Article

Not peer-reviewed version

# Feature Selection by Mutual Information

Philip E. Cheng * , Juin-Der Lee , Alexander N. Savostyanov , Sheng-Kai Lee , Michelle Liou *

Posted Date: 23 July 2025

doi: 10.20944/preprints202507.1954.v1

Keywords: data log-likelihood; maximum entropy; maximum likelihood; mutual information; Pythagorean law

*Article*

# Feature Selection by Mutual Information

**Philip E. Cheng [1], Juin-Der Lee [1,2], Alexander N. Savostyanov [3], Sheng-Kai Lee [1] and Michelle Liou [1,*]**

[1] Institute of Statistical Science, Academia Sinica, Taipei
[2] GoGoX, Hong Kong
[3] Laboratory of Psychological Genetics, Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
* Correspondence: mliou@stat.sinica.edu.tw

**Abstract**

Mutual information (MI), a crucial component in statistical inference and an essential tool for data analysis, has been largely overlooked for seven decades in the statistical literature. Emerging from the analysis of data information within the realms of biological, engineering and physical sciences, essential working MI formulas have been involved with asymmetric expressions of terms found in both MI and Shannon entropy, consequently leading to a reduction in effective statistical inference. The innovative observation of the equivalence among the three principles: maximum entropy, maximum likelihood, and minimum MI, has offered new insights into the geometry of data likelihood and established a new framework for statistical inference by Cheng et al. (2008, 2010). Advanced data analysis, in contrast to the existing methods, is established based on the MI identities and the fundamental Pythagorean law of conditional MI. This article presents the new methodology by elaborating its effective applications to feature selection in genetics for predicting patients with depressive disorders.

**Keywords:** data log-likelihood; maximum entropy; maximum likelihood; mutual information; Pythagorean law

## 1. Introduction

In the exploration of the mathematical theory surrounding discrete message communication, Shannon (1948) investigated the uncertainty associated with information transmission by utilizing a measured quantity known as "entropy," which is grounded in the principles of statistical mechanics. Kullback and Leibler (1951) introduced the term "mutual information (MI)," which represents the "difference" between the sum of the component entropies of two random variables and their joint entropy; it is also referred to as relative entropy or KL-divergence between the two variables, involving either discrete or continuous distributions. When the definition of MI is expanded to encompass a vector of three variables, McGill (1954) substituted the original MI (a symmetric difference) with an asymmetric difference of entropy terms, which he termed "transmitted information (TI)", also known as "interaction information". The former designation will be employed in this article, because the term "interaction" among three variables has been recognized as a classic symmetric measure in statistics since the 1930s. Upon breaching the conventional set-theory presentation of the Venn diagram of the log-likelihoods of three variables, an asymmetric TI can be negative-valued, whereas the symmetric MI is always non-negative. When dealing with four or more variables, TI can present multiple formulas for the same quantity. This complicates effective statistical inference without a thorough theoretical examination in the existing literature. Over the past seventy years, TI has been extensively analyzed in numerous articles that cover both empirical and theoretical research across a wide range of scientific fields, including biology, computer science, engineering, econometrics, medicine, physics, and psychology. Notably, it has frequently been utilized in the investigation of feature selection and network inference, as evidenced by works such as McMahon et al. (2014), Timme et al. (2014), Vergara and Estévez (2014), Chan et al. (2017), and Ince et al. (2017), among others. However, very few studies of TI, built upon probability distributions, have ever been scrutinized through the lens of statistical inference (p-values) alongside data analysis in the literature.

On the other hand, it can be shown that the original MI in two variables, along with its direct extension in the symmetric version to cases of three or more variables, presents three significant advancements in the realm of general statistical inference. This is particularly relevant for discrete multivariate data analyzed through multiway contingency tables. First, the invariant Pythagorean law (IPL), which was established based on two-variable MI, serves to unify the power analysis involved in testing both independence and alternative hypotheses, without the need of employing non-central Chi-square distributions as recommended in the classic inference (Cheng et al, 2008). With three variables, the IPL delineates the intrinsic two-step likelihood ratio (LR) test, which enhances the power analysis for assessing conditional independence—specifically, the relationship between two variables while controlling for a third variable—by logically partitioning the process into testing interaction and partial association (Cheng et al, 2010). Second, the symmetric MI measure of a random vector directly produces the intrinsic formulas of information-equivalent MI identities based on the multivariate log-likelihood. These identities are inherently valuable, as each MI identity constitutes an orthogonal decomposition of the LR deviance, which is utilized for testing mutual independence among the components of the vector. This revises the traditional non-orthogonal decompositions of the corresponding Chi-square test statistic and LR deviance within the same hypothesis testing framework, as noted by Bishop et al. (1975).

