A Combined Modeling of Generalized Linear Mixed Model and LASSO Techniques for Analizing Monthly Rainfall Data

A Muslim1, M Hayati1,2, B Sartono1, K A Notodiputro1*
1Department of Statistics, Bogor Agricultural University, Bogor, Indonesia.
2Department of Statistics, The university of Nahdlatul Ulama Lampung, Sukadana, Indonesia.
*email: khairlnotodiputro@gmail.com

Abstract. The rainfall pattern is always interesting to be investigated. This paper discusses the performance of three methods in modeling the rainfall data namely the LASSO (Least Absolute Shrinkage and Selection Operator) and GLMM (Generalized Linear Mixed Model) methods as well as a combination of GLMM and LASSO techniques. The rainfall data is usually collected on a regular basis, hence it is longitudinal data. The GLMM methods are usually employed to analyze longitudinal data, especially when number of explanatory variables is small. If the number of explanatory variables is large and if these variables are correlated then the GLMM estimation will be suffered by ill condition problems. These problems may be overcome by adding L1 penalty and start doing variable selection and shrinkage simultaneously. In this paper a combination of GLMM and LASSO techniques is evaluated by using monthly rainfall data, a high dimensional data, collected during 1981-2014 in Indramayu sub-district. The results showed that a combination of GLMM and LASSO methods is superior when compared with GLMM and LASSO methods separately. This claim is supported by evidence that MSE of the combined method is smaller than MSEs of the other two methods for various λ (lambda).

1. Introduction
The pattern of rainfall around the world today is strongly influenced by climate change. This occurs because the average air temperature is higher resulted in a higher evaporation rate, the moisture content is higher, and resulted in the hydrological cycle and faster [5]. Many quantitative studies have been conducted to provide evidence of climate change with regard to temperature and trends of hydro-meteorological data performed. Data analysis using various statistical methods.

The regression model describes a phenomenon (response variable) based on other phenomena (explanatory variables). The classical regression model was developed with the assumption that the response variable is normally distributed. However, in data modeling, response variables are not always normally distributed. For data that have exponential family distributions, developed Generalized Linear Model (GLM) to overcome the abnormalities of response variables. The main benefits of GLM are twofold. First, regression is no longer limited to normal data, but extended to exponential family distributions. This allows for appropriate modeling, for example, the number of frequencies or binaries. Second, the GLM model gives the additive effect of explanatory variables on the average transformation of the count, not the average count itself. GLM uses a link function that connects the average response variable count with the linear form of the predictor variable \( g(\mathbb{E}(Y)) = X' \beta \). \( g \) is a monotonic and deferred relationship function [1]. Since GLM is developed based on the
probability distribution of the response variable, the connecting function depends on this distribution. GLM is proposed to unify regression approaches for a wide range of such longitudinal, discrete and continuous data.

Response variables in repeatedly grouped data are usually correlated. Therefore, it is necessary to use GLM expansion by taking into account the correlation. This can be done by inserting a random effect on a linear predictor. GLM by inserting other than fixed effects but incorporating random effects is called Generalized Linear Mixed Model (GLMM / random effect model).

Several studies on rainfall have been conducted such as [2] using the Gamma and Weibull Distribution to predict rainy rainfall in India that provides good conformity. Previous study [3] predicted the precipitation in the Daqing Mountains by using the Multivariate Regression Model. Other research modeled rainfall based on the Statistical Downscaling with Gamma Distribution and Elastic Net Regularization which focused on GLM models coupled with Lasso penalty constraints ($L_1$) [4].

In this study, consideration of the addition of Lasso penalty constraints ($L_1$) to the fixed effect of the GLMM model and called GLMMLASSO. The GLMMLASSO application was applied to the Indramayu District rainfall modeling using monthly rainfall response from January 1981 to April 2014 and precipitation as a large-scale predictor obtained from interpolated combinations of surface and satellite observation data in the grid form of GPCP (Global Precipitation Climatology Project) 2.2. The use of GLMMLASSO will reduce the model's complexity. This method can be used in large dimensional data (High Dimensional Data) which has the potential to use the predictors/variables freely in large quantities [5].

