Fused Lasso For Modeling Monthly Rainfall In Indramayu Sub Distric West Java Indonesia

F Novkaniza1,2, M Hayati1,3, B Sartono1, K Anwar Notodiputro1
1Department of Statistics, Bogor Agricultural University, Bogor, Indonesia
2Department of Mathematic, Indonesia University, Depok, Indonesia
3Department of Statistics, The University of Nahdlatul Ulama Lampung, Sukadana, Indonesia.
khairlnotodiputro@gmail.com

Abstract. Rainfall data is usually of high variation that needs advanced statistical techniques to properly model the data while taking into account its variation. Statistical Downscaling methods are well known in climate modeling especially to analyze the relationship between the large-scale climate data with small-scale climate data. In this setting, the large-scale covariates are correlated and require techniques for shrinking the regression coefficients. Fused LASSO (Least Absolute Shrinkage and Selection Operator) is a very popular shrinking method. The fused LASSO is a generalization of the LASSO penalty in a sense that new penalty parameters are added to enforces sparsity in both the coefficients and their successive differences. This addition of new parameters is desirable in applications especially if the covariates can be ordered in some meaningful way. In this paper the Fused LASSO is employed to model the average monthly rainfall data at Indramayu Sub-District collected from January 1981 until April 2014. The rainfall data is treated as a response variable whereas the precipitation data is considered as large-scale covariates. These covariates are obtained from a combination of interpolate surface observations and satellite data based on GPCP (Global Surface Climatology Project) version 2.2. The results showed that based on the AIC and BIC loss function the Fused LASSO method can selects 28.57 % significant grids.

1. Introduction
Climate change that occurs can affect rainfall patterns in the whole world. Climate change can be affected by temporal variability of precipitation. Therefore, it is important to obtain definition knowledge of the changes in magnitude and frequency of precipitation. However, climate change is currently affecting precipitation patterns throughout the world because higher average air temperatures result in higher evaporation rates, higher water vapor contents and consequently an accelerated hydrological cycle [2]. Much quantitative research about temperatures and hydro data trends of meteorology done to provide evidence of climate change and analyzed using various methods of statistics.

Indonesia as one country that is located in the tropical area has agricultural productivity, especially rice which in, depending on the elements of the climate change that are the rainfall. The rainfall is one of the elements of the climate change in tropical regions that have high variation so that it is difficult to be done sounding. Modeling rainfall through a statistical modeling is very important to improve the productivity of rice in Indonesia. One of the statistical modeling that can be used for modeling the rainfall of a specific region is the Statistical Downscaling (SDS) modeling.
The 4th International Seminar on Sciences  
IOP Publishing 
IOP Conf. Series: Earth and Environmental Science 187 (2018) 012046  
doi:10.1088/1755-1315/187/1/012046

SDS model is a statistical model who studied the relationship between the circumstances of some variables which represent "large scale (global)", with the situation of some variables which represent "small scale (local)". Generally, the data on a large scale is the outer data General Circulation Models (GCM) and small-scale data in the form of local climate such as rainfall data in a specific region observation post [3].

![Image: Statistical Downscaling](image.png)

Figure 1 : Statistical Downscaling

The monthly rainfall data is a time series data so that the statistical analysis is required to detect stationarity of time series. One of the methods that can be used to detect stationarity is based on a linear regression model. Because the linear regression model uses time-dependent covariates as predictors and projects the response variables via estimated regression coefficient, the model can combine many possible factors, including both climatological and geomorphological ones for analysis of nonstationarity [1]. Due to the flexibility of linear regression model, many statistical methods have been developed based on the basic concept of the regression model, one of which is Fused LASSO. The Fused LASSO, first proposed by [4] is a regression model with penalty function as the generalization of the LASSO penalty that can produce better estimator for regression coefficient under the condition of regularity.

Precipitation is the fall of water either in liquid form or frost from the atmosphere to the surface of the earth. In the meteorology, precipitation can be interpreted as all forms of products from condensing water vapor in the atmosphere and then will fall as rain. Many quantitative and non-stationary studies on precipitation have recently been conducted to provide clear evidences of climate change. [2] proposed the fused lasso penalty function to detect the change point of the annual maximum precipitation which can be generally fitted by using the Generalized Extreme Value (GEV) distribution in South Korea. In this paper, we focus on Fused LASSO model for rainfall modeling in Indramayu District, with precipitation as large-scale correlate covariates and average monthly rainfall as response variable. Covariates were taken from the output of GPCP version 2.2 in the $7 \times 7$ gridded domain above Indramayu District from January 1981 until April 2014. Missing data is not used in the model.

