Journal of Big Data

RESEARCH Open Access

Albatross analytics a hands-on into practice: statistical and data science application

Rezzy Eko Caraka[1,2*], Youngjo Lee[2*], Jeongseop Han[2], Hangbin Lee[2], Maengseok Noh[3], Il Do Ha[3], Prana Ugiana Gio[4] and Bens Pardamean[5,6]

*Correspondence: rezzy.eko.caraka@brin.go.id; youngjo@snu.ac.kr

[1] Research Center for Data and Information Sciences, Research Organization for Electronics and Informatics, National Research and Innovation Agency, Bandung, West Java 40135, Indonesia
[2] Lab Hierarchical Likelihood, Department of Statistics, College of Natural Science, Seoul National University, 56-1 Mountain, Sillim-dong, Gwanak-gu, Seoul, Republic of Korea
Full list of author information is available at the end of the article

Abstract

Albatross Analytics is a statistical and data science data processing platform that researchers can use in disciplines of various fields. Albatross Analytics makes it easy to implement fundamental analysis for various regressions with random model effects, including Hierarchical Generalized Linear Models (HGLMs), Double Hierarchical Generalized Linear Models (DHGLMs), Multivariate Double Hierarchical Generalized Linear Models (MDHGLMs), Survival Analysis, Frailty Models, Support Vector Machines (SVMs), and Hierarchical Likelihood Structural Equation Models (HSEMs). We provide 94 types of dataset examples.

Keywords: Albatross analytics, R shiny, Data science, Statistics, Application

Introduction

The application of statistical data processing has grown during the last decade, starting from traditional methods to advanced methods, including machine learning and extensive data analysis. The objective of statistical inference is to draw conclusions about a study population based on a sample of observations. Recently, subjective specific beliefs have been developed by introducing random effects in various components of models [1]. Different study problems involve specific sampling techniques and a statistical model to describe the analyzed situation.

Albatross Analytics is a statistical and computational data analysis program belonging to the open-source software class built after the R program package with the S programming language. Albatross Analytics is currently under a project by HGLM's worldwide group. In particular, it provides a new unified state-of-the-art statistical package from basic analysis to advance analysis including various random-effect models (HGLMs, DHGLMs, MDHGLMs, and frailty models) whose implementations are generally difficult.

Meanwhile, the basis of Albatross analytics in R software is clear R software was first worked on by Robert Gentleman and Ross Ihaka of the University of Auckland's Statistics Department in 1995 [2, 3]. Most of the functionality and capabilities of Albatross Analytics can be obtained through Add—packages/libraries.

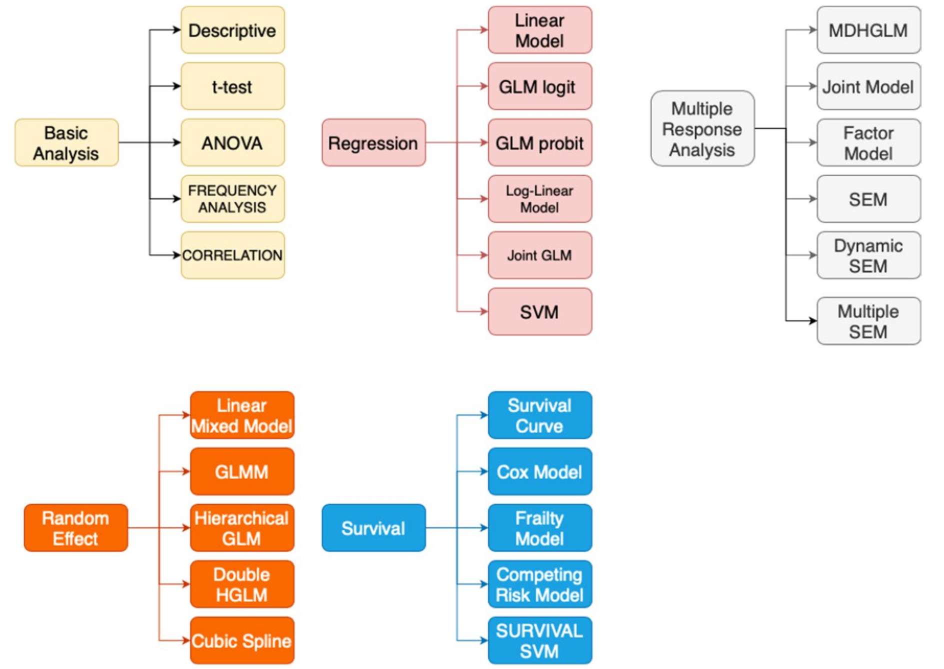


Fig. 1 Features in albatross analytics

A library is a collection of commands or functions that can perform specific analyses. For instance, this implements double hierarchical generalized linear models in which the mean, dispersion parameters for the variance of random effects, and residual variance (overdispersion) can be further modeled as a random-effect model may use DHGLM [4], MDHGLM by Lee [5–7]. This package allows various models for multivariate response variables where each response is assumed to follow double hierarchical generalized linear models. See also further HGLM applications for machine learning [4], schizophrenic behavior data [8], variable selection methods [9], non-Gaussian factor [10], factor analysis for ordinal data [11], survival analysis [12], longitudinal outcomes and time-to-event data [13], and recent advanced topics [14–17].

The FRAILTYHL package fits semi-parametric frailty and competing risk models using the h-likelihood. This package allows lognormal or gamma frailties for random-effect distribution, and it fits shared or multilevel frailty models for correlated survival data. Functions are provided to format and summarize the FRAILTYHL results [18]. The estimates of fixed effects and frailty parameters and their standard errors are calculated. We illustrate the use of our package with two well-known data sets and compare our results with various alternative R-procedures. Refers to the application of semi-competing risks data [19], and clustered survival data [20, 21]. This paper addresses and explains what Albatross Analytics is and include how to use it in statistical and data science application. The advantage of Albatross Analytics is the user can analyze and interpret the data easily. Meanwhile, Fig. 1 shows the feature of Albatross Analytics, including fundamental analysis, random effect, regression, survival analysis, and multiple response analysis. This paper aims to express the application of Albatross Analytics software to handle statistical analysis in broad areas. Long story short, we provide illustrative examples. A

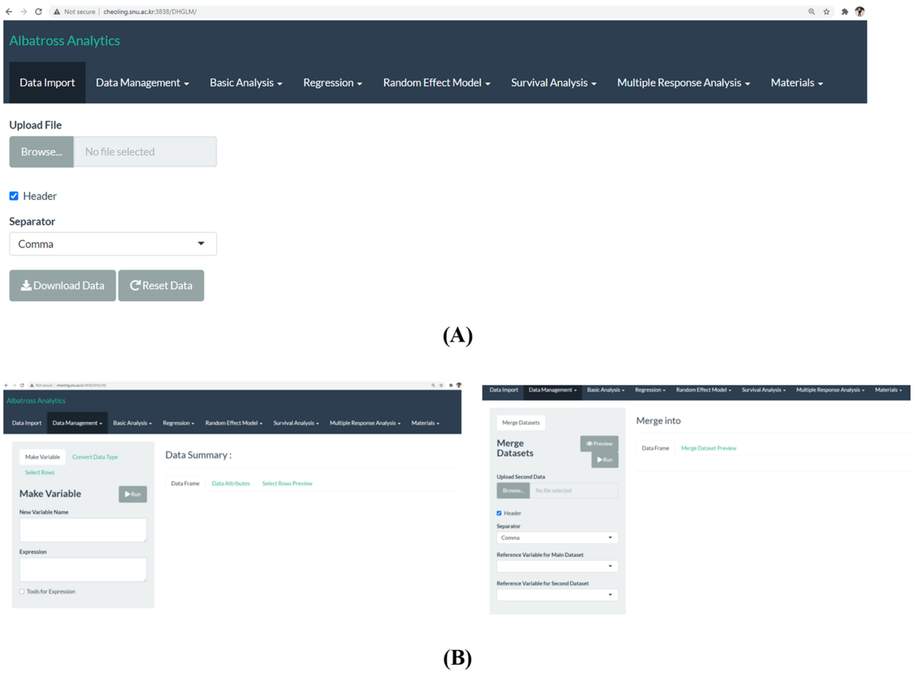


Fig. 2 Main feature (**A**) and Data management in albatross analytics (**B**)

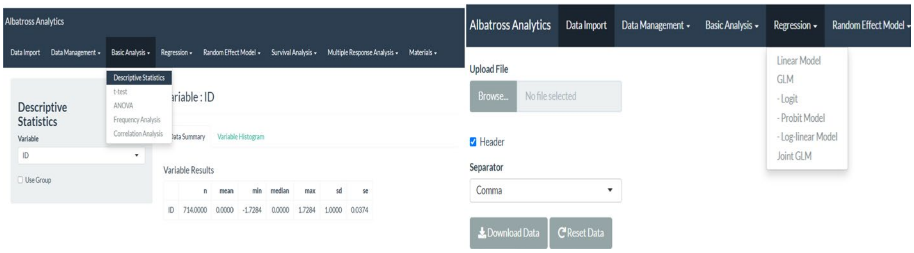
hands-on various applications including HGLM, DHGLM, MDHGLM, Survival Analysis, Frailty Models, Support Vector Machine, and Structural Equation Models.

Illustrative examples

Data management

In today's world, data is the driving factor behind all establishments. As institutions keep collecting so much data, there is a need to handle the quality of the data becoming more notable by the day. Data Quality Management is the set of measures applied by a technical team or a database management system to enable good new knowledge [22–24]. The above collection of techniques is decided to carry out during the data management pathway, from data capture to execution, dissemination, and interpretation [24–26]. In line with this, the data management is the process of processing, managing, and maintaining data quality [27, 28]. Effective data management can increase the efficiency of research work [26, 29]. Figure 2a describes the main features available in Albatross Analytics. In the import data section, users can maximize this feature to upload data to be processed where the possible files are in excel and txt formats, respectively. For instance, Fig. 2b explains how to make a new variable feature, merge the dataset, and add new variables.

Each expression or variable has a data type such as numeric, integer, complex, logical, and character. The data types in Albatross analytics are expressed in class. A class is a combination of data types and the operations performed against the dataset type. The Albatross analytics look at the data as objects having attributes or properties. Data properties are defined by data type.



Descriptive Statistics

Variable: ID

□ Use Group

ariable : ID

Variable Results

	n	mean	min	median	max	sd	se
ID	714.0000	0.0000	-1.7284	0.0000	1.7284	1.0000	0.0374

Upload File

No file selected

☑ Header

Separator: Comma

Fig. 3 Basic analysis and regression

Basic analysis and GLMs

Descriptive statistics have been used to identify the specific characteristics of the data in the interpretation. We provide simple details of the findings and the procedures followed in Fig. 3. Alongside primary frequency distribution, we form the basis of almost all quantitative analyses of the results. Descriptive statistics shall be used to present practical explanations understandably. Descriptive statistics allow one to interpret enormous amounts of data in a structured way.