2. LASSO (least absolute shrinkage and selection operator)

On the second discussion, it will cope with the method of Lasso

2.1 LASSO (Least Absolute Shrinkage and Selection Operator)

LASSO is one of the predictor shrinkage regression techniques. LASSO can be used to solve problems in microarray data. The LASSO method is a penalized regression technique that predicts the regression coefficient by minimizing the sum of the error squares with a $\lambda_1$ constraint. LASSO shrinks the coefficient (parameter $\beta$) which correlates to zero or close to zero, resulting in an estimate with a smaller variant and a more representative end model. LASSO is one of the techniques to produce sparse statistical model that is statistic model which has few nonzero parameter, so it is easier to estimate and interpret than "dense" model such as multiple linear regression model.

Estimation of parameters on LASSO are:

$$\hat{\beta} = \arg \max_{\beta} l(\beta)$$

with $\|\beta\|_1 \leq s, s \geq 0$, and $\|\cdot\|_1$ is norm of $(\xi_k)$.

The $\beta$-lasso estimation is also derived by solving the optimum problem of the following equation:

$$\hat{\beta} = \arg \max_{\beta} [l(\beta) - \lambda \|\beta\|_1]$$

with $\lambda \geq 0$ is tuning parameter.

2.2 Generalized linear mixed model (GLMM)

Let $y_{it}$ constitutes the observation $t$ in a cluster $i, i = 1, ..., n, t = 1, ..., T_i$ and in $y_i = (y_{ir_1}, ..., y_{iT_i})$, then $x_{it} = 1, x_{it1}, ..., x_{itp}$, is composed as covariate vector, which is related to fixed effect, and $z_{it} = 1, z_{it1}, ..., z_{itp}$ is the covariate vector which is related to random effect. It is therefore assumed that $y_{it}$ is conditionally independent with mean $\mu_{it} = E(y_{it}|b_i, x_{it}, z_{it})$ and variance $\text{var}(y_{it}|b_i) = \phi \nu(b_i)$ in which $\nu(\cdot)$ is a known variance function and $\phi$ is the scale parameter. The form of GLMM becomes:
\[ g(\mu_{it}) = x_{it}^T \beta + z_{it}^T b_i = \eta_{it}^{par} + \eta_{it}^{rand} \]  

\( g \) is a monotonic and continuously differentiable link function. \( \eta_{it}^{par} = x_{it}^T \beta \) is composed as the form of parametric linear within the parameter vector \( \beta^T = (\beta_0, \beta_1, ..., \beta_p) \) including intercept and \( \eta_{it}^{rand} = z_{it}^T b_i \) contains the certain cluster of random effects, wherein \( b_i \sim N(0, Q) \) and \( Q \) covariant matrix have measurement of \( q \times q \). So, the alternative form of GLMM becomes:

\[ \mu_{it} = h(\eta_{it}), \eta_{it} = \eta_{it}^{par} + \eta_{it}^{rand} \]  

\( h = g^{-1} \) is constituted as inverse of link function.

In GLMM, it is assumed that conditional density of \( y_{it} \) following explanatory variable is provided and random effect \( b_i \) is constituted as exponential family type

\[ f(y_{it}|x_{it}, b_i) = \exp \left\{ \frac{y_{it} \theta_{it} - \frac{1}{2} b^T Q (\theta_{it})^{-1} b}{\phi} + c(y_{it}, \phi) \right\} \]  

\( \theta_{it} = \theta(\mu_{it}) \) is constituted as natural parameter, \( k(\theta_{it}) \) is specific function depending on the type of exponential family, \( c(\cdot) \) is constituted as the log normalization constant and \( \phi \) is constituted as dispersion parameter.