Lastly, a crucial aspect in the foundation of statistical inference is observed with the definition of MI. It is the equivalence among the three principles: maximum likelihood, minimum MI, and maximum entropy. It is recognized that the equivalence between the latter two principles directly stems from the MI definition, while the equivalence between the first two principles is elucidated in the definition of the LR test for independence in two- and three-way contingency tables, as explained in Cheng et al. (2006, 2008, 2010). Indeed, this insight provides fresh perspectives on the geometry of data likelihood information, which serves as a fundamental element of statistical inference ever examined by Eguchi and Copas (2006). Recognizing the possible inconvenient negative TI values, Amari (2001, Section IV) explained the difficulty in extending the MI formulas of discrete random variables to the case of continuous distributions for valid analysis in Riemannian manifolds. In the case of $n$-dimensional discrete or continuous multivariate distributions, he showed that orthogonal MI decompositions hold with $2^n - 1$ components of log-liner models, based on the $e$-flat coordinate system of an exponential family manifold. However, it is also understood that, in the Euclidean geometry, the orthogonality between those $2^n - 1$ components of a log-linear model can be replaced with a more concise and convenient equation of orthogonal MI decomposition, when $n$ is greater than three (Liou et al., 2023). This particularly important observation will be illustrated in performing the task of feature selection in this article.

It is observed that there are benefits by implementing the new geometry of MI identity in the inference of LR statistics, especially in the context of discrete multivariate analysis. It is advised to employ the orthogonal MI decomposition of data log-likelihood in the analysis of feature selection and model selection through logistic (logit) models with discrete (discretized) multivariate data. Specifically, the MI orthogonal decomposition along with the inherent IPL can be utilized to reframe the conventional discrete multivariate data analysis. This approach facilitates the construction of a valid logistic regression model through preliminary feature selection, followed by evaluating an appropriate MI identity to achieve the selection of a parsimonious regression model.

In view of the aforementioned illustrations, it is evident that there exists an implicit relationship between the selection of features in loglinear or logistic models and the examination of graphical or network models within the field of computer sciences. In applications, there are instances where specific covariates or features may exhibit significant interaction effects with the response variable. It is therefore important that MI analysis can assist in verifying proper conditions for valid inference with regard to the interaction effects involving discrete or discretized features. The current study will demonstrate the novel feature selection algorithm through a real example, which encompasses single nucleotide polymorphism (SNP) genotyping data from 165 pairs of alleles aimed at predicting patients who have been clinically diagnosed with depressive disorders.

## 2. Methodology

In the examination of generalized linear models (GLM), a variety of techniques for feature selection have been investigated, primarily focusing on the identification of powerful features or predictors in models involving only main effects. These traditional approaches encompass ridge regression, the Cp statistic, the Akaike Information Criterion (AIC), and the Bayesian Information Criterion (BIC), all of which were established in the 1970s. Subsequently, discussions on stepwise variable selection procedures emerged to improve the implementation of these methods. Within the broad field of applied sciences, most research has primarily concentrated on the identification of main-effect models or the selection of features to facilitate simplified data interpretation. In essence, conventional studies have largely emphasized the selection of features or predictors, as well as various dimension reduction techniques, while potential interaction effects between the features (or predictors) and the response variable have often been overlooked. The significant literature addressing this concern includes contributions from Linhart and Zucchini (1986), Fahrmeir and Tutz (1994), Breiman (1995), and Burnham and Anderson (2010), among others.

In the 21st century, a new range of feature selection techniques emerged. Researchers encountered the intricate challenge of analyzing data characterized by high-dimensional parameters, primarily consisting of non-informative attributes such as brain images and genetic data, along with limited sample sizes from patients in biological and medical settings. Meanwhile, a variety of techniques for feature selection have been investigated in the context of high-dimensional biological, economic, and genetic datasets characterized by relatively small sample sizes. This situation prompted the search for optimal feature selection criteria that utilize various penalty functions. Notable methods developed during this time include Lasso, adaptive Lasso, elastic Net, Scad, and the Dantzig selector, which are rooted in the foundational work of Tibshirani in 1996 and subsequent studies. Each of these variable selection criteria generally aims to identify a uniquely optimal set of predictors customized to the observed data, regarded as the most effective main-effect GLM in regression analysis. In scenarios involving high-dimensional data with small sample sizes under specified likelihood distributions, the Lasso-type methods have been popular in both empirical and theoretical research by yielding various dimension reduction results. However, these methods usually overlooked the inspection of potential significant interaction effects between the selected predictors and the response variable, because interactions are difficult to identify or assess through $p$-value inference by these methods. In contrast to the challenges posed by collinearity and interaction effects in linear regression analysis, it is essential, adhering to the principle of parsimony, to identify the minimal sets of predictors that result in the least number of significant interaction effects based on specific information criteria. In this regard, the proposed MI analysis can be validated as one of the most suitable candidate methods.

In the present research, it is advised to employ the orthogonal MI decomposition of data log-likelihood for the selection of variables and models when addressing discrete or partially discretized multivariate data. Initially, the MI variable selection technique is utilized to pinpoint the key predictors, followed by an examination of appropriate MI identities for effective model selection. Within the framework of linear regression models that utilize multivariate normal distributions, the MI and CMI components can be expressed through logarithmic ratios of “determinants of variances’” and logarithmic functions of “partial correlations” (cf. Whittaker, 1990, Chapter 6). Thus far, in applying MI analysis to datasets that include both continuous and discrete predictors, it is evident that challenges in the examination of MI and CMI terms continue, particularly in the attainment of valid $p$-value inference (Cheng & Liou, 2024). A provisional solution may employ appropriate discretization of numerical variables, allowing direct application of MI analysis to the selection of a valid model with precise data interpretation. Further investigations into discrete MI analysis that integrates both discrete and numerical variables are yet to be conducted.