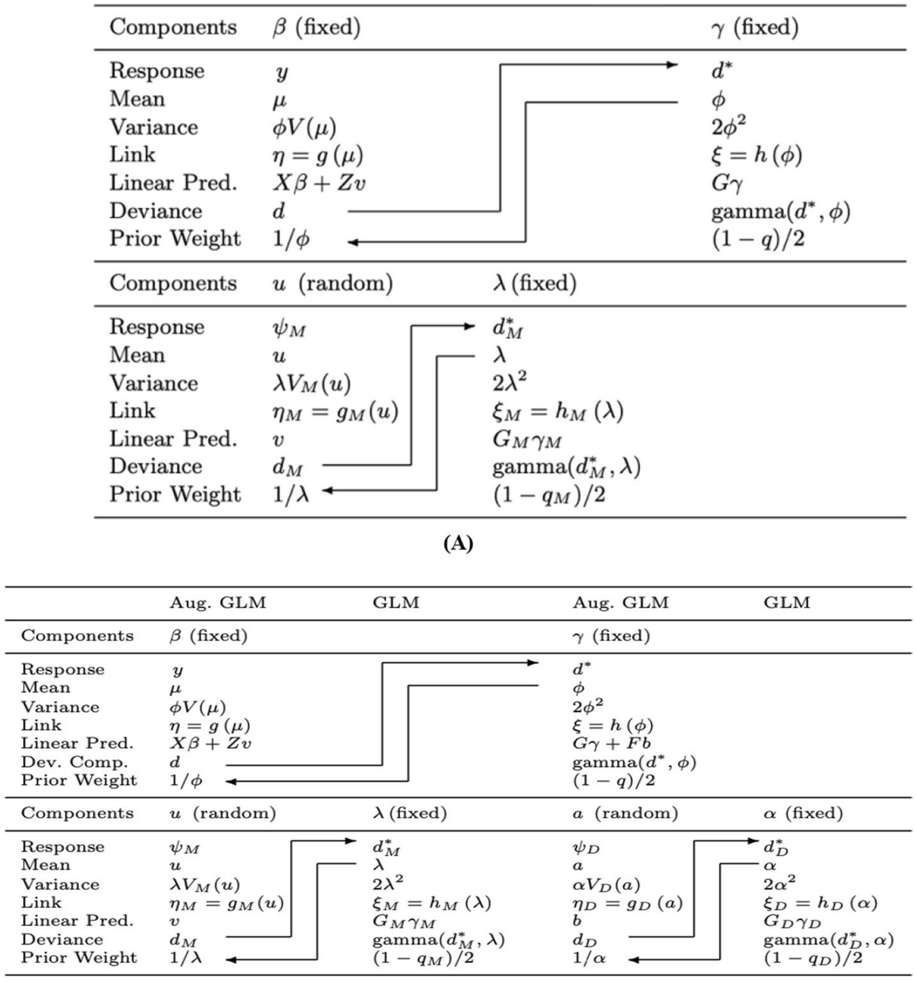
The *t*-test can be used to compare the means of two groups of data with the type of interval scale variable. Sometimes We will come across a study that aims to compare the mean of a sample with the mean of the entire population. Research models like this are rare, but the researcher can still provide valuable assumptions. We can do two kinds of tests, including *z*-test and a *t*-test. The condition we need to pay attention to is the population's standard deviation. If we know the standard deviation, we get it using the *z*-test. This will be found very rarely or never. Therefore, the most frequently used test is the *t*-test because we do not need to know the standard deviation of the population we study.

Furthermore, the use of the *t*-test on two samples is divided into two types based on the characteristics of the two samples. The first is the *t*-test on two independent samples. This means that the two samples to be studied came from two different groups and were given further treatment.

During the research, the use of analysis of variance is fundamental. One of the assumptions that must be met is that the population variances are the same, so we need to test the hypothesis. The purpose of the analysis of variance (ANOVA) is to determine the similarity of several population means. One-way ANOVA may be used if only one factor is involved. Two types of tests can be used in ANOVA testing, including formal and visual tests.

Meanwhile, the statistical test can be conducted by model checking plot. If the plot does not form a specific pattern, it is said that the homogeneity of the variance is fulfilled. We know the characteristics of each variable using descriptive analysis. In addition, we may see the relationship between variables, either normal or non-normal data [30]. With this correlation test, we want to know the similarity of the trends of the two variables. When the value of variable increases, it will also be accompanied by an increase or decrease in the value of other variables [31].

One main factor determines the test method used, namely the distribution of the data to be tested. We can use the parametric correlation test if the data distribution is normal,



Components	β (fixed)	γ (fixed)
Response	y	d^*
Mean	μ	ϕ
Variance	$\phi V(\mu)$	$2\phi^2$
Link	$\eta = g(\mu)$	$\xi = h(\phi)$
Linear Pred.	$X\beta + Zv$	$G\gamma$
Deviance	d	$\text{gamma}(d^*, \phi)$
Prior Weight	$1/\phi$	$(1-q)/2$

Components	u (random)	λ (fixed)
Response	ψ_M	d_M^*
Mean	u	λ
Variance	$\lambda V_M(u)$	$2\lambda^2$
Link	$\eta_M = g_M(u)$	$\xi_M = h_M(\lambda)$
Linear Pred.	v	$G_M\gamma_M$
Deviance	d_M	$\text{gamma}(d_M^*, \lambda)$
Prior Weight	$1/\lambda$	$(1-q_M)/2$

(A)

	Aug. GLM	GLM	Aug. GLM	GLM
Components	β (fixed)		γ (fixed)	
Response	y		d^*	
Mean	μ		ϕ	
Variance	$\phi V(\mu)$		$2\phi^2$	
Link	$\eta = g(\mu)$		$\xi = h(\phi)$	
Linear Pred.	$X\beta + Zv$		$G\gamma + Fb$	
Dev. Comp.	d		$\text{gamma}(d^*, \phi)$	
Prior Weight	$1/\phi$		$(1-q)/2$	
Components	u (random)	λ (fixed)	a (random)	α (fixed)
Response	ψ_M	d_M^*	ψ_D	d_D^*
Mean	u	λ	a	α
Variance	$\lambda V_M(u)$	$2\lambda^2$	$\alpha V_D(a)$	$2\alpha^2$
Link	$\eta_M = g_M(u)$	$\xi_M = h_M(\lambda)$	$\eta_D = g_D(a)$	$\xi_D = h_D(\alpha)$
Linear Pred.	v	$G_M\gamma_M$	b	$G_D\gamma_D$
Deviance	d_M	$\text{gamma}(d_M^*, \lambda)$	d_D	$\text{gamma}(d_D^*, \alpha)$
Prior Weight	$1/\lambda$	$(1-q_M)/2$	$1/\alpha$	$(1-q_D)/2$

(B)

Fig. 4 HGLM algorithm (**A**), and DHGLM algorithm (**B**)

including Pearson's correlation coefficient. Besides, if the data distribution is not normal, we can use Kendall's rank correlation and Spearman's rank correlation, which are non-parametric correlation tests.

Regression analysis tests the causal relationship between variables—one variable as the independent variable and one other variable as the dependent variable. Numerous regression approaches, including Poisson regression, were used during the 1970s. Linear regression and logistic regression require a unique estimation algorithm by maximizing the likelihood. Figure 4 explains that Albatross Analytics provides features for using the Linear model, GLM Logit Model, GLM Probit Model, Log-linear Model, and joint GLM.

GLM describes a family of models where the response comes from the exponential family of distributions. The method used to t-test or F-test and inferences of these models is maximum likelihood (ML). In the GLM family of models, an IWLS algorithm can compute the ML estimates and their standard errors. Hence, the computational machinery developed for least-squares estimation for linear models can fit GLMs, but the statistical method is based on ML.

Hands-on and application albatross analytics

Hierarchical generalized linear models (HGLMs)

Albatross Analytics' distinct advantage is its unified analysis of random effect models. Various random effect models can be represented as HGLMs and estimated by h-likelihood procedures [32]. HGLMs are defined as follows:

(1) Conditional on random effects u, the responses y follows a GLM family, satisfying

$$E\left(y|u\right) = \mu \text{ and } var\left(y|u\right) = \phi V(\mu),$$

for which the kernel of the log-likelihood is given by

$$\sum \left\{ \mathrm{y}\theta - \mathrm{b}(\theta) \right\} / \phi,$$

where $\theta = \theta(\mu)$ is the canonical parameter. The linear predictor takes the form in Eq. (1):

$$\eta = g(\mu) = X\beta + Zv, \tag{1}$$

where $v = v(u)$, for some monotone function $v(\cdot)$ and the link function $g(\mu)$.

(2) The random component u follows a (conjugate) distribution to a GLM family of distributions with parameter λ.

To infer the HGLM, Lee and Nelder [32] proposed using the h-likelihood. The h (log-) likelihood is defined as Eq. 2:

$$h = \log f_{\beta,\phi}(y|v) + \log f_{\lambda}(v). \tag{2}$$

The GLM attributes of an HGLM are summarized in Fig. 4.

In Bissell's fabric study, the response variable y is the number of faults in a bolt of the fabric of length l. Table 1 represents the results of the fabric study. Figure 6 illustrates the negative binomial model fitted via Poisson-gamma HGLM with saturated random effects for the complete response. In addition, the model checking plot is presented in Fig. 5.

Table 1 Results for fabric study

Estimates from the mean model				
	Estimate	**Std.error**	**t-value**	**p-value**
Intercept	− 3.77988	0.01443	− 2.61933	0.00881
log (*l*)	0.94236	0.00226	4.17445	0.00003
Estimates from the dispersion model				
		Estimate		**Std.Error**
log (λ)		− 2.07637		0.02103
Likelihood				
− 2ML	**− 2RL**	**cAIC**	**Scaled Deviance**	**df**
175.75601	179.91906	172.76006	14.33786	14.43461

