

ISBN 978-979-19256-0-0



# PROCEEDING

**THE 3RD INTERNATIONAL CONFERENCE  
ON MATHEMATICS AND STATISTICS**

**BOGOR, 5 - 6 AUGUST 2008**

*Mathematics and Statistics: bridge for academia, business,  
and government in the entrepreneurial era*



Department of Statistics  
University of Mathematics  
Bogor, Indonesia



RESEARCH  
MATHEMATICS AND  
STATISTICS  
IN BOGOR, 2008



Department of Mathematics  
Faculty of Science, University  
of Indonesia

Inventory Model With Gamma Distribution	Hadi Sumadibrata, Ismail Bin Mohd	642
Accuracy Analysis Of Naive Bayesian Anti-Spam Filter	Ruslam, Armin Lawi, And Sri Astuti Thamrin	649
A New Method For Generating Fuzzy Rules From Training Data And Its Application In Financial Problems	Agus Maman Abadi, Subanar, Widodo, Samsubar Saleh	655
The Application Of Laws Of Large Numbers In Convergence Concept In Probability And Distribution	Georgina M. Tinungki	662
An Empirical Bayes Approach for Binary Response Data in Small Area Estimation	Dian Handayani, Noor Akma Ibrahim, Khairil A. Notodiputro, MOhd. Bakri Adam	669
Statistical Models For Small Area Estimation	Khairil A Notodiputro, Anang Kurnia, and Kusman Sadik	677
Maximum Likelihood Estimation For The Non-Separable Spatial Unilateral Autoregressive Model	Norhashidah Awang, Mahendran Shitan	685
Small Area Estimation Using Natural Exponential Families With Quadratic Variance Function (Nef-Qvf) For Binary Data	Kismiantini	691
Using An Extended And Ensemble Kalman Filter Algorithm For The Training Of Feedforward Neural Network In Time Series Forecasting	Zaqiatud Darojah, M. Isa Irawan, And Erna Apriliani	696
Estimation Of Outstanding Claims Liability And Sensitivity Analysis: Probabilistic Trend Family (PTF) Model	Arif Herlambang, Dumaria R Tampubolon	704
Expected Value Of Shot Noise Processes	Suyono	711
Modelling Malaysian Wind Speed Data Via Two Paramaters Weibull	Nur Arina Basilah Kamisan, Yong Zulina Zubairi, Abdul Ghapor Hussin, Mohd. Sahar Yahya	718
Application Of Latin Hypercube Sampling And Monte Carlo Simulation Methods: Case Study The Reliability Of Stress Intensity Factor And Energy Release Rate Of Indonesian Hardwoods	<u>Yosafat Aji Pranata</u> And Pricillia Sofyan Tanuwijaya	726
The Development Of Markov Chain Monte Carlo (Mcmc) Algorithm For Autologistic Regression Parameters Estimation	Suci Astutik, Rahma Fitriani, Umu Sa'adah, And Agustin Iskandar	734
A Note About Dh-Fever Estimation With ARIMAX Models	Elly Ana, Dwi Atmono Agus W	741
Evaluation Of Additive-Innovational Outlier Identification Procedure For Some Bilinear Models	Ismail, M.I., Mohamed, I.B., Yahya, M.S.	745