Consider the selection of variables and construction of models through introducing the MI analysis. Let $\{X_1, \cdots, X_m, Y\}$ denote the set of data variables, where $Y$ is assumed to be a discrete response variable (or target) of interest and $\{X_1, \cdots, X_m\}$ denote the available set of $m$ discrete (discretized) features or predictors of $Y$. A general form of the MI identity can be expressed as $I(X_1, \cdots, X_m, Y) = I(X_1, \cdots, X_m) + I((X_1, \cdots, X_m), Y)$. An interested reader may consult Eqn. (9) in the work of Cheng and Liou (2024) for details regarding the information decomposition rule. Denote this general form as $I(X_1, \cdots, X_m, Y) = I(\mathbf{X}) + I(\mathbf{X}, Y)$ and $\mathbf{X} \equiv (X_1, \cdots, X_m)$. The details of the MI identity is

$$\begin{aligned} I(\mathbf{X}, Y) &= I(X_1, Y) + I(X_2, Y|X_1) + \ldots + I(X_m, Y|(X_1, \ldots, X_{m-1})) \\ &= I(X_1, Y) + Int(X_2, Y, X_1) + Par(X_2, Y|X_1) + \cdots \\ &\quad + Int(X_t, Y, (X_1, \ldots, X_{t-1})) + Par(X_t, Y|(X_1, \ldots, X_{t-1})) + \cdots \\ &\quad + Int(X_m, Y, (X_1, \ldots, X_{m-1})) + Par(X_m, Y|(X_1, \ldots, X_{m-1})), \quad (2.1) \end{aligned}$$

which expresses the entire association effects between $Y$ and $\mathbf{X}$.

### 2.1. Forward Feature Selection

A stepwise forward variable selection procedure is defined using the MI ratio (MIR), which is the ratio of an MI or CMI estimate to its effective *df*, given that all empty rows and columns have been deleted from the data.

**Step 1: Select significant features**. For simplicity, assume that the final selected feature (predictor) set of $k$ predictors is denoted by $\{\mathrm{X}_{(1)}, \cdots, \mathrm{X}_{(k)}\}$, $k \leq m$. Denote the first selected feature by $X_{(1)}$, which yields the maximum MIR estimate or the least $p$-value (the most significant predictor for the target $Y$) among all candidate features in $\mathbf{X}$. That is, $X_{(1)}$ is chosen to satisfy that

$$MIR_1 = \frac{\hat{I}\left(X_{(1)}, Y\right)}{df\left[\hat{I}\left(X_{(1)}, Y\right)\right]} = max_{X_i \in \boldsymbol{X}} \left\{ \frac{\hat{I}(X_i, Y)}{df\left[\hat{I}(X_i, Y)\right]} \right\}, (2.2)$$

where

$$p\text{-value of } \{\hat{I}\left(X_{(1)}, Y\right)\} = min_{X_i \in \boldsymbol{X}} \{p \ value \ of \ \hat{I}(X_i, Y)\}. (2.3)$$

Here, each sample MI estimate $\hat{I}(X_i, Y)$ approximates the Chi-square distribution with degree of freedom $df[\hat{I}(X_i, Y)]$, assuming the sample size is sufficiently large. It is observed that (2.2) does not necessarily imply (2.3); however, this implication holds true when the sample size is adequately large. Let $\mathbf{X}^{(\mathbf{0})} (= \emptyset)$ denote the empty set. For $t \geq 1$, let $\mathbf{X}^{(t)} = \{\mathrm{X}_{(1)}, \cdots, \mathrm{X}_{(t)}\}$ denote the set of selected predictors at the $t^{th}$ stage. The procedure selects a new feature $X_{(t+1)}$ which yields the greatest MI ratio estimate

$$MIR_{t+1} = max_{X_i \in \boldsymbol{X} \setminus \boldsymbol{X}^{(t)}} \left\{ \frac{\hat{I}\left(\boldsymbol{X}^{(t)} \cup X_i, Y\right)}{df\left[\left(\boldsymbol{X}^{(t)} \cup X_i, Y\right)\right]} \right\} \equiv \frac{\hat{I}\left(X_{(t+1)}, Y\right)}{df\left[\hat{I}\left(X_{(t+1)}, Y\right)\right]}, (2.4)$$

among the unselected features such that $\mathbf{X}^{(t+1)} = \boldsymbol{X}^{(t)} \cup X_{(t+1)} = \{X_{(1)}, \cdots, X_{(t+1)}\}$. In general, formula (2.4) selects a new feature $X_{(t+1)}$ which gives the most significant CMI estimate $\hat{I}\left(X_{(t+1)}, Y|\boldsymbol{X}^{(t)}\right)$ with the least $p$ value. The forward selection procedure proceeds until no member in $\boldsymbol{X} \setminus \mathbf{X}^{(t)}$ can be selected when $t = k$, and yields the final feature set $\mathbf{X}^{(k)} = \{\mathrm{X}_{(1)}, \cdots, \mathrm{X}_{(k)}\}$.

It is possible that two candidates of a new feature $X_{(t+1)}$ (for some $t \geq 1$) may yield rather close MIR estimates in (2.4) and close $p$ values when the sample size is not sufficiently large. This is the case that two or more sets of competitive features may yield different acceptable main-effect logistic (regression) models.

### 2.2. Backward Feature Deletion

The backward deletion procedure is used to remove dispensable features from a set of selected variables in Step 1.