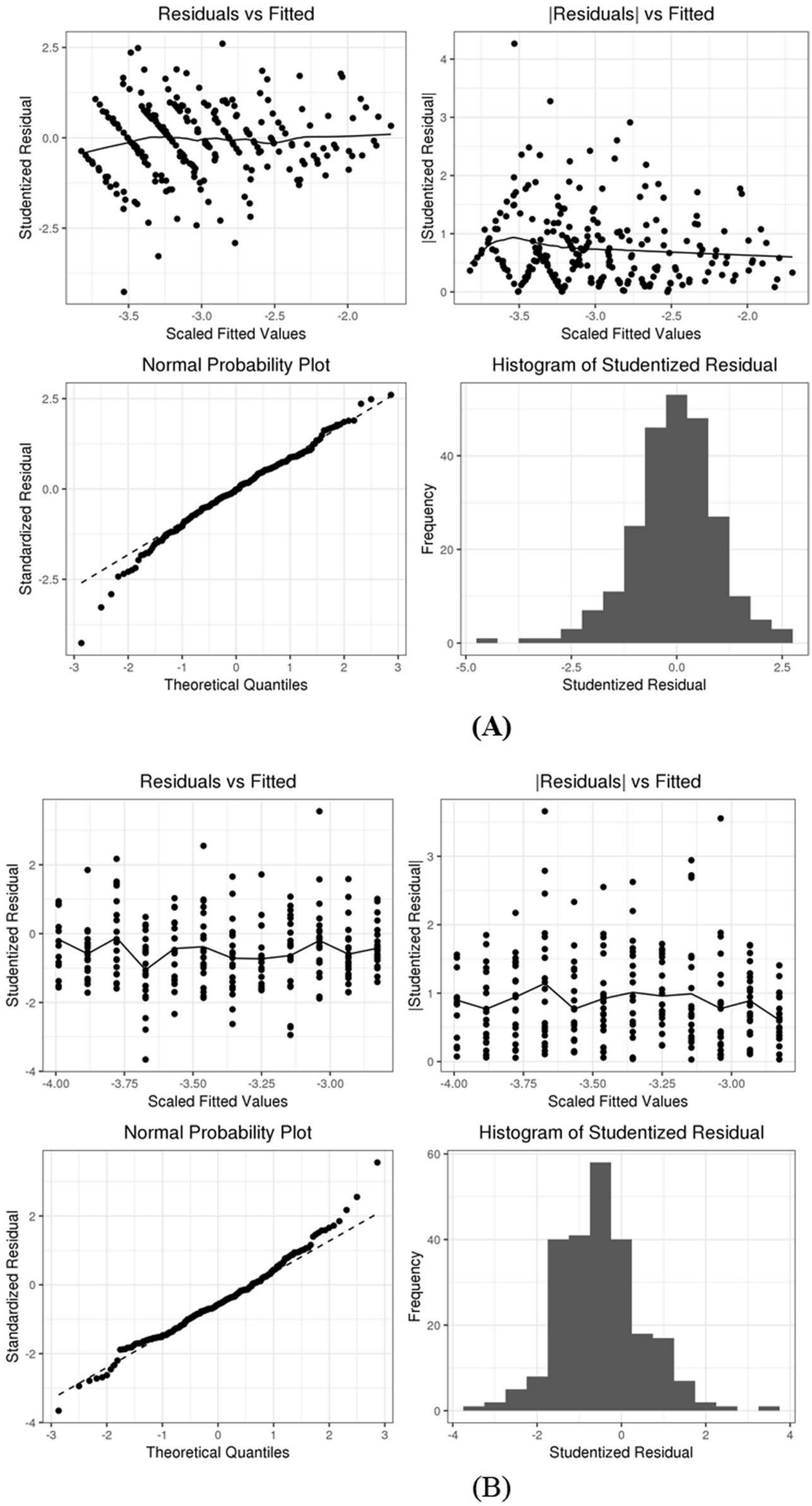


Fig. 5 Model checking plots of DHGLM for mean (**A**), and model checking plots of DHGLM for dispersion (**B**)

Double hierarchical generalized linear models (DHGLMs)

HGLM can be extended by allowing additional random effects in their various components. Lee and Nelder [32] introduced a class of double HGLMs (DHGLMs) in which random effects can be specified in both the mean and the residual variances. Heteroscedasticity between clusters can be modeled by introducing random effects in the dispersion model as heterogeneity between clusters in the mean model. With DHGLMs, it is possible to have robust inference against outliers by allowing heavy-tailed distribution. Many models can be unified and extended further by the use of DHGLMs. These also include models in the finance area such as autoregressive conditional heteroscedasticity (ARCH) models, generalized ARCH (GARCH), and stochastic volatility (SV) models. Models can be further extended by introducing random effects in the variance terms. Suppose that conditional on the pair of random effects (a, u), the response y satisfies.

$$E\left(y|a,u\right) = \mu \text{ and } var\left(y|a,u\right) = \phi V(\mu).$$

The critical extension is to introduce random effects into the component ϕ:

(1) Given u, the linear predictor for μ takes the HGLM form in Eq. 1 where $g(\cdot)$ is the link function, X and Z are model matrices, $v = g_M(u)$ for some monotone function, $g_M(u)$ are the random effects, and β are the fixed effects. Moreover, dispersion parameters λ for u have the GLM form in Eq. 3

$$\xi_M = h_M(\lambda) = G_M\gamma_M, \tag{3}$$

where $h_M()$ is the link function, G_M is the model matrix and γ_M is fixed effects.

(2) Given a, the linear predictor for ϕ takes the HGLM form as described in Eq. 4

$$\xi = h(\phi) = G\gamma + Fb, \tag{4}$$

where $h()$ is the link function, G and F are model matrices, $b = g_D(a)$ for some monotone function, $g_D(a)$ are the random effects, and γ are the fixed effects. Moreover, dispersion parameters α for a have the GLM form, as shown in Eq. 5.

$$\xi_D = h_D(\alpha) = G_D\gamma_D \tag{5}$$

where $h_D(\cdot)$ is the link function, G_D is the model matrix and γ_D is fixed effects. Here, the labels M and D stand for mean and dispersion, respectively. The GLM attributes of a DHGLM are summarized in Fig. 4.

However, We illustrate an example of how to fit the DHGLM. Hudak [33] presented crack growth data, listed in Lu [34]. Each of 21 metallic specimens was subjected to 120,000 loading cycles, with the crack lengths recorded every 10,000 cycles. Let l_{ij} be the crack length of the i-th specimen at the j-th observation and $y_{ij} = l_{ij} - l_{ij-1}$ be the corresponding increment of crack length (response variable) measured in inches, which always has a positive value. A detailed description of the model can be found in Table 2, and Fig. 5a and b represent the mean and the dispersion, respectively [5]. Compared to an HGLM, DHGLM gives model checking plots for mean and dispersion, respectively.

Table 2 Results of DHGLM for crack growth data

Estimates from the mean model				
	Estimate	**Std. error**	**t-value**	**p-value**
intercept	− 5.64457	0.00007	− 429.0492	0.00000
crack	2.40596	0.00005	238.59171	0.00000
		Estimate		**Std. error**
$log\,(\lambda)$		− 3.44556		0.00983
Estimates from the dispersion model				
		Estimate		**Std. error**
Intercept		− 3.01495		0.00735
Cycle		− 11.44552		0.07700
		Estimate		**Std. error**
$log\,(\alpha)$		− 0.40365		0.00229
Likelihood				
− 2ML	**− 2RL**	**cAIC**	**Scaled deviance**	**Df**
− 1910.493	− 1602.520	− 1620.783	215.58663	215.58663

Table 3 Descriptive Statistics for crack growth data

Dose (g/kg)	**Dams**	**Live**	**Malformations**		**Weight (g)**	
			No	**%**	**Mean**	**S.D**
0.00	25	297	1	0.34	0.972	0.0976
0.75	24	276	26	9.42	0.877	0.1041
1.50	22	229	89	38.86	0.764	0.1066
3.00	23	226	129	57.08	0.704	0.1238

Multivariate double hierarchical generalized linear models (MDHGLM's)

Using h-likelihood, multivariate models are directly extended by assuming correlations among random effects in DHGLMs for different responses. The use of h-likelihood indicates that interlinked GLM fitting methods for HGLMs can be easily extended to fit multivariate HGLMs (MDHGLMs). Moreover, the resulting algorithm is numerically efficient and gives statistically valid inferences. In this paper, we present the example for MDHGLM. For more details, see [35] Meanwhile, Price et al. [36] presented data from a study on the developmental toxicity of ethylene glycol (EG) in mice. Table 3 summarizes the data on malformation (binary response) and fetal weight (continuous response) and shows clear dose-related trends concerning both responses.

To fit the EG data, the following bivariate HGLM is considered:

(1) $y_{1ij}|w_i \sim N\left(\mu_{ij}, \phi\right), \mu_{ij} = x_{1ij}\beta_1 + w_i,$

(2) $y_{2ij}|u_i \sim Ber\left(p_{ij}\right), logit(p_{ij}) = x_{2ij}\beta_2 + u_i,$ and

(3) $(w_i, u_i)^T \sim BVN(\mathbf{0}, \mathbf{\Sigma}), cor(w_i, u_i) = \rho.$

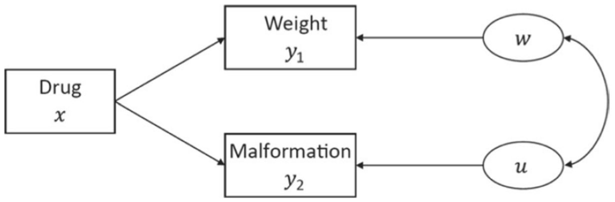


Fig. 6 Path diagram for MDHGLM towards weight and misinformation

Table 4 Results for malformation model

Estimates from the mean model				
	Estimate	**Std. error**	**t-value**	**p-value**
Intercept	0.97800	0.01410	69.36273	0.00000
Dose	− 0.16300	0.02537	− 6.42366	0.00000
Dose^2	0.02500	0.00794	3.14776	0.00165
		Estimate		**Std. error**
log (λ)		− 4.95674		0.21902
Estimates from the dispersion model				
		Estimate		**Std. error**
Intercept		− 4.95674		0.21902
Likelihood				
− 2ML	**− 2RL**	**cAIC**	**Scaled deviance**	**df**
− 2173.975	− 2150.607	− 2330.449	942.09806	942.09806

Table 5 Results for weight model

Estimates from the mean model				
	Estimate	**Std. error**	**t-value**	**p-value**
Intercept	− 5.85700	0.77059	− 7.60063	0.00000
Dose	4.74202	0.94058	5.04161	0.00000
Dose^2	− 0.88501	0.23378	− 3.78563	0.00015
		Estimate		**Std. error**
log (λ)		0.61297		0.33902
Likelihood				
− 2ML	**− 2RL**	**cAIC**	**Scaled deviance**	**df**
719.47338	724.13384	689.55801	606.88179	986.66086

Figure 6 shows the path diagram of the model for the EG data. The malformation model information is given in Table 4, with cAIC for the evaluation models. In line with this, we get the result for the weight model in Table 5 and correlation in Table 6.

Table 6 Correlation towards crack growth

Estimates for correlation	w_i	u_i
w_i	1	− 0.61933
u_i	− 0.61933	1

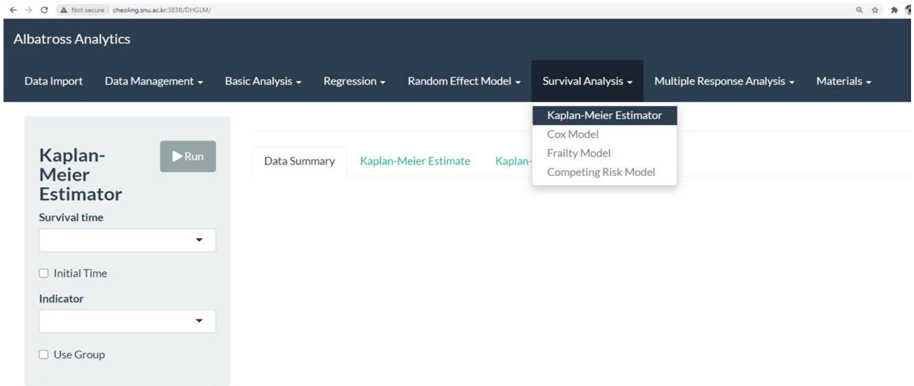


Fig. 7 Survival analysis feature

Survival analysis

Albatross Analytics also provides features for survival analysis, which represent in Fig. 7 by including incomplete data caused by censoring in survival time (time-to-event) data including Kaplan–Meier Estimator, Cox Model, Frailty Model [7], and Competing Risk Model [19, 37]. More instances of the Kaplan Meier curve describe the relationship between the estimated survival function at time t and the survival time. The vertical axis represents the estimated survival function, and the horizontal axis represents the survival time.