Model By Spectral Regression Methods	Iriawan, Suhartono	
Application Of Cluster Analysis To Developing Core Collection In Plant Genetic Resources	Sutoro	875
Small Area Estimation With Time And Area Effects Using A Dynamic Linear Model	Kusman Sadik And Khairil Anwar Notodiputro	880
Statistical Analysis Of Wind Direction Data	Ahmad Mahir Razali, Arfah Ahmad, Azami Zaharim And Kamaruzzaman Sopian	886
Generalized Additive Mixed Models in Small Area Estimation	Anang Kurnia, Khairil A. Notodiputro, Asep Saefuddin, I Wayan Mangku	891
Kernel Principal Component Analysis In Data Visualization	Ismail Djakaria, Suryo Guritno, Sri Haryatmi	898
GARCH Models And The Simulations	Nelson Nainggolan, Budi Nurani Ruchjana And Sutawanir Darwis	906
Rainfall Prediction Using Bayesian Network	Hera Faizal Rachmat, Aji Hamim Wigena, and Erfiani	911
Identifying Item Bias Using The Simple Volume Indices And Multidimensional Item Response Theory Likelihood Ratio (Irt-Lr) Test	Heri Retnawati	916
Ordinary Kriging And Inverse Distance Weighting For Mapping Soil Phosphorus In Paddy Field	Mohammad Masjkur, Muhammad Nuraidi and Chichi Noviant	924
K-Means Clustering Visualization On Agriculture Potential Data For Villages In Bogor Using Mapserv	Imas S. Sitanggang, Henri Harianja, and Lailan Syaufina	932
Some Methods To Estimate The Number Of Components In A Mixture	M. A. Satyawan, A. H. Wigena, Erfiani	941
A Probabilistic Model For Finding A Repeat Triplet Region In DNA Sequence	Tigor Nauli	947
Application Of Spherical Harmonics In Determination Of Tec Using Gps Observable	Mardina Abdullah, Siti Aminah Bahari, Baharudin Yatim, Azami Zaharim, Ahmad Mahir Razali	954
Testing Structure Correlation Of Global Market By Statistic Vvsv	Erna Tri Herdiani, and Maman A. Djauhari	961
Exploring the MAUP from a spatial perspective	Gandhi Pawitan	967
Estimation of RCA(1) Model using EF: A new procedure and its robustness	1Norli Anida Abdullah, 2Ibrahim Mohamed, 3Shelton Peiris	996
Second Order Linear Elliptic Operators In The Unit Square	Abdul Rouf Alghofari	1008

## POSTER

Study Of Fractional Factorial Split-Plot Experiment	Sri Winarni, Budi Susetyo, and Bagus Sartono	1012
Improving Model Performance For Predicting Poverty Village Category Using Neighborhood Information In Bogor	Bagus Sartono, Utami Dyah S, and Zulhelmi Thaib	1019
Ammi Models On Count Data: Log-Bilinear Models	Alfian Futuhul Hadi H. Ahmad Ansori Mattjik I Made Sumertajaya Halimatus Sa'diyah	1026
Prediction Of Oil Production Using Non Linear Regression By Sdpro Software (Special Program Package)*)	Budi Nurani R , and Kartlos J. Kachiashvili	1038
An Implementation Of Spatial Data Mining Using Spatial Autoregressive (Sar) Model For Education Quality Mapping At West Java*)	Atje Setiawan A. , Retantyo Wardoyo , Sri Hartati , and Agus Harjoko	1045
Validation Of Training Model For Robust Tests Of Spread	Teh Sin Yin, and Abdul Rahman Othman	1056
Spectral Approach For Time Series Analysis	Kusman Sadik	1063
The ACE Algorithm for Optimal Transformations in Multiple Regression	Kusman Sadik	1066
The Relation Between The Students' Interaction And The Construction Of Mathematical Knowledge	Rini Setianingsih	1069
Application of Auto Logistic Regression Spatial Model using Variogram Based Weighting Matrix to Predict Poverty Village Category	Utami Dyah Syafitri, Bagus Sartono, Vinda Pratama	1075

# The ACE Algorithm for Optimal Transformations in Multiple Regression

Kusman Sadik

Department of Statistics, Institut Pertanian Bogor Jl. Meranti, Wing 22 Level 4,  
 Kampus IPB Darmaga, Bogor 16680 - Indonesia  
 e-mail : [kusmans@ipb.ac.id](mailto:kusmans@ipb.ac.id)

**Key Words:** Alternating conditional expectations, non-restrictive transformations, parametric transformations, multivariate analysis, generalized additive models.

