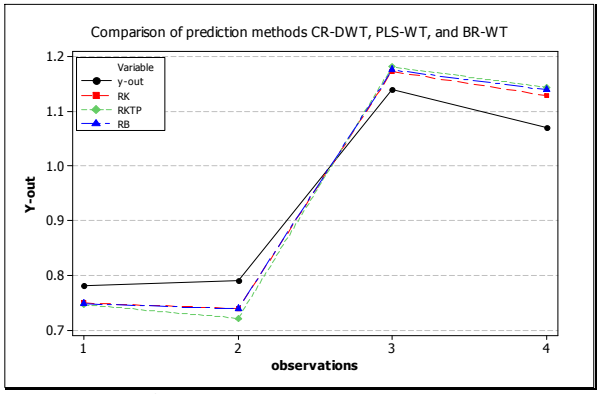

Fig. 4. Diagram of the data scatter observations (y-outsample) with Predicted results using BR-WT, CR-WT, and PLS-WT

TABLE 1.
CONCENTRATION GINGEROL OBSERVATION, ALLEGATION AND THE RESULTS OF PREDICTION RESULTS

Observation	Prediction Results		
$y_{insample}$	CR-WT	PLS-WT	BR-WT
0.63	0.7157	0.7139	0.7153
0.72	0.6914	0.7138	0.7125
0.58	0.5548	0.5328	0.55
0.53	0.6019	0.6006	0.6107
0.52	0.6049	0.6092	0.6075
0.54	0.532	0.5391	0.5325
0.79	0.7426	0.7338	0.7411
0.78	0.7483	0.7457	0.7477
0.63	0.5745	0.5636	0.5741
0.63	0.6084	0.6185	0.6176
0.78	0.742	0.7366	0.7404
1.26	1.2849	1.2887	1.2865
1.6	1.615	1.6092	1.6111
1.18	1.189	1.1903	1.1897
1.24	1.1879	1.1878	1.1878
0.53	0.5466	0.5564	0.5476
R^2	97.9	97.7	97.9
RMSE	0.0488	0.0517	0.0497
Observation	Prediction Results		
$y_{insample}$	CR-WT	PLS-WT	BR-WT
0.78	0.7485	0.7455	0.7472
0.79	0.7384	0.7205	0.7376
1.14	1.1725	1.1811	1.1755
1.07	1.1291	1.1438	1.1387
R^2_{predik}	99.2	98.8	99
RMSEP	0.02	0.0255	0.0254