Cox proportional hazards (PH) regression is used to describe the relationship between the hazard function of survival time and independent variables which are considered to affect survival time. Cox regression is a common regression used in survival analysis because it does not assume a particular statistical distribution (e.g., baseline hazard) of the survival time.

Cox's PH model is widely used to analyze survival data. This method is helpful with its semi-parametric existence, whereby baseline hazards are non-parametric, and treatment effects are estimated parametrically. A partial likelihood has usually been used to accommodate such a semi-parametric form. However, it can also be fitted with Poisson GLM methods. Moreover, they are sluggishly led to many nuisance parameters induced by non-parametric measurement hazards. Meanwhile, using the h-likelihood theory, we can prove that Poisson HGLM methodologies could be used for such kinds of modeling techniques. That being said, this method is again sluggish since the number of nuisance parameters in non-parametric baseline hazards grows with the number of events.

Example 1 using incomplete data caused by censoring in survival data

In Fig. 7, we study the analysis of incomplete data caused by censoring survival data. Cox's PH model is widely used to analyze survival data. Frailty models with a non-parametric baseline hazard extend the PH model by allowing random effects in hazards and have been widely adopted for the analysis of correlated or clustered survival data using h-likelihood theory; we can show that Poisson HGLM algorithms can be used to fit the frailty models [12, 38–43].

Data consist of right-censored observation from q subjects, with n_i observations each $(i = 1, \ldots, q)$, $n = \Sigma_i n_i$ as the total sample size, T_{ij} as survival time for the j-th observation of the i-th subject $(j = 1, \ldots, n_i)$, C_{ij} as the corresponding censoring time, $y_{ij} = \min\{T_{ij}, C_{ij}\}, \delta_{ij} = I(T_{ij} \leq C_{ij})$, and u_i as observed frailty for the i-th subject. The conditional hazard function of T_{ij} given u_i is of the form in Eq. 6

$$\lambda_{ij}(t|u_i) = \lambda_{0j}(t)\exp\left(x_{ij}^T\beta\right)u_i \tag{6}$$

Here $\lambda_0(\cdot)$ is an unspecified baseline hazard function and $\beta = \left(\beta_1, \ldots, \beta_p\right)^T$ is a vector of regression parameters for the fixed covariates x_{ij}. Here, the term $x_{ij}^T\beta$ does not include an intercept term because of identifiability. Then, we assume that the frailties u_i are i.i.d random variables with a frailty parameter α. We often assume gamma or log-normal distribution for u_i; that is, it is gamma frailty with $E(u_i) = 1$ and $\mathrm{var}(u_i) = \alpha$ and log-normal frailty with $v_i = \log u_i \sim N(0, \alpha)$. Meanwhile, the multi-component frailty models can be expressed in Eq. 7, with the linear predictor

$$\eta = X\beta + Z^1v^1 + Z^kv^k + \cdots + Z^kv^k \tag{7}$$

X is $n \times p$ model matrix for β, andZ^r is $n \times q_r$ model matrices corresponding to the frailties v^r. At the same time, $v^{(r)}$ and $v^{(i)}$ are independent for $r \neq I$. Also, Z^r has indicator values such that $Z_{st}^{(r)} = 1$ if observation s is a member of the subject t in the r-th frailty component, and 0 otherwise.

To the illustration, below we present two examples. Example 1 considers the dataset of the recurrence of infections in kidney patients using a portable dialysis machine. The data consist of the first and second recurrences of kidney infection in 38 patients. The catheter is later removed if the condition occurs and can be removed for other reasons, which we regard as censoring (about 24%).

In Example 1, the variables consist of 38 patients (*id*), time until infection since the catheter insertion (*time*), and a censoring indicator (1, infection; 0, censoring) for *status*, age of the patient (*age*), sex (*sex*) of the patient (1, male; 2, female), disease types (*disease*) following GN, AN, PKD, other, and estimated frailty (*frail*). The survival times (1st and 2nd infection times) for the same patient are likely to be correlated because of shared frailty describing the common patient's effect. We thus fit log-normal frailty models with two covariates, *sex,* and *age*. Here, we consider the patient as frailty. Figure 8 presents the Kaplan–Meier plot for the estimated survival probability of the sex (sex1, male; sex2, female). This shows that the female group has overall higher survival (i.e., less infectious) probabilities than ones in the male group. Table 7 summarizes the estimated results of the log-normal frailty model. We show the estimated frailty in Fig. 9. For further discussions in survival analysis, see [18].

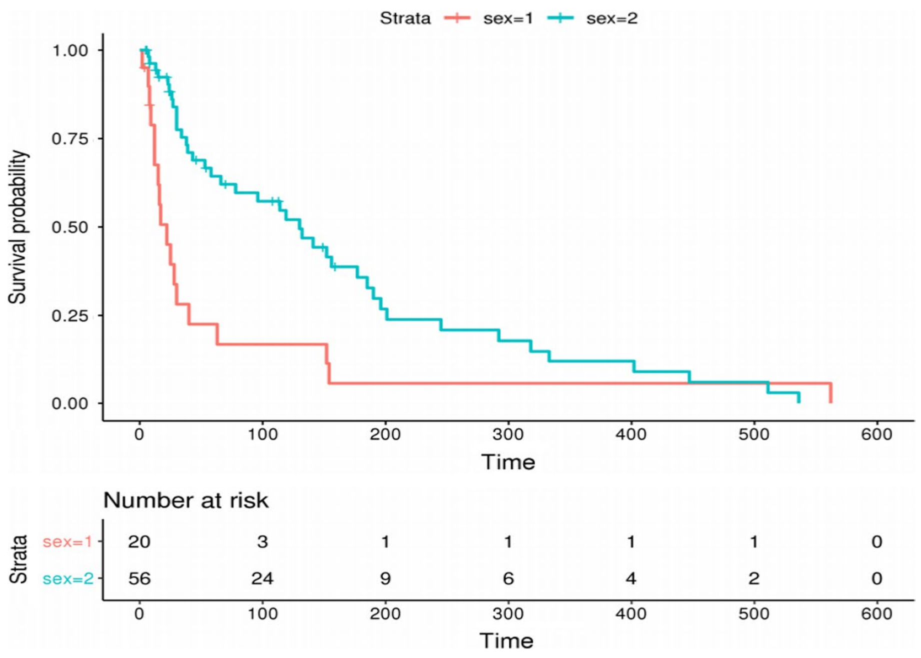


Fig. 8 Survival probability towards sex

Table 7 Model description FOR LOG-normal frailty

Model description				
	Model	**Number of data**	**Number of events**	**Method**
1	Log normal frailty model	76	58	HL (0,1)
Estimates from the mean model				
	Estimate	**Std.error**	**t-value**	**p-value**
Sex	− 1.38043	0.43082	− 3.20419	0.00135
Age	0.00488	0.01209	0.40412	0.6812
Estimates from the dispersion model				
		Estimate		**Std. error**
Id		0.53448		0.38842
Likelihood				
$-2\,h_0$		$-2\,h_p$		$-2\,p_{\beta,v}(h_p)$
330.40166		390.7718		371.54037
AIC				
cAIC		**mAIC**		**rAIC**
362.45706		370.70076		373.54037

Example 2 placebo-controlled rIFN-g in the treatment of CGD

Example 2, in the following case examples, consists of a placebo-controlled rIFN-g in the treatment of CGD [44, 45]. One hundred twenty-eight patients from 13 centres were tracked for around 1 year. The survival times are the recurrent infection times of each

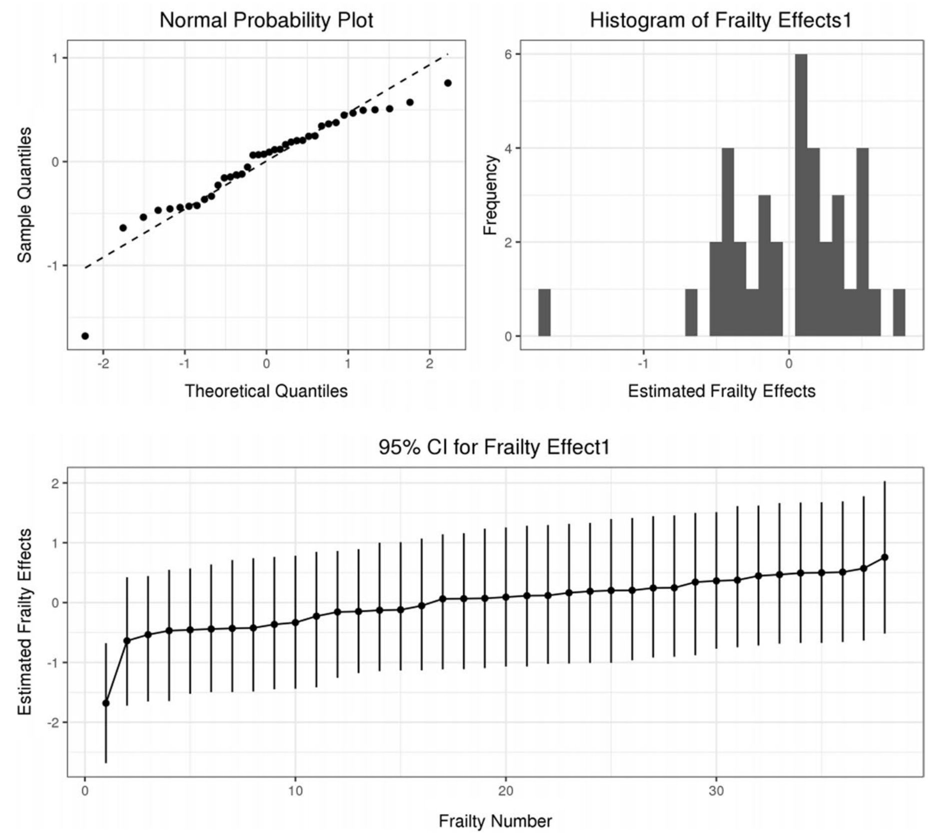


Fig. 9 Estimated frailty in the kidney infection data

patient. Censoring occurred at the last observation for all patients, except one, who experienced a severe infection on the date he left the study. About 63% of the data were censored. The recurrent infection times for a given patient are likely to be correlated. Also, each patient belongs to one of the 13 centres. The correlation may be attributed to the patient effect and centre effect. Meanwhile, the recurrent infection times of each patient or censoring time (*tstart–tstop),* 128 patients (*id*), 13 centers (*center*), rIFN-g or placebo (*treat*), censoring indicator (1, *infection observed;* 0, *censored*) for *status,* data of randomization (*random*) information about patients at study entry (*sex, age, height, weight*), the pattern of inheritance (*inherit*), use of steroids at study entry 1(*yes*), 0(*no*) (*steroids*), use of propylac antibiotics at study entry. 1(*yes*), 0(no) (propylac), categorization of the centers into four groups (*hos.cat*), and observation number within-subject (*enum*). We fit multilevel log-normal frailty with two frailties and a single covariate, treatment. Here, the two frailties are random center and patient terms, with their structures given in Eq. 8.

$$\begin{aligned} &\eta = X\beta + Z^1 v^1 + Z^2 v^2 \\ &v^1 \sim N\left(0, \alpha_1 I_{q1}\right) \\ &v^2 \sim N\left(0, \alpha_2 I_{q2}\right) \end{aligned} \tag{8}$$

Here ν^1 is center frailty, and ν^2 is patient frailty. For testing the need for a random component i.e.,($\alpha_1 = 0 \text{ or } \alpha_2 = 0$) we use the deviance $-2p_{\beta,\text{v}}\left(h_p\right)$ and fit the following four models.