One of methods to maximize GLMM may be implemented by the application of penalized quasi likelihood (PQL) (Breslow and Clayton (1993), Lin and Breslow (1996), and Breslow and Lin (1995)) [5]. Covariant matrix \( Q(\phi) \) on random effect \( b_i \) depends on an unknown parameter vector \( \phi \). On the basic concept of penelized, the joint likelihood function is defined by parameter vector of covariant structure \( \phi \) along with dispersion parameter \( \phi^T = (\phi, \theta^T) \) and parameter vector \( \theta^T = (\theta^T, b^T) \) with the function of likelihood log are described as follow:

\[ l(\delta, \gamma) = \sum_{i=1}^{n} \log(\int f(y_{i}|\delta, \gamma)p(b_i, \gamma)db_i) \]  

Where, \( p(b, \gamma) \) is the density of the random effects, Breslow and Clayton (1993) derive an approach

\[ l^{app}(\delta, \gamma) = \sum_{i=1}^{n} \log(f(y_{i}|\delta, \gamma)) - \frac{1}{2} b^T Q (\phi)^{-1} b \]  

within the form of penalty \( b^T Q (\phi)^{-1} b \) is due to the approximation based on Laplace method [5]

3. Generalized Linier Mixed Model LASSO (GLMMLASSO)

By employing the equation of the log-likelihood (7) is expanded to include the penalty term \( \lambda \sum_{i=1}^{p} |\beta_i| \) and becomes the form of penelized log likelihood by Breslow and Clayton (1993).

\[ l^{pen}(\beta, b, \gamma) = l^{pen}(\delta, \gamma) = l^{app}(\delta, \gamma) - \lambda \sum_{i=1}^{p} |\beta_i| \]  

With \( \delta \) obtained from by optimizing function.

\[ \delta = \arg \max_{\delta} l^{pen}(\delta, \hat{\gamma}) = \arg \max_{\delta} [l^{app}(\delta, \hat{\gamma}) - \lambda \sum_{i=1}^{p} |\beta_i|] \]  

The penalty used in the equation (8) and (9) is regarded as the approach of partially-penalized approach if parameter vectors used \( \delta^T = (\beta^T, b^T) \) are all calculated.
3.1 Algorithm Gradient Ascent—GLMMLASSO

On the equation (8), penalized log likelihood \( \ell_{\text{pen}}(\delta, \gamma) \) is hardly to be differentiable, the derivation can be defined as the following equation:

\[
l'_{\text{pen}}(\delta; v, \gamma) = \lim_{t \to 0} \frac{1}{t} \left( \ell_{\text{pen}}(\delta + tv, \gamma) - \ell_{\text{pen}}(\delta, \gamma) \right)
\]

(10)

Algorithm gradient ascent uses the series of Taylor approximations and estimates every movement of penalized log likelihood \( \ell_{\text{pen}} \) on equation (8) for each estimation \( \hat{\delta} \) with the second order of Taylor Estimation.

\[
\ell_{\text{pen}}(\hat{\delta} + ts_{\text{pen}}(\hat{\delta}, \gamma), \gamma)
\approx \ell_{\text{pen}}(\hat{\delta}, \gamma) + tl'_{\text{pen}}(\hat{\delta}; s_{\text{pen}}(\hat{\delta}, \gamma), \gamma) + 0.5t^2l''_{\text{pen}}(\hat{\delta}; s_{\text{pen}}(\hat{\delta}, \gamma))
\]

(11)

\( t > 0 \) and \( s_{\text{pen}}(\ldots) \) and \( l'_{\text{pen}}(\ldots) \) are taken into Goeman algorithm.