**Step 2: Delete CMI terms.** Let $\mathbf{X}^{(t)} = \{X_{(1)}, \cdots, X_{(t)}\}$ denote the selected set of features at stage $t$ of Step 1. For $t \geq 1$, suppose that a new feature is selected to yield the set $\mathbf{X}^{(t+1)}$. Now, it is possible to find the particular $X'$ (certain $X_{(i)}$) in $\mathbf{X}^{(t)}$ such that $\hat{I}(X', Y|(\mathbf{X}^{(t+1)} \backslash X'))$ yields the greatest insignificant $p$-value, if it exists, then delete this dispensable feature $X'$. That is, delete $X'$ in $\mathbf{X}^{(t)}$ according to the insignificant estimate:

$$min_{X_j \in \boldsymbol{X}^{(t)}} \left\{ \frac{\hat{I}\left(X_j, Y|(\boldsymbol{X}^{(t+1)} \backslash X_j)\right)}{df[\hat{I}\left(X_j, Y\right)|(\boldsymbol{X}^{(t+1)} \backslash X_j)]} \right\} = \left\{ \frac{\hat{I}\left(X', Y|(\boldsymbol{X}^{(t+1)} \backslash X')\right)}{df[\hat{I}(X', Y)|(\boldsymbol{X}^{(t+1)} \backslash X')]} \right\}. \quad (2.5)$$

Continue the stepwise deletion procedure (2.5) with the selected set $\mathbf{X}^{(k)} = \{X_{(1)}, \cdots, X_{(k)}\}$ of features in Step 1, until it stops and yields the final set of features. Note that the stepwise forward selection (2.4) in Step 1 is processed with the stepwise backward deletion (2.5) simultaneously to accomplish the feature selection procedure.

### *2.3. MI Main-Effect Model Selection*

Finally, the MI model construction procedure is defined as follows.

**Step 3: Delete insignificant interaction terms.** Assume that the final set of features $\{\mathrm{X}_1, \dots, \mathrm{X}_k\}$ (representing $\{X_{(1)}, \cdots, X_{(k)}\}$ for simplicity of notation) is selected after Step 2. Rearrange the final selected features in the MI identity (2.1) where $m$ is replaced by $k$, such that the highest order interaction estimate $\widehat{Int}(\mathrm{X}_k, Y, (\mathrm{X}_1, \dots, \mathrm{X}_{k-1}))$ is the least significant and most likely deleted. The procedure continues with deleting insignificant higher-order interaction terms as much as possible; the remaining variables $\{X_1, \dots, X_{t-1}, X_t\}$are arranged such that $\widehat{Int}(\mathrm{X}_t, Y, (\mathrm{X}_1, \dots, \mathrm{X}_{t-1}))$ is the least significant, for $2 \leq t < k$, with the reduced formula (2.1). Finally, the procedure stops when reaching the last estimate $\widehat{Int}(\mathrm{X}_2, Y, \mathrm{X}_1)$, with the significant $\widehat{Int}(X_1, Y)$.

It is a cautionary remark that Step 3 may be skipped if all significant interaction effects between the selected features (predictors) and the target (response) variable can be ignored, that is, a valid main-effect model is all what is desired in application. Nonetheless, such a practice is in general not recommended.

**Step 4: Model construction.** Use the selected feature set $\mathbf{X}^{(k)} = \{X_{(1)}, \cdots, X_{(k)}\}$ from the final stage of Step 1, to determine, by the results of Step 3, the retained significant interaction and partial association (main) effects in the regression model. This concludes the final model selected by the rearranged MI identity examined in Step 3.

To summarize, the proposed MI feature selection procedure essentially consists of Steps 1 and 2. Once the features (predictors) are selected, Step 3 is used to identify the significant main and indispensable interaction effects such that Step 4 follows to yield the desired regression model. It is important that the analysis of Step 3 must follow the Pythagorean law: the interaction and partial association effects of each CMI estimate were evaluated by dividing the usual test level $\alpha = 0.05$ (or 0.10) into two separate levels $\alpha_1$ and $\alpha_2$, such that the two-step LR test (cf. Cheng *et al*, 2010) is legitimately executed based on the Pythagorean law.

## 3. Empirical Example

ICBrainDB is a database developed between 2013 and 2021, encompassing both healthy individuals and clinical patients, which includes adults and children. Participants were recruited from the regions of Novosibirsk, Yakutia, Tuva, and Mongolia. Additionally, clinical patients suffering from major depression, anxiety, or stress disorders from Novosibirsk were involved in this research. All participants underwent psychological assessments utilizing questionnaires designed to evaluate personality traits, cultural characteristics, and the risks associated with the development of depression and anxiety disorders. The database also includes findings from single nucleotide polymorphism (SNP) analysis across 165 pairs of alleles linked to various neurotransmitter functions, cellular activities, and immune system responses (Ivanov et al., 2019; 2022). The subsequent section outlines the types of data categorized by ethnic and cultural groups. In this research, data from 664 adults (248 males; average age 24.71± 9.63) were selected based on their complete records related to the 165 pairs of alleles.