M1 Cox's model without frailty $(\alpha_1 = 0 \text{ or } \alpha_2 = 0) : -2p_{\beta,\text{v}}\left(h_p\right) = 707.48$

M2 model without patient effect $(\alpha_1 > 0 \text{ or } \alpha_2 = 0) : -2p_{\beta,\text{v}}\left(h_p\right) = 703.66$

M3 model without center effect $(\alpha_1 = 0 \text{ or } \alpha_2 > 0) : -2p_{\beta,\text{v}}\left(h_p\right) = 692.99$

M4 multilevel model $(\alpha_1 > 0\text{or}\alpha_2 > 0) : -2p_{\beta,\text{v}}\left(h_p\right) = 692.95.$

Table 8 represents the model description. The deviance difference (692.99 − 692.95=0.04) between M3 and M4 ($0.04 < 2.71 = \chi^2_{0.10}(1)$) indicates the absence of the random center effects, and the deviance difference between M2 and M4 (10.71) shows the necessity of random patient effects. In addition, the deviance difference between M1 and M3 (14.49) presents the random patient effect with or without random center effects. All of the three criteria (cAIC, mAIC and rAIC) also choose M3 among the M1–M4. Figure 10 presents the estimated frailty effects of this study. The explanations of model evaluation toward these three criteria can be seen in the Appendix.

Support vector machine using H likelihood

Support Vector Machine (SVM) is a supervised learning method for classification and regression using non-linear boundaries by feature space [4, 46–49]. We present a Support Vector Machine (SVM) based on the HGLM method [4]. The match between the observed response and the model output is optimized. The output model is a feature or prognostic function also referred to as a utility function and more specifically in medical research it is called the prognostic index or health function, defined in Eq. 9:

Table 8 Model description for log normal frailty

Model description

	Model	Number of data	Number of events	Method
1	Log normal frailty model	203	76	HL(1,1)

Estimates from the mean model

	Estimate	Std.error	t-value	p-value
treatIFN-g	− 1.18425	0.34065	− 3.47642	0.00051

Estimates from the dispersion model

	Estimate	Std.Error
center	0.03003	0.15720
id	1.00206	0.50880

Likelihood

$-2\,h_0$	$-2\,h_p$	$-2\,p_v\left(h_p\right)$	$-2\,p_{\beta,v}\left(h_p\right)$
603.31409	853.69944	692.62963	692.94167

AIC

cAIC	AIC	rAIC
684.91665	698.62963	696.94167

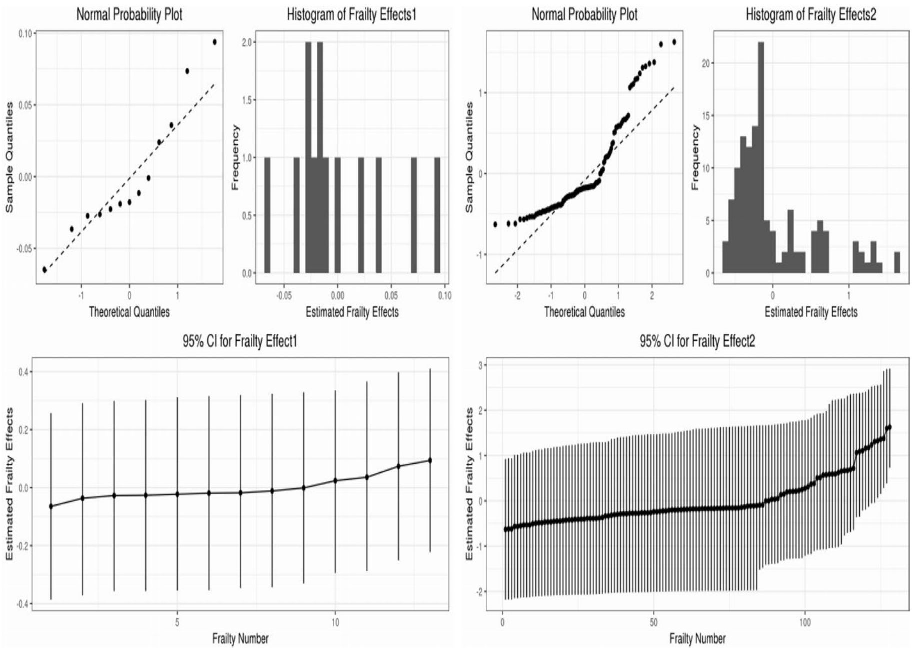


Fig. 10 Estimated frailty effects in the CGD recurrent infection data

$$u(x) = w^T\varphi(x) \tag{9}$$

Here $u : \mathbb{R}^d \to \mathbb{R}$, w is a vector of unknown d parameters and $\varphi(x)$ is the transformation of the covariates x. In non-linear SVM, the transformation function used is "Kernel Trick", see: [50–52] Kernel Trick calculates the scalar product in the form of a kernel function. The SVM model is implied with a constraint function that will get the right margin. The constraint function of the SVM model is shown in Eq. 10. If there is an error in ranking it is given by the slack variable $\xi_{ij} \geq 0$. The formulation of the SVM model is described in Eq. 10: cantered depression, and the latent person-

$$\begin{aligned} &\min_{w,\xi} \frac{1}{2}w^T w + \frac{\gamma}{2}\sum_{i<j} v_{ij}\xi_{ij} \\ &\text{constraint function} \begin{cases} w^T\varphi(x_j) - w^T\varphi(x_i) \geq 1 - \xi_{ij}, & \forall i < j \\ \xi_{ij} \geq 0, & \forall i < j \end{cases} \end{aligned} \tag{10}$$

with a regularization parameter $\gamma \geq 0$. v_{ij} is an indicator function of whether or not two subjects with observations i and j are comparable; it is 1 if i and j are comparable and 0 otherwise. In this paper, we use the dataset of the anatomy of an Abdominal Aortic Aneurysm (AAA), Aortic Anatomy on Endovascular Aneurysm Repair (EVAR), see [53]. The variables are described as follows: Y = Sex, X_1 = Age, X_2 = Aortic type Fusiform (1), Saccular (2), X_3 = Proximal neck length, X_4 = Proximal neck diameter, X_5 = Proximal neck angle, and X_6 = Max. Aneurysmal sac. We set the response variable towards simulation by following the Bernoulli distribution with 500 observations. In each scenario, the process of generating data is repeated 100 times. The parameter values used are: $\gamma = 0.7, Cost = 8$. Verbose takes advantage of a per-process runtime setting. Meanwhile, the SVM parameter setting is as follows:

- First simulation:
  Cluster method: "kmeans", cost = 8, lambda = 1, centers = 2, verbose = 0.
- Second simulation:
  Cluster method: "kernkmeans", cost = 8, lambda = 1, centers = 2, verbose = 0.
- Third simulation:
  Cluster.method = "kernkmeans", cost = 8, lambda = 1, centers = 3, verbose = 0.
- Fourth simulation:
  Cluster.method = "kernkmeans", cost = 8, lambda = 1, centers = 4, verbose = 0.

There are two types of model evaluation criteria: the classification stage and the HGLM analysis stage. Evaluation of the model's goodness at the classification stage uses AUC and is determined using the values contained in the confusion matrix, with

$$Sensitivity = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$Specitifity = \frac{True\ Negative}{False\ Positive + True\ Negative}$$

This simulation shows that HGLM performs better with high sensitivity because some of the data used is a binary case that SVM cannot handle. For more information on step construction using hierarchical likelihood towards SVM. Table 9 represents that the use of Ensemble SVM reduces the accuracy and other measures. When the mixture patterns exist in the predictor, Ensemble SVM improves SVM performance in two scenarios. Ensemble SVM performed almost as well as logistic regression, except for sensitivity. There is a decrease in performance in the Ensemble SVM model in the multicollinearity condition and linear combination between the predictor variables. Meanwhile, HGLM still has a good performance, which is represented in Fig. 11a and b, respectively.

Table 9 Models accuracy compariso

Methods	AUC	Accuracy	Sensitivity	Specificity
hglm	0.9877261	0.98922	0.9843113	0.9925881
svm	0.9753548	0.98356	0.9774139	0.9857814
kern_csvm2	0.9842441	0.98828	0.9797909	0.9918297
kmean_svm2	0.9862619	0.99012	0.9840275	0.9925610
kern_csvm3	0.9852699	0.98816	0.9760351	0.9929982
kmean_svm3	0.9874062	0.99046	0.9823998	0.9937159
kern_csvm4	0.9860299	0.98870	0.9768563	0.9932874
kmean_svm4	0.9894770	0.99046	0.9769520	0.9956495