4. The modelling of monthly rainfall data in Indramayu sub-district in 1981-2014 by utilizing Generalized Linier Mixed Model LASSO (GLMMLASSO)

4.1 Data

The Data that used for this research is rainfall data from Indramayu sub-district as response variable and the corresponding precipitation as explanatory variable. The precipitation is value calculation of interpolation surface data combination and the satellite in the form of GPCP grid (Global Precipitation Climatology Project) 2.2 version and as abbreviated for GPCP, is constituted as higher scale covariate. The data of GPCP is acquired from the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, on the website http://www.esrl.noaa.gov/psd/). Covariate data is described in 7 × 7 grid domain (49 covariate) on the coordinate system 101.25°–116.25° East lon. and 13.75°–1.25° North Lat, with the grid as wide as 2.5° x 2.5°. On such a position, the site of Indramayu Sub-district lies below the central grid of the given region (Figure. 1).

To identify the influence of 49 precipitated covariate variables; from \( pr11 \) to \( pr77 \) against response variable of Indramayu’s monthly rainfall from January 1980 up to April 2014, it is employed with GLMMLASSO and then compared to Lasso and GLMM with the random effect in which the weather is used as the experimental estimation. The season traditionally consists of two sorts; rainy from September up to March and dry from April up to August.

This research mainly utilizes the model of GLMM, which is described in the formulation below:

\[
g(\mu_{ij}) = \eta_{ij} = x_{ij}^T \beta + z_{ij}^T b_i
\]

Where:

\( y_{ij} \sim \text{Gaussian} (\mu, \sigma^2) \)

\( b_i \sim \text{Gaussian} (0, Q) \)

\( X_i \) is precipitated variable \( i=1,..,49 \)

\( Y_{ij} \) is response variable, as the rainfall on the precipitation-\( i \), season -\( j \)

\( Z_{ij} \) is covariate vector that is correlated to the season records.

Link function employed here is the identity meanwhile GLMMLASSO is acquired by calculation of log likelihood on equation (6) with Gaussian response, then getting equation (7) and in the form of penalized likelihood on the equation. To employ estimated parameter, it is acquired by optimally attempting equation (9) to the direction of algorithm gradient ascent.

To compare a goodness of fit for the model by associating Root Mean Square Error (RMSE). Additionally, the Software for application is software R equipped with the package of GLMMLasso ().
GLMM for lme4 package equipped with glmer function() and Lasso for package glmnet equipped with cv.glmnet function(). To associate among three models of Lasso, GLMM, and GLMMLASSO, data is substituted by providing Box Cox transformation on the worth of Lambda which is used for transformation worth 0.22.

4.2 Result
The monthly rainfall data in Indramayu Sub-district cannot be designed by the application of GLMM, having been carried out by the application. It is due to the calculation is not convergent caused by its higher eigen value. Hence, this paper will attempt to implement the model of Lasso and GLMMLASSO.

The rainfall data consists of 49 covariate variables that is employed for modeling. In Figure 2 and 3, it likely seems that the model of LASSO is managed to select variables and simplifies the model, in
which a large number of shrinkage covariates shrinkage to zero point. There are numerous covariate variables positively active to non-equal value of zero or its surrounding as reflected in pr15, pr22, pr31, pr32, pr37 and pr77.

In the model of GLMM LASSO, the influence of additional constraint $L_1$ at different levels of the lambda, is demonstrated in Figure 2 and 3. On the same point as zero, the total of shrinkage variable is approximately zero or remains one to zero point. In addition, the process of shrinkage begins with lambda which equals to 16, and the shrinkage is relatively stable on the lambda 45. There are many more covariates to zero point. It deals with as the result of constraint accomplishment $L_1$ in making the model simpler, so that it decreases complexity of estimation.