### *3.1. Ethnic/Culture Groups*

The following ethnic groups were included among the 664 participants: (1) 95 individuals from Barnaul (Altai Krai, Russia): psychological and genetic data derived from non-clinical subjects with an assessment of the propensity for depression; (2) 180 Tuvans (Kyzyl, Tuva): psychological and genetic data obtained from non-clinical subjects with an evaluation of the tendency towards depression; (3) 46 Evenks (Northern region of the Sakha Republic): psychological and genetic data based on non-clinical subjects with an analysis of the inclination to depression; (4) 50 Yakuts (the Sakha Republic, Russia): psychological and genetic data sourced from non-clinical subjects with a review of the likelihood of depression; (5) 293 individuals from Novosibirsk (including Russians, Ukrainians, Jews, Tatars, and Tajiks residing in Siberia, among whom 54 were patients): psychological and genetic data collected from both clinical and non-clinical subjects with an assessment of the tendency for depression (Note: Clinical adults were diagnosed with depression and anxiety disorders).

### *3.2. Psychological Questionnaires*

In the database, all examined subjects provided personal information (gender and age) and completed the 40-item State-Trait Anxiety Inventory (STAI) (Spielberger, 2010) and 21-item Beck Depression Inventory (BDI). The Beck depression inventory (Beck et al., 1996) was included to assess the severity of depressive symptoms, which was shown to have positive correlations with depression disorders. The inventory has emerged as the most widely utilized approach for assessing depressive symptomatology in non-clinical populations. The inquiries included in the inventory are presented in Table S1 in the supplement. Anxiety is a personality trait, which has strong association with the risk of depression. The STAI has been used to assess explicit anxiety levels (Shek, 1993; Tyc et al., 1995) as a reflection of the degree of mental tension induced by stressful experiences, predicted future dangers, or anticipated failures. Because STAI-Trait scores have a high correlation with scores on BDI, we selected 20 items in the STAI-State scale as predictors to assess how a subject feels at the time s/he takes the inventory. Depression is a symptom that may fluctuate over time. The STAI-State scale similarly indicates the present condition and is more closely linked to the disorders. The questions in the scale can be found in Table S2 in the supplement.

### *3.3. SNP Genotyping Data*

A bioinformatic search was conducted for candidate genes that maximized the genetic effects on the risks of depression, anxiety and other affective disorders (Ivanov et al., 2022; Ivanov et al., 2019). A list of the 165 pairs of alleles available in the database can be found in Table S3 in the supplement. The collection of genotyping data for compiling lists of candidate genes in the ICBrainDB had a focus on identifying genes that differentiated phenotypic effects in different groups of people, that deviated from the standard methods for identifying genes with the most matching possible effects for all groups of people examined. The selection of a list of genes was based on the literature which suggested heterogeneity in phenotypic manifestation in different human populations. The 165 pairs of alleles can be classified according to their functional roles; for instance, Table S4 in the supplement presents the classification of genes based on their associations with serotonin and dopamine. Other genes can also be classified according to neuronal functions, cell signaling, cell functions, or immune functions. In feature selection, a pair of single nucleotide variants (SNVs) corresponding to a SNP gene were summed to obtained the total SNVs for data analysis.

### *3.4. Results*

The continuous SNP data were categorized by the Chi-squared automatic interaction detection (CHAID) procedure, which classified data into mutually exclusive, exhaustive, subsets that best describe the dependent variables (Kass, 1980). By using these CHAID splits, important genes were selected by the forward step and deleted by the backward step. Due to the extensive number of genes, the 165 SNP genes were grouped into 5 homogeneous clusters using the k-mean method based on original continuous data. Discretized SNP data within each cluster were input into the MI selection algorithm, and this cycle was repeated across all clusters. The selected genes were then competed with each other in the second cycle to conclude the feature selection procedure. In the first cycle, there were eight SNP genes selected, namely, ADCY2, DARPP32, GPR6, HTR2C, OXTR, PENK, TF, and WFS1. In the second cycle, the MI selection algorithm was applied to these eight candidates, resulting in the selection of only one gene, namely, ADCY2 (or HTR2C). With the exception of OXTR, the eight genes chosen in the first cycle shared a common characteristic: patients displayed lower counts of SNVs in comparison to controls. Table 1 shows the cross-classified tables between the target and gene classification groups for the eight candidates.

**Table 1.** The cross-classified tables between the target groups and genes.

| Genes | ADCY2 | | DARPP32 | | GRP6 | | HTR2C | | OXTR | | PENK | | TF | | WFS1 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Splits | ≤ 0 | > 0 | ≤ 1 | > 1 | ≤ 0 | > 0 | ≤ 0 | > 0 | ≤ 11 | > 11 | ≤ 0 | > 0 | ≤ 0 | > 0 | ≤ 10 | > 10 |
| Patients | 54 | 0 | 54 | 0 | 54 | 0 | 54 | 0 | 27 | 27 | 54 | 0 | 54 | 0 | 54 | 0 |
| Controls | 241 | 369 | 546 | 64 | 529 | 81 | 241 | 369 | 510 | 100 | 249 | 361 | 354 | 256 | 280 | 330 |

