Quality Technology & Quantitative Management
Vol. 11, No. 1, pp. 61-70, 2014

QTQM

The Rank Minrelation Coefficient

Patrick E. Meyer*

Bioinformatics and Systems Biology Unit, Faculty of Sciences, Université de Liège (ULg)
27 Bld du Rectorat, 4000 Liège, Belgium
(*Received* August 2013, *accepted* February 2014)

Abstract: Bivariate (or pairwise) information measures such as mutual information or correlation are heavily used in variable selection and network inference algorithms mainly because they are faster and require fewer samples than multivariate (or multidimensional) strategies. This paper proposes a new relevance measure that aims at improving the detection of relevant variables based on pairwise measures. The new measure is called the rank minrelation coefficient because of its connection to the rank correlation coefficient. However, on the contrary to correlation, the minrelation is not symmetric. More explicitly, if a variable X exhibits a minrelation to Y then, as X increases, Y is likely to increases too, but, if X decreases, little can be said on Y values (except that the uncertainty on Y actually increases). In this paper, we introduce a new rank coefficient and an associated relevance measure that targets the detection of a characteristic dependency, connected to the concept of probabilistic implication. Finally, we show through several key examples and experiments that this new coefficient is competitive in order to select relevant variables, in particular when compared to correlations and mutual information.

Keywords: Correlation, information measure, probabilistic implication, probabilistic logic.

1. Introduction

Many methods of variable selection, such as ranking [6], mRMR [19], FCBF [22], and network inference, such as relevance networks [4], Aracne [15], CLR [7], MRNET [16], rely on pairwise estimation of relevance. There are important advantages in using pairwise strategies. First, it does not require a large amount of samples to estimate a bivariate dependency accurately. Second, with n variables, there are only $n(n-1)/2$ possible pairs and both computation and memory requirements are manageable even with very large datasets. These two advantages are extremely valuable in fields such as bioinformatics where the ratio n (variables) / m (samples) is very high. As a result, improving the detection of relevant variables using a bivariate measure could importantly improve both variable selection and network inference algorithms that are based on a pairwise measure. The objective of this paper is precisely to improve the detection of relevant variables using a bivariate measure that rely on the concept of probabilistic implication.

Let us first provide some examples where our measure could be particularly relevant:

(1) When the price of aluminum increases the prices of cars is likely to increase too. However, if the price of aluminum drops, the price of cars might stay high because of other components or some technological and economical considerations (i.e. other relevant variables). However, if the price of cars become really low, then it is likely that

* Corresponding author. E-mail: Patrick.Meyer@ulg.ac.be

the price of their components, including aluminum, are considered as low too (i.e. relatively to their average values).

(2) An increase in the level of adrenaline leads to an increased heart rate, but a low level of adrenaline do not prevent a high heart rate (because of other relevant variables) and a high heart rate do not mean a high adrenaline level. However, a low level of adrenaline is likely to be observed in a person having a low heart rate (w.r.t. to his/her usual heart rate).

(3) Let $Y = X_1.X_2$ with $X_1, X_2 \in [0,1]$. In such case, a low X_1 implies a low Y but a high X_1 has little information on Y (because a low X_2 automatically means a low Y whatever the value of X_1). However, a high Y (w.r.t. its average) automatically implies a high X_1.

In those examples, where variable dependencies are not quite correlations, looking for correlations might be fastidious because in order to observe joint variations, it might require that no other effect impacts the measured variables or that all those effects cancel each other. Indeed, Spearman's correlation relies on a symmetric concordance between ranking values of X and Y. However, our minrelation coefficient relies on an asymmetric concordance between squared ranks of X and Y. As a result, if a variable X exhibits a minrelation to Y then, as X increases, Y is likely to increases too, but not the other way around.

In the next Section 2, the concept of probabilistic implication is introduced. In Section 3, a rank coefficient that ranges between -1 and 1 is designed in view of detecting probabilistic implications. In Section 4, a relevance measure is proposed, based on the properties of our rank coefficient. Finally, in Section 5, the new coefficient and its associated relevance measure are benchmarked for variable selection tasks, on (i) toy examples, (ii) publicly available synthetic datasets, and (iii) public real datasets.

2. Probabilistic Implication

Let m samples of $X \in [0,1]$ and $Y \in [0,1]$ be drawn uniformly with the condition that $\forall (x_i, y_i),\ x_i \leq y_i,\ i \in \{1,...,m\}$ (see Figure 1).

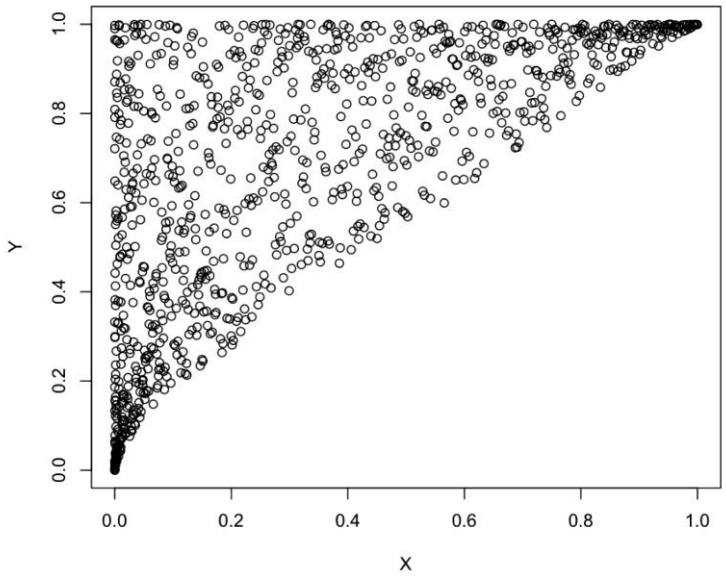

Figure 1. Typical plot of a probabilistic implication. For a linear model, the variance of $Y|X$ decreases as X increases and symmetrically the variance of $X|Y$ decreases as Y decreases.