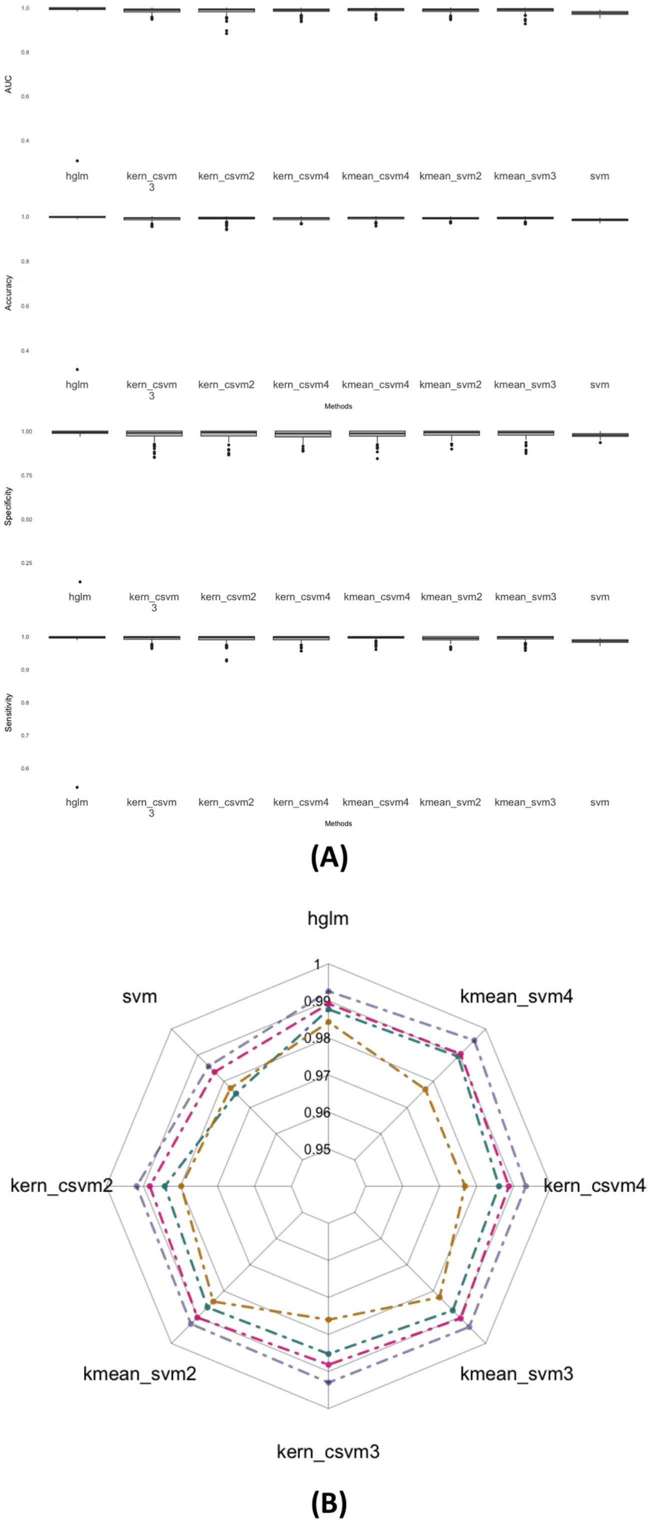


Fig. 11 AUC, accuracy, sensitivity, and specificity (**A**), and accuracy of all scenarios (**B**)

Using H-likelihood to structural equation model (HSEMs)

The is widely used in multidisciplinary fields [41]. To account for the information, [42, 43] performs the style of frequentist model averaging in structural equation modeling, non-linear structural equation modeling towards ordinal data [44] and partial least square [45, 46] and robust nonlinear with the interaction between exogenous and endogenous latent variables [47]. With an example we present a SEM method based on h-likelihood, called "hsem" [52].

In application, [48] uses two-level dynamic SEM on longitudinal data at Mplus. In this paper, we explicitly discuss how to use h-Likelihood in SEM. This data set consists of 50 repetitions on regular time scales for 100 individuals. For the response variable, the urge to smoke is on a standardized scale so that 0 corresponds to the average where the standard deviation is 1. Smokers can feel drastic mood changes. Starting from feeling happy then turning into sadness, this can show the characteristics of a person who is depressed. For those addicted, smoking can give a calm mind for a moment. The second model will answer the question; latent person predicts smoke, mean cantered depression, and the latent person-mean centered lag-1 urge to smoke. The model Eq. 11 is given as follows:

$$\begin{aligned} urge_{ti} &= \beta_{0i} + \beta_{1i} Time_{ti} + e_{ti}, \\ \beta_{0i} &= \gamma_{00} + u_{0i} \\ \beta_{1i} &= \gamma_{10} + u_{1i} \\ \boldsymbol{u}_i &= \begin{pmatrix} u_{0i} \\ u_{1i} \end{pmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & 0 \\ 0 & \tau_{11} \end{bmatrix} \right) \\ e_{ti} &\sim N\left(0, \sigma^2\right) \end{aligned} \tag{11}$$

Figure 12 represents the path diagram by using hsem. This same standard progression path across all respondents was defined through the fixed-effect model. In contrast, the person-specific random effects are used to catch the variance of each participant from the expected path. Meanwhile, the path diagram represents within-level and between-level models. As more instance, we provide the R package hsem [54].

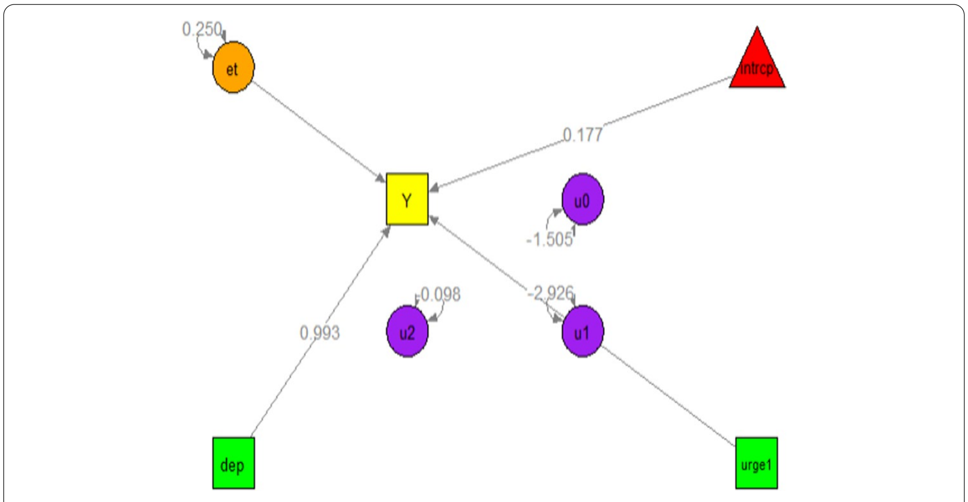


Fig. 12 Path diagram

Short review albatross analytics

This paper explains how Albatross software can be used for alternative multidisciplinary data processing. We offer model estimation, model checking plots, and visualization features to interpret information. Through data and R code, instances would further reveal the benefit of the HGLM model for particular statistical cases. The h-likelihood approach is distinct from both classical frequentist and Bayesian frameworks, while this encompasses inference of both fixed and random unknowns. The main benefit over classical frequentist approaches is that it would be possible to infer unobservable quantities, such as random effects, and therefore, observations could be rendered. Whenever a statistical model has been selected for the research, the likelihood contributes to the direction of inferential statistics.

Throughout direct ties with the establishment of h-likelihood, the nomenclature has already been used in which a wide variety of likelihoods have now been established. Most are through theoretical computation of GLM and GLMM, e.g., quasi-likelihood and extended quasi-likelihood. Many others are used to show the linkage of conventional frequentist estimation and Bayesian inference by the following other terms such as joint likelihood, extended likelihood, and adjusted profile likelihood. We demonstrate whether h-likelihood was an essential likelihood that marginal and REML probabilities and statistical probabilities are extracted. The extended probability theory underlies the h-likelihood system and demonstrates how it holds from classical and Bayesian probability.

Generalizations on random effects are of great application in simulations. For example, a typical example is that there are frequent observations of hospital admissions by patients and that the life of these patients can be expected. This might include a survival experiment with unexpected results for patients, and the variance of the estimates indicates the variability of the random effect.

During the first few examples, we demonstrate experiments using normal, log-normal, gamma, Poisson, and binomial HGLMs. Binary models are used to compare with application areas, while the dhglm package is fast and yields consistent results. Descriptions using HGLMs, including organized dispersion, are given below. We also line up models including correlated random effects and structural equation models.

The likelihood implies that probability models will offer an effective way to interpret the data if the model is accurate. It is also necessary to validate the model to verify the interpretation of the results. That being said, it could be hard to ascertain all the model assumptions. During the simulation using h-likelihood, SEM's normal assumption in binary GLMMs can give serious biases if the normal assumption on random effects is incorrect.

Conclusion and future research

The likelihood inferences for specific models may be susceptible to data leakage outliers. If the data size is limited, we can review the data carefully to detect outliers, but it can be difficult for large-scale data to identify outliers or degraded data. A commonly cited drawback of the probability approach is that it is not resilient to model distribution predictions or the existence of outliers or data degradation. It is advantageous to build models that are likely to have stable inferences against such violations. That's also feasible by believing that the model encompasses a wide variety of distributions. We are leaving future studies to combine h-likelihood in the deep learning [39, 40, 55–59], and using this framework towards spatial and remote sensing [60–64], hybrid forecasting [65–80], and more advanced disease detection cases using image detection [81–90].

Appendix A

H-likelihood theory for the frailty model

The h-likelihood gives a straightforward way of handling non-parametric baseline hazards. The h-likelihood under the frailty model is defined by:

$$h = h(\beta, \lambda_0, \alpha) = l_0 + l_1$$

Here,

$$l_0 = \sum\nolimits_{ij} \log f\left(y_{ij}, \delta_{ij} | u_i;\, \beta, \lambda_0\right) = \sum\nolimits_{ij} \delta_{ij} \left\{\log\left(\lambda_0 y_{ij}\right) + \eta_{ij}\right\} - \sum\nolimits_{ij} \wedge_0\left(y_{ij}\right) \exp\left(\eta_{ij}\right)$$

$$l_1 = \sum\nolimits_{ij} \log f\left(v_i;\, \alpha\right)$$

The functional form of $\lambda_0(t)$ is unknown. Hence, we consider $\wedge_0$(t) to be a step function with jumps at the observed event time/At the moment, $y_{(k)}$ is the k-th smallest distinct event time among the y_{ij}'s, and $\lambda_{0k} = \lambda_0\left(y_{(k)}\right)$. Thus, we proposed the use of the profile h-likelihood with λ_0 eliminated, $r^* = h|_{\lambda_0 = \widehat{\lambda}_0}$, given by

$$r^* = r^*(\beta, \alpha) = l_0^* + l_1.$$

Here,

$$l_0^* = \sum_{ij} \log f^*(y_{ij}, \delta_{ij} | u_i;\, \beta) = \sum_{ij} \log f(y_{ij}, \delta_{ij} | u_i;\, \beta, \widehat{\lambda}_0),$$

where,

$$\widehat{\lambda}_{0k}(\beta, v) = \frac{d_{(k)}}{\sum_{ij \epsilon R_k} \exp\left(\eta_{ij}\right)}$$

are solutions of the estimating equations, $\frac{\partial h}{\partial \lambda_{0k}} = 0$. However, $d_{(k)}$ is the number of events at $y_{(k)}$ and $R_{(k)} = \left\{\left(i, j\right) : y_{ij} \geq y_{(k)}\right\}$ is the risk set at $y_{(k)}$. In consequences, we proposed h-likelihood, called penalized partial likelihood (PPL) h_p, given by:

$$h_p(\beta, v, \alpha) = \sum\nolimits_{ij} \delta_{ij} \eta_{ij} - \sum\nolimits_k d_k \log\left\{\sum\nolimits_{ij \epsilon R_{(k)}} \exp\left(\eta_{ij}\right)\right\} + l_1$$

Appendix B

Calculation of scaled deviance test

The scaled deviance is defined by following equation:

$$D\left(y,\widehat{\mu}\right) = -2\left\{l\left(\widehat{\mu},\sigma^2;y|v\right) - l\left(y,\sigma^2;y|v\right)\right\}.$$

Here, the estimated degree of freedom, $d.f = n - tr\left(H^{-1}H^*\right)$, where H and H^* are the Hessian matrices of (β, v) based on l_0 and h, respectively.