In the table, a total of 241 controls and 54 patients exhibited less SNVs in the ADCY2 polymorphic site and were categorized together. The 241 controls with lower counts of SNVs were not significantly associated with any ethnic/culture groups. Conversely, 369 controls had greater SNVs relative to the reference genome. To distinguish between patients and controls among the 295 unclassified subjects, CHAID splits were identified for their responses to the 41 psychological items. CHAID splits are listed in Tables S1 and S2 for the BDI and STAI-State scales, respectively. There were 7 items insignificant in the Chi-squared tests and were eliminated from further analysis. The MI selection algorithm was applied to the rest of items based on the 295 unclassified subjects and ST1, ST11, Bdi1 and Bdi21 were selected by the algorithm. Incorporating ACDY2 (or HTR2C), ST1, ST11, Bdi1, Bdi21, and Age (with CHAID splits of ≤ 22, (22-39], and > 39) into a logistic regression model by considering only main effects resulted in an overall predictive accuracy of 97.4%. Note that the interaction effects (i.e., CMI) among these predictors were already examined and found to be insignificant before specifying the logistic model using only main effects (Cheng & Liou, 2024). In Table 1, the HTR2C gene produced an identical cross-classified table to that of ADCY2, and it is linked to serotonin. When substituting ADCY2 with HTR2C, the predictive accuracy continued to be 97.4%.

Conversely, the stepwise logistic regression applied to the continuous SNP data and 41 ordinal scale scores identified 13 predictors, namely, ADCY2, CALU, GRIK4, MAOA, PPP1R1B, ST4, ST11, ST16, ST17, ST20, Bdi13, Bdi21, and Age, with the predictive accuracy of 100%. Table 2 shows the cross-classified tables between the target groups and five selected genes. As shown in Table S4, the SNP data associated with serotonin and dopamine were cumulated separately. The stepwise logistic regression selected fourteen predictors, namely, serotonin, dopamine, ST4, ST5, ST6, ST11, ST16, ST17, Bdi3, Bdi5, Bdi6, Bdi12, Bdi17, and Age, with the predictive accuracy of 98.8%. As indicated in Tables S1 and S2, ST5 and Bdi6 were insignificant in CHAID Chi-Squared tests.

**Table 2.** The cross-classified tables between the target groups and genes.

| Genes | ADCY2 | | CALU | | | GRIK4 | | | | | | MAOA | | PPP1R1B | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Splits | ≤ 0 | > 0 | ≤ 5 | (5,6] | > 6 | ≤ 7 | (7,10] | (10,14] | (14,15] | (15,19] | >19 | ≤ 3 | > 3 | ≤ 0 | > 0 |
| Patients | 54 | 0 | 41 | 10 | 3 | 0 | 2 | 19 | 14 | 16 | 3 | 54 | 0 | 54 | 0 |
| Controls | 241 | 369 | 430 | 46 | 134 | 128 | 63 | 117 | 45 | 124 | 133 | 531 | 79 | 405 | 205 |

Adenylate cyclase 2 (ADCY2) and 5-Hydroxytryptamine Receptor 2C (HTR2C) showed greater SNVs in normal controls but none in depressed patients, a fact which might suggest a protective or compensatory genetic mechanism. ADCY2 is an enzyme involved in the cyclic AMP (cAMP) signaling pathway, which plays a crucial role in neurotransmitter signaling and mood regulation (X. Cheng et al., 2008; Gray, Nash, & Yao, 2024). Genetic research has pointed to variations in ADCY2 as a factor in mood disorders. A genome-wide association study has discovered polymorphisms in ADCY2 associated with bipolar disorder, indicating that a decrease in functional diversity within ADCY2 (as evidenced by lower counts of SNVs) could hinder cAMP-dependent signaling and increase the risk of developing depressive phenotypes (Sen et al., 2025; Chen et al., 2022). HTR2C encodes the 5-HT2C serotonin receptor, which modulates dopamine and serotonin release—both critical in mood and emotional processing (Brummett et al., 2014;). Clinical and molecular investigations have associated variations in HTR2C with depression and suicidal behavior. A modified burden of single nucleotide variants—particularly diminished diversity—in HTR2C may limit the range of RNA-edited receptor isoforms, interfere with serotonin signaling, and make carriers susceptible to mood dysregulation. In fact, irregularities in RNA editing and polymorphisms within HTR2C have consistently been identified in victims of suicide who were depressed, potentially contributing to the dysfunction of receptors in these individuals (Dracheva et al., 2008; Iwamoto, Bundo & Kato, 2009; Wang et al., 2000).

HTR2C and ADCY2 interact with environmental factors through their roles in neurochemical signaling pathways that are sensitive to stress, hormones, and metabolic cues. These two SNP genes are functionally connected through their roles in G protein-coupled receptor (GPCR) signaling pathways. Nucleotide variations in these genes in healthy individuals might enhance or stabilize signaling efficiency, potentially buffering against mood dysregulation. Conversely, the absence of such variations in depressed patients could indicate a lack of this protective modulation, making them more susceptible to dysregulated neurotransmission and depressive symptoms. This pattern suggests that certain nucleotide variations in ADCY2 and HTR2C may confer resilience rather than risk, highlighting the complexity of genetic contributions to psychiatric conditions.

## 4. Discussion

In the empirical example, the MI selection algorithm primarily pinpointed a limited set of genes exhibiting low SNVs among patients, while a greater level of SNVs was observed in most controls. The psychological item scores, Age, and ADCY2 (or HTR2C) did not interact with one another, as assessed by the conditional mutual information. Consequently, the logistic regression that incorporated the main effects of the chosen predictors through CHAID splits was deemed valid. Although the logistic regression based on the original SNP data and ordinal item scores achieved higher predictive accuracies of 100% and 98.8%, respectively, the selected genes did not share common characteristics, complicating their interpretation.