## 1. Introduction

We employed the Alternating Conditional Expectations (ACE) technique (Breiman & Friedman, 1985) to relax the assumption of model linearity. By generating non-restrictive transformations for both the dependent and independent variables, ACE develops regression models that can provide much better model fits compared to models produced by standard linear techniques such as Ordinary Least Squares. The ACE transformations can reveal new information and insights on the relationship between the independent and dependent variables.

The objective of fully exploring and explaining the effect of covariates on a response variable in regression analysis is facilitated by properly transforming the independent variables. A number of parametric transformations for continuous variables in regression analysis have been suggested (Box and Tidwell, 1962; Kruskal, 1965; Mosteller and Tukey, 1977; Cook and Weisberg, 1982; Carroll and Ruppert, 1988; Royston, 2000).

In this paper, we introduce the ACE algorithm for estimating optimal transformations for both response and independent variables in regression and correlation analysis, and illustrate through two examples that usefulness of ACE guided transformation in multivariate analysis. The power of the ACE approach lies in its ability to recover the functional forms of variables and to uncover complicated relationships.

## 2. The Alternating Conditional Expectation Methods

Non-parametric regression techniques are based on successive refinements by attempting to define the regression surface in an iterative fashion while remaining 'data-driven' as opposed to 'model-driven'. These non-parametric regression methods can be broadly classified into those which do not transform the response variable (such as Generalized Additive Models) and those which do ACE.

The ACE module produces an output of graphical transformations for the dependent and independent variables. ACE will also indicate the adjusted and imputed *p-value* of the model based on these graphical transformations (all *p-values* are really only the computational counterparts of the *p-values* in a standard regression model, but they are useful for between-model comparisons as well as for variables selection). An ACE regression model has the general form:

$$\theta(Y) = \alpha + \sum_{i=1}^p \phi_i(X_i) + \varepsilon$$

where  $\theta$  is a function of the response variable,  $Y$ , and  $\phi_i$  are functions of the predictors  $X_i$ ,  $i = 1, \dots, p$ . For a given dataset consisting of a response variable  $Y$  and predictor variables  $X_1, \dots, X_p$ , the ACE algorithm starts out by defining arbitrary measurable mean-zero transformations  $\theta(Y)$ ,  $\phi_1(X_1), \dots, \phi_p(X_p)$ . The error variance  $\varepsilon$  that is not explained by a regression of the transformed variable on the sum of

$$\varepsilon^2(\theta, \phi_1, \dots, \phi_p) = E \left\{ \left[ \theta(Y) - \sum_{i=1}^p \phi_i(X_i) \right]^2 \right\}$$

The ACE algorithm approaches this problem by minimizing the  $\varepsilon^2$ . If only there is a predictor variable  $X_j$ , so it needs minimizing  $\varepsilon^2(\theta, \phi_j)$ . For fixed  $\theta$ , the minimizing is  $\phi_j(X_j) = E\{\theta(Y) | X_j\}$ . For fixed  $\phi_j$  minimizing is  $\theta(Y) = E\{\theta(X) | Y\}$ . This is the key idea in the ACE algorithm, it begins with some starting functions and alternates between these two steps until convergence (Hastie dan Tibshirani, 1990).

transformed independent variables is (under the constraint,  $\int \phi_j(X_j) dX_j = 0$ )

### 3. Data Simulated

In this study, we apply the ACE technique to a synthetic example - case for which we know the correct answers - to demonstrate how the ACE algorithm can be used to identify the functional relationship between dependent and independent variables. This synthetic example is a multivariate case with four predictor and 250 observations generated from the following model

$$Y = 5 + \exp(X_1) + |X_2| + X_3^2 + X_4^3 + \epsilon$$

where  $X_1 \sim \text{uniform}(-5,5)$ ,  $X_2 \sim \text{uniform}(-20,20)$ ,  $X_3 \sim \text{uniform}(-5,5)$ , and  $X_4 \sim \text{uniform}(-4,4)$  and  $\epsilon$  is independently drawn from a standard normal distribution  $N(0, 1)$ . Note that the plots in Figure 1 do not reveal any obvious functional forms for either the dependent variable or predictors, even though  $X_1, X_2, X_3,$  and  $X_4$  are statistically independent. Under such circumstances, direct application of linear regression is not appropriate.