2. LASSO (Least Absolute Shrinkage and Selection Operator)

A linear regression model can be written as:

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i, \quad i = 1, \ldots, n$$  \hspace{1cm} (1)

with the errors $\epsilon_i$ having mean 0 and constant variance. We assume that the predictors or covariates are standardized to have mean 0 and unit variance and the response $y_i$ has mean 0. If number of covariates $p$ larger than $N$, then many methods have been proposed for regularized or penalized regression. The LASSO method is a regression technique that penalizes a least squares regression by the sum of the absolute values ($L_1 - norm$) of the regression coefficients. The form of this penalty encourages sparse solution (with many coefficients equal to 0). The LASSO shrinkages regression
coefficient (parameter $\beta$) that correlates to be zero or near zero, and variance of the estimator is smaller and produce a more representative model (Tibshirani, 1996). The LASSO is one of sparse statistical model that more easily to be estimated and interpreted, compared than the dense model as multiple linear regression models.

Suppose that $(X_1, Y_1), ..., (X_N, Y_N)$ the data is a random sample from the $N$ vectors of observation of the response variables with $p$ covariates. The LASSO finds the coefficients $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p)$ satisfying:

$$\min_{\beta} \| \mathbf{y} - \mathbf{X} \beta \|^2 \text{ subject to } \sum_j |\beta_j| \leq s$$

which is equivalent to:

$$\min_{\beta} \| \mathbf{y} - \mathbf{X} \beta \|^2 + \lambda_1 \sum_{j=1}^p |\beta_j|$$

(2)

The bound $s$ is called "tuning the parameter". The form (1) can be written as :

$$\min_{\beta} \left\{ \sum_i (y_i - x_i^T \beta)^2 + \lambda_2 \| \beta \|_1 \right\}$$

where $\lambda_1$ is a penalty or shrinkage parameter. The larger the value $\lambda_1$, then the greater shrinkage can be done. For sufficiently large $s$ and $p>N$, we obtain the least square solution or one of the many possible least squares solutions. But for smaller values of $s$, the solutions are sparse i.e. regression coefficients will be exactly 0 or near to 0, so LASSO can do the natural selection for covariates. This is interesting from data analysis viewpoints because of the LASSO can select the important covariates and discards the other covariates. The criterion and constraints condition in (1) are convex problem, and can be solved even for $p >> N$. The LASSO has a unique solution if there are no two covariates "perfectly collinear" and number of non-zero regression coefficient is at most min($N,p$).

LASSO does variabel selection but does not select groups of correlated variables. As the generalization of LASSO, Fused LASSO was proposed by [4] for handling correlated and ordered covariates in a linear regression model.

3. The Fused LASSO

Fused LASSO is a generalization of LASSO penalty. Suppose that there are $N$ observations and have the responses $y_1, y_2, ..., y_N$ and $X_{ij}$, $i = 1, 2, ..., N$, $j = 1, 2, ..., p$ covariates. The response variable $Y$ can be quantitative or equal 0 or 1 which represents 2 classes like "healthy" or "sick". In many problems covariates can be ordered, that is, $X_{ij}$ are realizations of covariate $X_j$ that can be ordered as $X_1, X_2, ..., X_p$, in some meaningful way. If ordering exists, neighbouring coefficient values $\beta_j$ are related and being piecewise constant over neighbouring values. It makes sense to encourage both block-sparsity and smoothness. This is the underlying principle of Fused LASSO. Fused LASSO has the same purpose by the method of the LASSO, to predict response $Y$ from covariate $X_1, X_2, ..., X_p$, especially on the problem of $p>>N$.

Fused LASSO defined by:

$$\hat{\beta} = \arg\min_{\beta} \left\{ \sum_i (y_i - \sum_j x_{ij} \beta_j)^2 \right\}$$

subject to

$$\sum_{j=1}^p |\beta_j| \leq s_1 \text{ and } \sum_{j=2}^p |\beta_j - \beta_{j-1}| \leq s_2$$

where $s_1 > 0, s_2 > 0$ denotes constraint or regularization parameters. Equation (3) can be written in the form of:

$$\sum_{i=1}^N (y_i - x_i^T \beta)^2 + \lambda_1 \sum_{j=1}^p |\beta_j| + \sum_{j=2}^p |\beta_j - \beta_{j-1}|$$

(4)

where:

$\lambda_1, \lambda_2 \geq 0$ penalty parameters.
\( \beta = (\beta_1, \beta_2, ..., \beta_p)^T \), \( x_i = (x_{i1}, x_{i2}, ..., x_{ip})^T \) and \( \lambda_1 \) and \( \lambda_2 \) are functions of the sample size \( N \).