When discrete variables are involved in the data, modern criteria for selecting useful features, predictors and valid models include AIC, BIC, and several Lasso-type penalty approaches. The theoretical foundations of feature selection have been thoroughly explored in the context of Generalized Linear Models (GLM), as highlighted by researchers such as Linhart and Zucchini (1986) and Burnham and Anderson (2010). The discourse surrounding these model and variable selection techniques has been extensively documented, with notable contributions from McCullagh and Nelder (1989), Fahrmeir and Tutz (1994), and Agresti (2013). Furthermore, a related area of study has focused on model selection in generalized additive models, as examined by Hastie (1986), Stone (1986), and Hastie and Tibshirani (1987). In this context, the method of average derivative estimation of scaled coefficients in econometrics was advocated by Stoker (1986), Power, Stock, and Stoker (1989), and Härdle and Stoker (1989), among others. In contrast to conventional methods, the MI feature selection algorithm presented in Section 2 typically selects a more limited number of predictors, regardless of the sample sizes, as illustrated in the empirical example. Moreover, the algorithm facilitates the evaluation of interaction effects among the selected predictors before establishing a valid logistic regression model to predict disease risk. The proposed algorithm is also recommended for feature selection in the context of big data analysis.

**Supplementary Materials:** The following supporting information can be downloaded at the website of this paper posted on Preprints.org.

**Author Contributions:** Conceptualization, P.-E.C. and M.L.; methodology, P.-E.C. and M.L.; software, J.-D.L.; validation, P.-E.C., J.-D.L.,A.N.S., S.-K.L., and M.L.; formal analysis, P.-E.C., J.-D.L., and M.L.; investigation, P.-E.C., J.-D.L., S.-K.L., A.N.S., and M.L.; resources, A.N.S.; data curation, J.-D.L., S.-K.L., A.N.S., and M.L.; original draft preparation, P.-E.C., J.-D.L., S.-K.L., A.N.S. and M.L.; review and editing, P.-E.C., J.-D.L., S.-K.L., A.N.S. and M.L.; All authors have read and agreed to the published version of the manuscript.

**Acknowledgments:** This work was supported by the National Science Council, Taiwan, under Grants NSC112-2410-H-001-068 and NSC113-2410-H-001-091.

## References

1. Agresti A. (2013). *Categorical data analysis*. 3rd ed., Wiley, New York.
2. Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans Auto Cont*., **19**, 716-723.
3. Amari, S. (2001). Information geometry on hierarchy of probability distributions. *IEEE Trans Info Th*., **47**, 1701-1711.
4. Beck, A. T., Steer, R. A., Ball, R., & Ranieri, W. F. (1996). Comparison of Beck Depression Inventories-IA and-II in psychiatric outpatients. *Journal of personality assessment*, *67*(3), 588-597.
5. Bishop, Y. M., Fienberg, S.E., and Holland, P. W. (1975). *Discrete Multivariate Analysis*: *Theory and Practice*, Cambridge, MA: MIT Press.
6. Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* **37**, 373–384
7. Brummett, B.H. et al. (2014). A putatively functional polymorphism in the HTR2C gene is associated with depressive symptoms in white females reporting significant life stress, PLoS ONE 9(12): e114451. DOI:10.1371/journal.pone.0114451
8. Burnham, K. P. and Anderson, D. R. (2010). *Model Selection and Multimodal Inference*: *A Practical Information-Theoretic Approach*, 2nd ed. New York: Springer.
9. Chan, T. E., Stumpf, M. P. H. and Babtie, A. C. (2017). Gene regulatory network inference from single-cell data using multivariate information measures. *Cell Systems*, 5, 251-267.
10. Chen Q, Ding J, An L, Wang H. (2022). Ca2+-stimulated adenylyl cyclases as therapeutic targets for psychiatric and neurodevelopmental disorders. Front Pharmacol. 13: 949384.
11. Cheng, P. E., Liou, J. W., Liou, M and Aston, J. A. D. (2006). Data information in contingency tables: A fallacy of hierarchical log-linear models. *J. Data Science*, **4**, 387-398.
12. Cheng, P. E., Liou, M., Aston, J. A. and Tsai, A. C. (2008). Information identities and testing hypotheses: Power analysis for contingency tables. *Statistica Sinica*, **18**, 535-558.
13. Cheng, P. E., Liou, M. and Aston, J. A. (2010). Likelihood ratio tests with three-way tables. *J. Am. Stat. Asso.*, **105**, 740-749.
14. Cheng, P. E. and Liou, M. (2024). Mutual information decomposition with applications. *Behaviormetrika*, https://doi.org/10.1007/s41237-024-00241-6.
15. Cheng, X., Ji, Z., Tsalkova, T., & Mei, F. (2008). Epac and PKA: a tale of two intracellular cAMP receptors. *Acta biochimica et biophysica Sinica*, *40*(7), 651-662.
16. Dracheva S, Patel N, Woo DA, Marcus SC, Siever LJ, & Haroutunian V. (2008). Increased serotonin-2C receptor mRNA editing: a possible risk factor for suicide. *Mol Psychiatry.* 13: 1001-1010.
17. Eguchi, S. & Copas, J. (2006). Interpreting Kllback-Leibler divergence with the Neyman-Pearson Lemma. *J. Multiv. Anal*., 97, 2034-2040.
18. Fahrmeir, L., Hamerle, A. and Tutz, G. (1984). *Multivariate Statistische Verfahren* [*Multivariate Statistical Analyses*]; Walter de Grnyter: Berlin, Germany. (In German)
19. Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modeling Based on Generalized Linear Models*; Springer: Berlin, Germany.