| Lasso | Lambda 0 | Lambda 1 |
|-------|----------|----------|
| pr11  | pr12     | pr13     |
| pr21  | pr22     | pr23     |
| pr31  | pr32     | pr33     |
| pr41  | pr42     | pr43     |
| pr51  | pr52     | pr53     |
| pr61  | pr62     | pr63     |
| pr71  | pr72     | pr73     |...|

| Lame 2 | Lame 3 | Lame 4 |
|--------|--------|--------|
| pr11   | pr12   | pr13   |
| pr21   | pr22   | pr23   |
| pr31   | pr32   | pr33   |
| pr41   | pr42   | pr43   |
| pr51   | pr52   | pr53   |
| pr61   | pr62   | pr63   |
| pr71   | pr72   | pr73   |...|

| Lampda 10 | Lampda 11 | Lampda 12 |
|-----------|-----------|-----------|
| pr11      | pr12      | pr13      |
| pr21      | pr22      | pr23      |
| pr31      | pr32      | pr33      |
| pr41      | pr42      | pr43      |
| pr51      | pr52      | pr53      |
| pr61      | pr62      | pr63      |
| pr71      | pr72      | pr73      |...|

| Lampda 45 | Lampda 50 | Lampda 60 |
|-----------|-----------|-----------|
| pr11      | pr12      | pr13      |
| pr21      | pr22      | pr23      |
| pr31      | pr32      | pr33      |
| pr41      | pr42      | pr43      |
| pr51      | pr52      | pr53      |
| pr61      | pr62      | pr63      |
| pr71      | pr72      | pr73      |...|

**Figure 4.** The constitution of grid selection, blue-colored for non active variables and white-colored for active variables

In Figure 4, it is demonstrated that the higher value of lambda is calculated, the more possible regression parameter tends to shrinkage zero point. It is indicated by a range of white-colored charts that means regression coefficient tends to find zero point or near to zero. From the results of modeling using GLMM LASSO, for example in lambda equal to 50, the active regression parameters are pr11, pr12, pr16, pr21, pr22, pr23, pr27, pr31, pr32, pr33, pr34, pr37, pr41, pr42, pr44, pr51, pr52, pr53, pr54, pr61, pr62, pr63, pr64, pr67, pr71, pr72, pr73 dan pr77.
In Table 1, it presents the RMSE values of both models. The GLMMLASSO model is better than the Lasso model, as can be seen that the RMSE GLMMLASSO value is smaller than that of LASSO, because GLMMLASSO is capable to handle the regression model with a very large covariate. So, our study indicates that GLMMLASSO model is better than the LASSO and GLMM models especially on data with very large covariate variables.

5. Conclusion
Based on the previous explanation it can be concluded that:
- The GLMMLASSO model performs better than GLMM and Lasso, as seen from smaller RMSE values, the GLMMLASSO model is able to handle regression models with covariate variables from too many regression models.
- GLMMLASSO model with lambda approaching maximum lambda, able to make many regression coefficient of shrinkage to zero or near zero almost

References
[1] Dobson J A 2002 An Introduction Generalized Linear 2nd edition (London:A CRC Press Company)
[2] Stephenson DB Kumar KR Doblas-Reyes FJ Royer JF Chauvin E and Pezzulli S 1999 J.Month.Weath.Rev 127 1954-66
[3] Ranhao…..?
[4] Soleh A M 2015 Pemodelan Linier Sebaran Gamma dan Pareto Terampat Dengan Regularisasi L1 Pada Statistical Down Scaling Untuk Pendugaan Curah Hujan Bulanan. Disertation. Institute Pertanian Bogor. Indonesia.
[5] Gad A M and Kholy R S 2012 Inte J. of Prob and Stat 1(3): 67-73 DOI:10.5923/j.ijps.20120103.03, Cairo.
[6] Jaiswal R K Lohani A K and Tiwari H L 2015 Environ Proc 2 (4) 729–49.
[7] Jeon J J Jang H Sung Eun SC 2016 J of Hydro 538:831–41
[8] Krishnamoorth K 2006 Handbook of Statistical Distributions with Applications (New York :Chapman&Hall/CRC).
[7] Permatasari SM Djuraidah A and Soleh AM 2016 Proc of The 2nd Int Conf on App Stat ISSN : 2579-4361. Indonesia

[8]
[10] Tibshirani R Saunders M Rosset S Zhu J and Knight K 2005 J. Roy. Stat. Soc. B, 67, Part 1, pp 91-108.