Appendix C

Conditional Akaike information criteria towards DHGLM

The conditional Akaike information for double HGLMs is defined as follows:

$$\begin{aligned}
&cAIC = -2E_{f(y,u,a)}E_{f1(y^*|u,a)} \\
&\log g\left\{y^*|\hat{\beta}(y),\hat{v}(y),\hat{\gamma}(y),\hat{b}(y)\right\} \\
&= \int -2\log g\left\{y^*|\hat{\beta}(y),\hat{v}(y),\hat{\gamma}(y),\hat{b}(y)\right\} \\
&f_1\left(y^*|u,a\right)f\left(y,u,a\right)dy^*dyduda
\end{aligned}$$

Here $f\left(y,u,a\right) = f_1\left(y|u,a\right)f_2(u)f_3(a)$ is the true joint distribution of y, u, and$a$. Meanwhile, $\widehat{\beta}\left(y\right)and\widehat{v}\left(y\right)$ are the estimators of fixed and random effects (β, ν) for the mean model, respectively. Here, $\widehat{\gamma}\left(y\right)$ and $\widehat{b}\left(y\right)$ are also the estimators of fixed and random effects(γ, b) for the dispersion model, respectively. At the same time, another two evaluation criteria are mAIC for marginal log-likelihood and rAIC for restricted log-likelihood [5], defined by:

$$mAIC = -2\log m + 2df_m \approx -2\log p_v(h) + 2df_m,$$

$$rAIC = -2\log r + 2df_r \approx -2\log p_{\beta,v}(h) + 2df_r,$$

Here df_m is the number of fixed parameters and df_r is the number of dispersion parameters. When we compare models with different fixed parameters, mAIC can be used, whereas rAIC can be used for dispersion parameter model selection.

Abbreviations

AAA: Abdominal aortic aneurysm; ANOVA: Analysis of variance; ARCH: Autoregressive conditional heteroskedasticity; AUC: Area under curve; cAIC: Conditional akaike information criterion; CGD: Chronic granulomatous disease; DHGLM: Double hierarchical generalized linear models; EG: Ethylene glycol; EVAR: Aortic anatomy on endovascular aneurysm repair; frailtyHL: Frailty models via hierarchical likelihood; GLM: Generalized linear model; HGLM: Hierarchical generalized linear models; IWLS: Iterated weighted least squares; mAIC: Marginal Akaike information criterion; mdhglm: Multivariate double hierarchical generalized linear models; ML: Maximum likelihood; rAIC: Restricted Akaike information criterion; REML: Restricted maximum likelihood; rIFN-g: Randomized trial of gamma interferon; SEM: Structural equation model; SV: Stochastic volatility; SVM: Support vector machine.

Author contributions

Conceptualization: REC, YL, JH, HL, MN, IDH. Methodology: REC, YL, JH, HL, MN, IDH. Project Administration: REC, YL, MN, IDH, BP. Software: REC, YL, JH, HL, MN, IDH, PUG. Validation: REC, YL, JH, HL, MN, IDH. Visualization: REC, YL, JH, HL, MN, IDH. Writing—original draft, review and editing: REC, YL, JH, HL, MN, IDH, PUG, and BP. All authors read and approved the final manuscript.

Funding

National Research Foundation of Korea fully supports this work under Grants [NRF-2019R1A2C1002408].

Data availability

The analysis codes and datasets used in this paper are available from the corresponding author upon reasonable request. Also, the reader can reach the Albatross Analytics website http://cheoling.snu.ac.kr:3838/DHGLM/ to perform graphical and statistical analysis.

Declarations

Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Author details

[1]Research Center for Data and Information Sciences, Research Organization for Electronics and Informatics, National Research and Innovation Agency, Bandung, West Java 40135, Indonesia. [2]Lab Hierarchical Likelihood, Department of Statistics, College of Natural Science, Seoul National University, 56-1 Mountain, Sillim-dong, Gwanak-gu, Seoul, Republic of Korea. [3]Department of Statistics, Pukyong National University, Nam-gu, Yongso-ro, Busan, Republic of Korea. [4]Department of Mathematics, Universitas Sumatera Utara, Medan 20155, Indonesia. [5]Bioinformatics and Data Science Research Center, Bina Nusantara University, Jakarta 11480, Indonesia. [6]Computer Science Department, BINUS Graduate Program—Master of Computer Science Program, Bina Nusantara University, DKI Jakarta 11480, Indonesia.

Received: 9 November 2021 Accepted: 15 May 2022
Published online: 26 May 2022

References

1. Lee Y, Rönnegård L, Noh M. Data analysis using hierarchical generalized linear models with R. 1st ed. Florida: Routledge; 2017.
2. R Development Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2011.
3. R Core Team. R software. Vienna: R Foundation for Statistical Computing; 2008. p. 409. https://doi.org/10.1007/978-3-540-74686-7.
4. Caraka RE, Lee Y, Chen RC, Toharudin T. Using hierarchical likelihood towards support vector machine: theory and its application. IEEE Access. 2020;8:194795–807.
5. Lee Y, Rnnegrd L, Noh M. Double HGLMs using the dhglm package. In: Noh M, editor. Data analysis using hierarchical generalized linear models with R. London: Chapman and Hall; 2017.
6. Felleki M, Lee D, Lee Y, Gilmour AR, Rönnegård L. Estimation of breeding values for mean and dispersion, their variance and correlation using double hierarchical generalized linear models. Genet Res. 2012;94:307–17.
7. Ha ID, Noh M, Lee Y. FrailtyHL: a package for fitting frailty models with h-likelihood. R J. 2012;4:28–37.
8. Lee Y, Noh M. Modelling random effect variance with double hierarchical generalized linear models. Stat Model. 2012;12:487–502.
9. Collignon O, Han J, An H, Oh S, Lee Y. Comparison of the modified unbounded penalty and the LASSO to select predictive genes of response to chemotherapy in breast cancer. PLoS ONE. 2018;13:15.
10. Noh M, Lee Y, Oud JHL, Toharudin T. Hierarchical likelihood approach to non-Gaussian factor analysis. J Stat Comput Simul. 2019;89:1555–73.
11. Jin S, Noh M, Lee Y. H-likelihood approach to factor analysis for ordinal data. Struct Equ Model. 2018;25:530–40.
12. Ha ID, Lee Y. A review of h-likelihood for survival analysis. Jpn J Stat Data Sci. 2021. https://doi.org/10.1007/s42081-021-00125-z.
13. Ha ID, Noh M, Lee Y. H-likelihood approach for joint modeling of longitudinal outcomes and time-to-event data. Biom J. 2017;59:1122–43.
14. Lee D, Lee Y. Extended likelihood approach to multiple testing with directional error control under a hidden Markov random field model. J Multivar Anal. 2016;151:1–13.
15. Lee W, Ha ID, Noh M, Lee D, Lee Y. A review on recent advances and applications of h-likelihood method. J Korean Stat Soc. 2021. https://doi.org/10.1007/s42952-021-00130-8.
16. Jin S, Lee Y. A review of h-likelihood and hierarchical generalized linear model. WIREs Comp Stat. 2020. https://doi.org/10.1002/wics.1527.
17. Caraka RE, Noh M, Chen RC, Lee Y, Gio PU, Pardamean B. Connecting climate and communicable disease to penta helix using hierarchical likelihood structural equation modelling. Symmetry. 2021;13:1–21.
18. Ha ID, Jeong J-H, Lee Y. Statistical modelling of survival data with random effects. Berlin: Springer; 2017.
19. Ha ID, Xiang L, Peng M, Jeong JH, Lee Y. Frailty modelling approaches for semi-competing risks data. Lifetime Data Anal. 2020;26:109–33.
20. Huang R, Xiang L, Ha ID. Frailty proportional mean residual life regression for clustered survival data: a hierarchical quasi-likelihood method. Stat Med. 2019;38:4854–70.
21. Ha ID, Kim JM, Emura T. Profile likelihood approaches for semiparametric copula and frailty models for clustered survival data. J Appl Stat. 2019;46:2553–71.
22. Taleb I, Serhani MA, Bouhaddioui C, Dssouli R. Big data quality framework: a holistic approach to continuous quality management. J Big Data. 2021. https://doi.org/10.1186/s40537-021-00468-0.
23. Shabbir MQ, Gardezi SBW. Application of big data analytics and organizational performance: the mediating role of knowledge management practices. J Big Data. 2020. https://doi.org/10.1186/s40537-020-00317-6.