doi:10.20944/preprints202507.1954.v1

20. Gray, M, Nash, K. R., & Yao, Y. (2024). Adenylyl cyclase 2 expression and function in neurological diseases. *CNS Neurosci Ther*. 30(7), e14880.
21. Hӓrdle, W. and Stoker, T. (1986). Investigating smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc.* **84**, 986-995.
22. Hastie, T. (1986). Generalized additive models. *Stat. Sci.* **1**, 297-318.
23. Hastie, T. and Tibshirani, R. (1987). Generalized additive models: some applications. *J. Amer. Statist. Assoc.* **82**, 371-386.
24. Ince, R. A. A., Bruno L. Giordano, B. L., Kayser, C., Rousselet, G. A., Gross, J. and Schyns, P. G. (2017). A Statistical Framework for Neuroimaging Data Analysis Based on Mutual Information Estimated via a Gaussian Copula. *Human Brain Mapping* **38**, 1541-1573.
25. Ivanov, R., Kazantsev, F., Zavarzin, E., Klimenko, A., Milakhina, N., Matushkin, Y. G., Savostyanov, A., & Lashin, S. (2022). ICBrainDB: An Integrated Database for Finding Associations between Genetic Factors and EEG Markers of Depressive Disorders. *Journal of personalized medicine*, *12*(1), 53.
26. Ivanov, R., Zamyatin, V., Klimenko, A., Matushkin, Y., Savostyanov, A., & Lashin, S. (2019). Reconstruction and analysis of gene networks of human neurotransmitter systems reveal genes with contentious manifestation for anxiety, depression, and intellectual disabilities. *Genes*, *10*(9), 699.
27. Iwamoto K, Bundo M, Kato T. Serotonin receptor 2C and mental disorders: genetic, expression, and RNA editing studies. RNA Biol. 2009; 6(3): 248–253.
28. Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **29**, 119-127.
29. Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Stat*. **22**, 79-86.
30. Linhart, H. and Zucchini, W. (1986). Finite sample selection criteria for multinomial models. *Statistische Hefte*. **27**, 173-178.
31. Liou, J. W., Liou, M. and Cheng, P. E. (2023). Modeling categorical variables by mutual information decomposition. *Entropy*. *25*(5), 750; https://doi.org/10.3390/e25050750.
32. Mallows, C. L. (1973). Some comments on $C_p$. *Technometrics*, **15**, 661–675.
33. McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed., Chapman & Hall, London.
34. McGill, W. J. (1954). Multivariate information transmission. *Psychometrika*, **19**, 97-115.
35. McMahon, S.S., Sim, A., Johnson, R., Liepe, J., and Stumpf, M.P.H. (2014). Information theory and signal transduction systems: from molecular information processing to network inference. *Semin. Cell Dev. Biol.* **35**, 98–108.
36. Powell, J. L., Stock, J. H. and Stoker, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica* **57**, 1403-1430.
37. Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* 6, 461–464.
38. Sen P et al. (2025). A bipolar disorder-associated missense variant alters adenylyl cyclase 2 activity and promotes mania-like behavior. Mol Psychiatry. 30(1): 97–110.
39. Shannon, C. E. (1948). A mathematical theory of communication. *Bell Sys. Tech. Journal* **27**, 379-423; 623-656.
40. Shek, D. T. (1993). The Chinese version of the State-Trait Anxiety Inventory: Its relationship to different measures of psychological well-being. *Journal of clinical psychology*, *49*(3), 349-358.
41. Spielberger, C. D. (2010). State-Trait anger expression inventory. *The Corsini encyclopedia of psychology*, 1-1.
42. Stoker, T. M. (1986). Consistent estimation of scaled coefficients. *Econometrica* **54**, 1461-1481.
43. Stone, C. J. (1986). The dimensionality reduction principle for generalized additive models. *Ann. Stat.* **14**, 590-606.
44. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* 1996, *58*, 267–288.
45. Timme, N., Wesley, A. Flecker, B. and Beggs, J. M. (2014). Synergy, redundancy, and
46. multivariate information measures: an experimentalist’s perspective. *J. Comput. Neurosci*. **36**, 119-140.
47. Tyc, V. L., Fairclough, D., Fletcher, B., Leigh, L., & Mulhern, R. K. (1995). Children’s distress during magnetic resonance imaging procedures. *Children’s Health Care*, *24*(1), 5-19.
48. Vergara, J. R. and Estévez, P. A. (2014). A review of feature selection methods based on mutual information. *Neural Computing and Applications*, **24**, 175-186.

49. Vrshek-Schallhorn, S. (2015). Additive genetic risk from five serotonin system polymorphisms interacts with interpersonal stress to predict depression. *Journal of Abnormal Psychology*, 124(4), 776-790.
50. Wang Q, O'Brien PJ, Chen CX, Cho DS, Murray JM, Nishikura K. Altered G protein-coupling functions of RNA editing isoform and splicing variant serotonin2C receptors. *J Neurochem* 2000; 74:1290-300.
51. Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley, Chichester.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.