The first constraint encourages "sparsity in coefficients", and the second constraint encourages "sparsity in their differences", i.e. similarity of neighbouring coefficient. The following is a diagram of thus schematic Fused LASSO for the case \( N > p = 2 \).

![Diagram of Fused LASSO](image)

In Figure 2 illustrates the idea of Fused LASSO, search for contours of the function of the losses for the number of square error (\( \bigcirc \)) that meet the (\( \bigotimes \)) and \( \sum_j |\beta_j| = s_1 \sum_j |\beta_j - \beta_{j-1}| = s_2 \) (\( \bigoplus \)). The term *fused* is borrowed from the terminology "fusion", introduced by Land and Friedman (1996), who proposed the use of a penalty of the form \( \sum_j \|\beta_j - \beta_{j-1}\|^\alpha \leq s_2 \) for the value of \( \alpha = 0.1, 2 \). But they did not consider the use of two penalties on both \( \sum_j |\beta_j| \) and \( \sum_j |\beta_j - \beta_{j-1}| \) as on Fused LASSO.

To obtain the regression coefficient of Fused LASSO model i.e. \( \hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_p) \) that meet the constraint form on (4), first we select candidates values of penalty parameters \( \lambda_1 \) and \( \lambda_2 \). Based on the value of this candidates, we search for optimal values that minimize selected loss function. There are 4 loss functions can be used; AIC, BIC, Squared and GCV (Generalized Cross Validation). In the searching process for optimal lambda, loss function is used to evaluate the performance of the model for every candidate pairs of lambda given. From the optimal lambda, we obtained regression coefficient as solution of Fused LASSO model.

4. The modelling of monthly rainfall data in Indramayu sub-district in 1981-2014 by utilizing Fused LASSO

In this research, we use Indramayu Sub-District Monthly Rainfall data, We will describe the data set used for the application of fused LASSO

4.1 Data Source.
The productivity of agriculture especially the rice is very depending on the elements of the climate change that is the rainfall. The rainfall is one of the elements of the climate change in tropical regions that have high variation so that it is difficult to estimate. But estimation of the rainfall is important to improve the productivity of rice in Indonesia. West Java Province is one of the largest rice producer in Indonesia. Rice production in West Java in the year 2014, reached ± 11 million tons of unhusked rice dry grind, a decline of 3.7% compared to 2013.
Indramayu District is one of the districts in West Java which has an area of rice harvest 225983 Ha. This makes the Indramayu District with extensive as the broadest rice harvest in West Java. This district border by Laut Jawa in the north Kabupaten Cirebon to south-east Kabupaten Majalengka and Kabupaten Sumedang in the south and Kabupaten Subang in the west. Indramayu District consists of 33 districts, which is divided into a number of 315 desa and kelurahan. As the central government is located in the sub-district Indramayu. Precipitating covariates as the results of interpolate combination of surface observations and satellite data in grid form from the GPCP (Global Surface Climatology Project) version 2.2 and abbreviated as GPCP, used as large-scale covariate. We obtain GPCP data from the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, through its website at http://www.esrl.noaa.gov/psd/). Precipitating data were taken in 7×7 gridded domain (49 covariates) on the coordinates system 101.25° - 116.25° BT and 13.75° - 13.75° LU with the width of the grid 2.5° x 2.5°. In this position, Indramayu District is located at the bottom of the grid among the districts chosen (Figure 2).

Figure 3: Domain of Covariates

The Fused LASSO method used to analyze the relationship between 49 precipitating covariate variable (p11 until p77) and response variable, the average monthly rainfall in Indramayu District from January 1981 until April 2014. Because in the Fused LASSO covariates must be ordered, first we compute value of the absolute correlation between each precipitating covariate with monthly rainfall. Covariate that have the largest absolute correlation values until the smallest, are sorted and they are denoted by $X_1, ..., X_{49}$, and then we use the Fused Lasso procedure.