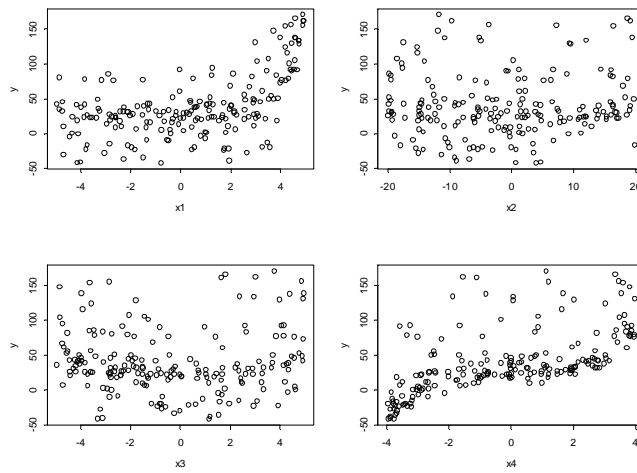


Figure 1. Scatterplots of simulated dataset (each predictor v.s. dependent variable)

To check if the ACE algorithm can recover these functions, we applied the algorithm to this simulated data set and the results are plotted in Figure 2. Clearly, ACE is able to recover the corresponding functions. A regression of the transformed dependent variable on all the transformed covariates results in all parameter coefficients of the independent variables being positive and close to:

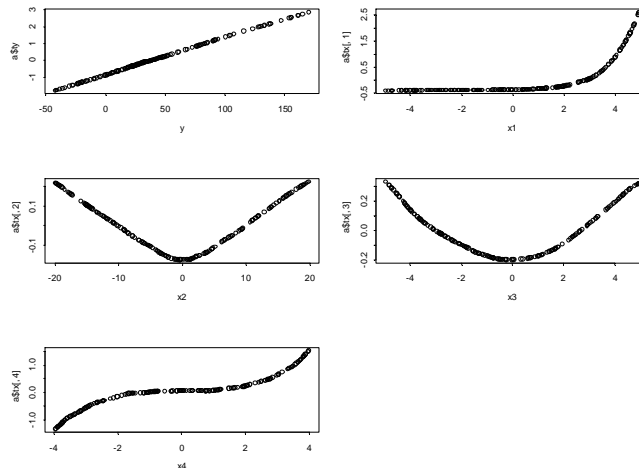


Figure 2. Scatterplots of ACE algorithm results (optimal transformation)

This results is indicating that the optimal parametric transformations have achieved. The ACE transformed variables has an adjusted  $R^2$  of 0.995, considerably better than the value of 0.651 obtained using OLS. Note that in theory ACE cannot produce a worse fit than ordinary regression, because if no transformations are

found to be necessary (i.e., the ordinary regression model is appropriate), then ACE would simply suggest nearly linear transformations for all the variables.

#### **4. Conclusions**

The ACE algorithm is a non-parametric automatic transformation method that produces the maximum multiple correlation of a response and a set of predictor variables. The approach solves the general problem of establishing the linearity assumption required in regression analysis, so that the relationship between response and independent variables can be best described and existence of non-linear relationship can be explored and uncovered. An examination of these results can give the data analyst insight into the relationships between these variables, and suggest if transformations are required.

The ACE plot is very useful for understanding complicated relationships and it is an indispensable tool for effective use of the ACE results. It provides a straightforward method for identifying functional relationships between dependent and independent variables. There will often be a number of potential candidates for transformation of a variable suggested by the ACE plot that fit the data well according to  $R^2$ .