24. Hu KH, Hsu MF, Chen FH, Liu MZ. Identifying the key factors of subsidiary supervision and management using an innovative hybrid architecture in a big data environment. Financ Innov. 2021. https://doi.org/10.1186/s40854-020-00219-9.
25. Shah SIH, Peristeras V, Magnisalis I. DaLiF: a data lifecycle framework for data-driven governments. J Big Data. 2021. https://doi.org/10.1186/s40537-021-00481-3.
26. Caraka RE, Chen RC, Huang SW, Chiou SY, Gio PU, Pardamean B. Big data ordination towards intensive care event count cases using fast computing GLLVMS. BMC Med Res Methodol. 2022. https://doi.org/10.1186/s12874-022-01538-4.
27. Daki H, El Hannani A, Aqqal A, Haidine A, Dahbi A. Big Data management in smart grid: concepts, requirements and implementation. J Big Data. 2017. https://doi.org/10.1186/s40537-017-0070-y.
28. Colombo P, Ferrari E. Access control technologies for big data management systems: literature review and future trends. Cybersecurity. 2019. https://doi.org/10.1186/s42400-018-0020-9.
29. Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: management, analysis and future prospects. J Big Data. 2019. https://doi.org/10.1186/s40537-019-0217-0.
30. Lee Y, Nelder J. Modelling and analysing correlated non-normal data. Stat Model. 2001;1:3–16.
31. Lee D, Lee W, Lee Y, Pawitan Y. Sparse partial least-squares regression and its applications to high-throughput data analysis. Chemom Intell Lab Syst. 2011;109:1–8.
32. Lee Y, Nelder JA. Hierarchical generalized linear models. J Royal Stat Soc Ser B. 1996. https://doi.org/10.1111/j.2517-6161.1996.tb02105.x.
33. Hudak Jr SJ, Saxena A, Bucci RJ, Malcolm RC. Development of standard methods of testing and analyzing fatigue crack growth rate data. 1978.
34. Lu CJ, Meeker WO. Using degradation measures to estimate a time-to-failure distribution. Technometrics. 1993;35:161–74.
35. Lee Y, Molas M, Noh M. mdhglm: multivariate double hierarchical generalized linear models. 2018.
36. Price CJ, Kimmel CA, Tyl RW, Marr MC. The developmental toxicity of ethylene glycol in rats and mice. Toxicol Appl Pharmacol. 1985;81:113–27.
37. Ha ID, Christian NJ, Jeong JH, Park J, Lee Y. Analysis of clustered competing risks data using subdistribution hazard models with multivariate frailties. Stat Methods Med Res. 2016;25:2488–505.
38. Ha ID, Lee Y, Song JK. Hierarchical-likelihood approach for mixed linear models with censored data. Lifetime Data Anal. 2002;8:163–76.
39. Hao L, Kim J, Kwon S, do Ha I. Deep learning-based survival analysis for high-dimensional survival data. Mathematics. 2021;9:1–18.
40. Kim JM, do Ha I. Deep learning-based residual control chart for binary response. Symmetry. 2021;13:1–15.
41. Ha ID, Youngjo L. Multilevel mixed linear models for survival data. Lifetime Data Anal. 2005;11:131–42.
42. Lee Y, Ha ID. Orthodox BLUP versus h-likelihood methods for inferences about random effects in Tweedie mixed models. Stat Comput. 2010;20:295–303.
43. Ha ID, Lee Y. Estimating frailty models via poisson hierarchical generalized linear models. J Comput Graph Stat. 2003. https://doi.org/10.1198/1061860032256.
44. Crowder M, Fleming TR, Harrington DP. Counting processes and survival analysis. J Royal Stat Soc Ser A. 1994. https://doi.org/10.2307/2983370.
45. Fleming TR, Lin DY. Survival analysis in clinical trials: past developments and future directions. Biometrics. 2000. https://doi.org/10.1111/j.0006-341X.2000.0971.x.
46. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20:273–97.
47. Drucker H, Burges CJC, Kaufman L, Smola A, Vapnik V. Support vector regression machines. Adv Neural Inf Process Syst. 1996;9:155–61.
48. Williams CKI. Learning with kernels: support vector machines, regularization, optimization, and beyond. J Am Stat Assoc. 2003. https://doi.org/10.1198/jasa.2003.s269.
49. Fradkin D, Muchnik I. Support vector machines for classification. DIMACS series in discrete mathematics and theoretical computer science. Citeseer. 2006;70:13–20.
50. Schölkopf B. The kernel trick for distances. Adv Neural Inform Process Syst. 2001;13:301–7.
51. Wang J, Lee J, Zhang C. Kernel trick embedded Gaussian mixture model. Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics). 2003;2842:159–74.
52. Hofmann M. Support vector machines—kernels and the kernel trick. Universität Bamberg. 2006;26(3):1–16. http://www.cogsys.wiai.uni-bamberg.de/teachingarchive/ss06/hs_svm/slides/SVM_Seminarbericht_Hofmann.pdf.
53. Caraka RE, Nugroho NT, Tai S-K, Chen RC, Toni T, Bens P. Feature importance of the aortic anatomy on endovascular aneurysm repair (EVAR) using Boruta and Bayesian MCMC. Commun Math Biol Neurosci 2020.
54. Caraka RE, Noh M, Lee Y. Package 'hsem'. R project; 2021. p. 1–7.
55. Moutarde F. Deep-learning: general principles + convolutional neural networks. 2018
56. Czum JM. Dive into deep learning. J Am Coll Radiol. 2020. https://doi.org/10.1016/j.jacr.2020.02.005.
57. Wilson AG, Hu Z, Salakhutdinov R, Xing EP. Deep Kernel learning. Artificial intelligence and statistics (AISTATS). 2016;370-378. http://arxiv.org/abs/1511.02222.
58. Benuwa BB, Zhan YZ, Ghansah B, Wornyo DK, Banaseka KF. A review of deep machine learning. Int J Eng Res Africa. 2016;24:124–36.
59. Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw. 2015;61:85–117.
60. Sakti AD, Rinasti AN, Agustina E, Diastomo H, Muhammad F, Anna Z, et al. Multi-scenario model of plastic waste accumulation potential in indonesia using integrated remote sensing, statistic and socio-demographic data. ISPRS Int J Geo-Inform. 2021. https://doi.org/10.3390/ijgi10070481.
61. Syahid LN, Sakti AD, Virtriana R, Wikantika K, Windupranata W, Tsuyuki S, et al. Determining optimal location for mangrove planting using remote sensing and climate model projection in southeast Asia. Remote Sens. 2020;12:1–29.

62. Sakti AD, Fauzi AI, Takeuchi W, Pradhan B, Yarime M, Vega-Garcia C, et al. Spatial prioritization for wildfire mitigation by integrating heterogeneous spatial data: a new multi-dimensional approach for tropical rainforests. Remote Sens. 2022;14:543.
63. Sakti AD, Fauzi AI, Wilwatikta FN, Rajagukguk YS, Sudhana SA, Yayusman LF, et al. Multi-source remote sensing data product analysis: investigating anthropogenic and naturogenic impacts on mangroves in southeast asia. Remote Sens. 2020;12:1–29.
64. Sakti AD, Rahadianto MAE, Pradhan B, Muhammad HN, Andani IGA, Sarli PW, et al. School location analysis by integrating the accessibility, natural and biological hazards to support equal access to education. ISPRS Int J Geo-Inform. 2022. https://doi.org/10.3390/ijgi11010012.
65. Hippert HS, Bunn DW, Souza RC. Large neural networks for electricity load forecasting: are they overfitted? Int J Forecast. 2005;21:425–34.
66. Zhang G, Patuwo BE, Hu MY. Forecasting with artificial neural networks: the state of the art. Int J Forecast. 1998;14:35–62.
67. Leung MT, Chen A-S, Daouk H. Forecasting exchange rates using general regression neural networks. Comput Oper Res. 2000;27:1093–110.
68. Herliansyah R, Jamilatuzzahro J. Feed forward neural networks for forecasting indonesia exchange composite index. GSTF J Math Stat Oper Res. 2017. https://doi.org/10.5176/2251-3388-4.1.77.
69. Toharudin T, Pontoh RS, Caraka RE, Zahroh S, Lee Y, Chen RC. Employing long short-term memory and facebook prophet model in air temperature forecasting. Commun Stat Simulat Comput. 2021;early acces:1–12.
70. Pontoh RS, Solichatus Z, Hidayat Y, Aldella R, Jiwani NM, Sukono. Covid-19 modelling in south korea using a time series approach. Int J Adv Sci Technol. 2020;29:1620–32.
71. Lee Y, Nelder JA, Noh M. H-likelihood: problems and solutions. Stat Comput. 2007;17:49–55.
72. Livieris IE, Pintelas E, Pintelas P. A CNN–LSTM model for gold price time-series forecasting. Neural Comput Appl. 2020;32:17351–60. https://doi.org/10.1007/s00521-020-04867-x.
73. Brodersen KH, Gallusser F, Koehler J, Remy N, Scott SL. Inferring causal impact using bayesian structural time-series models. Ann Appl Stat. 2015;9:247–74.
74. Khodabakhsh AA. Forecasting multivariate time-series data using LSTM and mini-batches in the 7th international conference on contemporary issues in data science. Cham: Springer; 2019. p. 121–9.
75. Makridakis S, Spiliotis E, Assimakopoulos V. M5 accuracy competition: results, findings, and conclusions. Int J Forecast. 2022. https://doi.org/10.1016/j.ijforecast.2021.11.013.
76. Makridakis S, Hibon M. The M3-competition: results, conclusions and implications. Int J Forecast. 2000. https://doi.org/10.1016/S0169-2070(00)00057-1.
77. Theodorou E, Wang S, Kang Y, Spiliotis E, Makridakis S, Assimakopoulos V. Exploring the representativeness of the M5 competition data. Int J Forecast. 2021. https://doi.org/10.1016/j.ijforecast.2021.07.006.
78. Makridakis S, Spiliotis E, Assimakopoulos V. The M4 Competition: 100,000 time series and 61 forecasting methods. Int J Forecast. 2020. https://doi.org/10.1016/j.ijforecast.2019.04.014.
79. Caraka RE, Chen RC, Yasin H, Pardamean B, Toharudin T, Wu SH. Prediction of status particulate matter 25 using state Markov chain stochastic process and HYBRID VAR-NN-PSO. IEEE Access. 2019;7:161654–65.
80. Caraka RE, Chen RC, Yasin H, Lee Y, Pardamean B. Hybrid vector autoregression feedforward neural network with genetic algorithm model for forecasting space-time pollution data. Indonesian J Sci Technol. 2021;6:243–66.
81. Aswale VA, Shaikh JA. Detection of microaneurysm in fundus retinal images using SVM classifier. IJEDR. 2017;5:175–80.
82. Pardamean B, Cenggoro TW, Rahutomo R, Budiarto A, Karuppiah EK. Transfer learning from chest X-ray pre-trained convolutional neural network for learning mammogram data. Proc Comput Sci. 2018;135:400–7. https://doi.org/10.1016/j.procs.2018.08.190.
83. Novitasari DCR, Hendradi R, Caraka RE, Rachmawati Y, Fanani NZ, Syarifudin A, et al. Detection of covid-19 chest X-ray using support vector machine and convolutional neural network. Commun Math Biol Neurosci. 2020.
84. Whi W, Ha S, Kang H, Lee DS. Hyperbolic disc embedding of functional human brain connectomes using resting state fMRI. bioRxiv. 2021. https://doi.org/10.1101/2021.03.25.436730.
85. Lee D, Kang H, Kim E, Lee H, Kim H, Kim YK, et al. Optimal likelihood-ratio multiple testing with application to Alzheimer's disease and questionable dementia data analysis, statistics and modelling. BMC Med Res Methodol. 2015;15:1–11.
86. Kim JY, Oh D, Sung K, Choi H, Paeng JC, Cheon GJ, et al. Visual interpretation of [18F]Florbetaben PET supported by deep learning-based estimation of amyloid burden. Eur J Nucl Med Mol Imag. 2021;48:1116–23.
87. Choi H, Ha S, Kang H, Lee H, Lee DS. Deep learning only by normal brain PET identify unheralded brain anomalies. EBioMedicine. 2019;43:447–53. https://doi.org/10.1016/j.ebiom.2019.04.022.
88. Whi W, Park JY, Choi H, Paeng JC, Cheon GJ, Kang KW, et al. Predicting outcome of repair of medial meniscus posterior root tear with early osteoarthritis using bone single-photon emission computed tomography/computed tomography. Medicine. 2020;99: e21047.
89. Bae S, Choi H, Whi W, Paeng JC, Cheon GJ, Kang KW, et al. Spatial normalization using early-phase [18F]FP-CIT PET for quantification of striatal dopamine transporter binding. Nucl Med Mol Imag. 2020;54:305–14.
90. Whi W, Huh Y, Ha S, Lee H, Kang H, Lee DS. Characteristic functional cores revealed by hyperbolic disc embedding and k-core percolation on resting-state fMRI. Sci Rep. 2022. https://doi.org/10.1038/s41598-022-08975-7.

Publisher's Note