4.2 Result
In this section, we will describe the the result of modelling Fused LASSO using Indramayu Sub-District Monthly Rainfall data, with some step in modelling.

4.2.1. Fused LASSO Model.
Because the average monthly rainfall data is time series data, we have to check stationary condition for the data and it can be done by formed an autocorrelation function plot as follows:
From the Figure 4, there is a seasonal pattern and it indicates non-stationary. The seasonal pattern exists because series is influenced by seasonal factor i.e. month. One way to make a time series stationary is by compute the differences between consecutive observations. This is known as differencing. Differencing can help stabilize the mean of a time series by removing changes in the level of a time series, and so eliminating seasonality. Because the seasonality factor is month, we use 12-differencing and as well as looking at the time plot of the data, from the autocorrelation function (ACF) plot, we can get stationary data in Figure 5.

After we get stationary, we use the 12-differencing average monthly rainfall data as response in Fused LASSO model. Furthermore, we calculated the value of the absolute correlation between each precipitating covariate (p11 – p77) with response and sorted from the largest value to the smallest to get new covariate order and denotes the new precipitating covariates as $X_1, ..., X_{49}$. 
Table 1: Covariate Precipitation ordered by Absolute Correlation Value

| Grid | 1  | 2  | 3  | 4  | 5  | 6  | 7  |
|------|----|----|----|----|----|----|----|
| 1    | p11| p12| p13| p14| p15| p16| p17|
| 2    | p21| p22| p23| p24| p25| p26| p27|
| 3    | p31| p32| p33| p34| p35| p36| p37|
| 4    | p41| p42| p43| p44| p45| p46| p47|
| 5    | p51| p52| p53| p54| p55| p56| p57|
| 6    | p61| p62| p63| p64| p65| p66| p67|
| 7    | p71| p72| p73| p74| p75| p76| p77|

Based on the Figure 6, we can see that the Fused LASSO provides smoothness and block sparsity for covariates. Most of the estimated regression coefficients shrinkage to 0 and neighbouring covariates in the same block have similar estimated regression coefficient. This is in accordance with the purpose of the Fused LASSO method. From the following figure 7, we can see the actual 12 differencing average monthly rainfall plots versus the fitted valued for each time observation.

Table 2. Fused LASSO model

| Fungsi Loss | Squared Loss | GCV Loss |
|-------------|--------------|----------|
|             | nfold=5      | nfold=10 | nfold=5 | nfold=10 |
| Optimal Lambda | $\lambda_1 = 50, \lambda_2 = 50$ | $\lambda_1 = 50, \lambda_2 = 10$ | $\lambda_1 = 40, \lambda_2 = 10$ |
| Residual Deviance | 12940000 | 12710000 | 12630000 | 12630000 |
| AIC | 12940000 | 12710000 | 12630000 | 12630000 |
| RMSE | 180.336954 | 178.6772116 | 178.1715303 | 178.1715303 |

| Loss Function | AIC Loss | BIC Loss |
|---------------|----------|----------|
|               | nfold=5 | nfold=10 | nfold=5 | nfold=10 |
| Optimal Lambda | $\lambda_1 = 50, \lambda_2 = 50$ | $\lambda_1 = 50, \lambda_2 = 50$ | $\lambda_1 = 50, \lambda_2 = 50$ | $\lambda_1 = 50, \lambda_2 = 50$ |
| Null Deviance | 10190000 | 10190000 | 10190000 | 10190000 |
| Residual Deviance | 9983000 | 9983000 | 9983000 | 9983000 |
| AIC | 9983000 | 9983000 | 9983000 | 9983000 |
| RMSE | 160.4937 | 160.4939 | 160.4939 | 160.4939 |
To know the structure of grid selection based on active or non active covariate from estimated regression coefficient in the Fused LASSO model, we made grid map in Figure 8:

**Figure 6.** Plot LASSO Coefficient VS Covariates

**Figure 7.** Monthly Rainfall Vs Fitted Value

**Figure 8 :** The Structure of Grid Selection Descriptions of active covariate (Brown) and non-active covariate (white)
Table 3. Percentage of Active Coefficient Fused LASSO

| Type of Cov | Loss Function |
|-------------|---------------|
|             | Squared GCV AIC BIC |
|             | nfold=5 nfold=10 nfold=5 nfold=10 nfold=5 nfold=10 nfold=5 nfold=10 |
| Active (%)  | 28.5714 28.5714 57.1429 57.1429 28.5714 28.5714 28.5714 28.5714 |

4.2.2. Fused Model Residual Analysis LASSO.
To analyze the Fused LASSO model, we have residual analysis based on the criteria in table and figure as follows:

Table 4. The Value of the Determination Coefficient Fused LASSO

| Loss Function |
|---------------|
| Criteria Squared Loss GCV Losses AIC Losses BIC Loss |
|               | nfold=5 nfold=10 nfold=5 nfold=10 nfold=5 nfold=10 nfold=5 nfold=10 |
| R² (%)        | 2.15857 2.15857 3.7699 3.7699 2.158567 2.1584 2.15838 2.15842 |

Based on the value of the determination coefficient R², the Fused LASSO model with GCV loss function has the highest value 3.7699%. It means all of precipitating covariates in the model can explain about 3.7% variation of the average monthly rainfall in Indramayu Sub-District. We also checked constant variance error assumptions are met or not, by plots between fitted value versus residual of Fused LASSO model (Figure 9).

Figure 9. Y Fitted VS a Residual

Based on the Figure 9 we can see that there is no specific trend or pattern so that the assumption of constant error variance can be met by the Fused LASSO model. In addition, because error also assumed independent and have normal distribution, we checked whether the assumption is fulfilled, by formed
the residual plots versus monthly index and normal Q-Q plot of residual Fused LASSO model on Figure of 10 and 11 as following:

Figure 10. Residual Plot VS Month Index

Based on the Figure 10, residual spread around 0 and already have a random pattern even though there are very high and small residual values. It indicated independent error assumption could be met by the residual. In the Figure 11, we can see seen that assumption error has normal distribution not met, especially for the residual values below and above normal line.

Figure 11. Normal Plots Q-Q
We conducted non-parametric test of normality by Kolmogorov Smirnov test with a null hypothesis is residual has normal distribution. The results of the normality test are presented in Table 5 as follows:

| Table 5. Normality Test of Fused LASSO Residual |
|-----------------------------------------------|
|                                | Squared Loss | GCV Loss |
|                                | nfold=5 | nfold=10 | nfold=5 | nfold=10 |
| Mean                          | -0.008406 | -0.008406 | -0.007028 | -0.007096 |
| StDev                        | 160.7    | 160.7     | 159.4   | 159.4   |
| N                            | 387      | 387       | 387     | 387     |
| KS                           | 0.119    | 0.119     | 0.125   | 0.125   |
| P-Value                      | <0.010   | <0.010    | <0.010  | <0.010  |

Using $\alpha = 0.05$, because P-Value < $\alpha$, so that it can be concluded that the assumption error has normal distribution is not met for Fused LASSO model. Based on the residual analysis, we know that the Fused LASSO model using AIC and BIC loss function are better than GCV loss function because it has AIC and smaller RMSE, although GCV has the highest coefficient of determination. The assumption of error has normal distribution and constant variance could not be met by the residual, but the independent error assumption can be fulfilled.

5. Conclusion
Fused LASSO is designed for the case in linear regression model when neighbouring coefficients are related or ordered covariates. Fused Lasso provides smoothness and block sparsity for correlated neighbouring covariates. In the modeling of average monthly rainfall in Indramayu Sub-District from January 1981 until April 2014 using Fused LASSO, we get 71.4286 % non active grid (sparse solution) based on Squared, AIC and BIC loss functions. But using GCV loss function, we get 42.85 % non active grid and has the highest coefficient of determination 3.77%. Because we get the stationary response data through differencing, we plan to apply Fused LASSO for another non-stationary time series data and another way for ordering covariates

6. References
[1] Jaiswal R K Lohani A K and Tiwari H L 2015 Environ Proc 2 (4) 729–49
[2] Jeon J Jang H Sung Eun SC 2016 J of Hydro 538:831–41.
[3] Soleh A M 2015 Pemodelan Linier Sebaran Gamma dan Pareto Terampat Dengan Regularisasi LI Pada Statistical Down Scaling Untuk Pendugaan Curah Hujan Bulanan. Disertation. Institute Pertanian Bogor. Indonesia
[4] Tibshirani R Saunders M Rosset S Zhu J and Knight K 2005 J. Roy. Stat. Soc. B, 67, Part 1, pp 91-108.