This particular triangular dependency pattern of Figure 1, has been characterized in [5] as a probabilistic version of an implication (called probabilistic implication). This can be intuitively explained by the fact that as X increases Y is likely to increase too (similarly to the logical rule known as *modus ponens* where the words *likely to increase* are replaced by *is true*) and as Y decreases, X is likely to decreases together with its uncertainty (similarly to the *modus tollens* rule) but nothing can be said on the other variables when X is minimal or Y is maximal (as in the logical implication $X \rightarrow Y$).

Indeed, [5] discretizes X and Y in two (or three) classes [False, (neutral), True], in order to detect the number of samples that satisfy the triangular truth table of the logical implication, see Table 1. The higher is $\hat{p}(X=1,Y=0)$ the less likely the implication, as can be seen in Table 1. Several papers [8, 9, 12, 11] have used the same reasoning (See equation (1)) in order to compute a probabilistic version of the logical implication,

$$p(X \rightarrow Y) = 1 - p(X=1, Y=0)\,. \tag{1}$$

Table 1. Truth table of the implication and the associated probability distribution with binary variables where x_0 and x_1 denote respectively $X=0$ and $X=1$. The probability distribution is obtained by sampling variables uniformly and rejecting samples that do not satisfy the implication truth table.

$X \rightarrow Y$	x_0	x_1
y_0	1	**0**
y_1	1	1

$p(X \rightarrow Y)$	x_0	x_1	$p(Y)$
y_0	0.33	0	0.33
y_1	0.33	0.33	0.66
$p(X)$	0.66	0.33	1

However, to the best of our knowledge, this paper brings the first method allowing the detection of probabilistic implications without requiring to discretize variables (into at most 3 classes). Also, when binary variables are independent, equation (1) will return a value of 0.75 (since $p(X=1,Y=0)$ would equal 0.25). This is not a desirable feature for a relevance measure. We will show that our proposed coefficient does not suffer from these drawbacks and is, as a result, able to compete with correlations and mutual information (see Section 5).

3. Minrelation

It is interesting to note that using the same approach as in Section 2, correlation could be seen as a statistical analog of the logical equivalence (assuming normally distributed variables). Indeed, when X or Y increases or decreases the other variable is likely to changes in the same way (similarly to the logical equivalence where the word "likely to increase" and "likely to decrease" can be replaced by “is true” and “is false”). We can also randomly sample X and Y in the $\{0,1\}$ domain and only accept samples that satisfy the equivalence truth table and obtain the probability distribution of Table 2.

Table 2. Truth table of the equivalence and associated probability distribution with binary variables where y_0 and y_1 denote respectively $Y=0$ and $Y=1$.

$X \rightarrow Y$	x_0	x_1
y_0	1	**0**
y_1	**0**	1

$p(X \leftrightarrow Y)$	x_0	x_1	$p(Y)$
y_0	0.5	0	0.5
y_1	0	0.5	0.5
$p(X)$	0.5	0.5	1

In order to assess how likely a set of measurements shows a probabilistic equivalence between two variables, it is straightforward to consider the *concordance probability* $p_c = p(x_0, y_0) + p(x_1, y_1) = p(X \leftrightarrow Y)$ as the probability of an equivalence. Similarly, the *discordance* probability reflects the anti-equivalence $p_d = p(x_1, y_0) + p(x_0, y_1)$.

We have $p_c + p_d = 1$ which is analog to the Boolean property stating that equivalence and anti-equivalence cannot be both true at the same time. The coefficient measuring the difference between concordance and discordance probabilities

$$\rho = p_c - p_d, \tag{2}$$

returns a value between $+1$ (perfect equivalence) and -1 (anti-equivalence) with 0 indicating that the variables are as equivalent than anti-equivalent, which is the case when two variables are independent. The Blomqvist's medial correlation coefficient is an example of such measure [17].

Many rank correlation coefficients can be formulated as an extension of equation (2) with the following formula,

$$\frac{C(F(X),F(Y)) - C(F(X),G(Y))}{C(F(X),F(Y)) + C(F(X),G(Y))}, \tag{3}$$

where $C()$ is a concordance function between rankings and where $F()$ and $G()$ are functions that maps values to ranks.

For example, taking $F()$ as the function that maps m values sorted in increasing order to their respective ranks $(1, 2, ..., m-1, m)$, $G()$ a function that maps values in increasing order to ranks in decreasing order $(m, m-1, ..., 2, 1)$ and $C()$ as the inverse of the sum of the squared differences between rankings,

$$C(F(X),F(Y)) = \frac{1}{\sum (F(X) - F(Y))^2}, \tag{4}$$

lead to $\rho(X,Y)$ Spearman's correlation coefficient. Kendall's tau can be expressed similarly [17].

In this paper, we define the minrelation coefficient $\iota(X,Y)$ based on equation (3) with $F()$ the function mapping values in increasing order to squared ranks in increasing order, i.e. $x_i = i^2$ with x_i denoting the *i-th* values of X (once sorted). Note that this weighting function of ranks has been previously used in [13]. However, in our proposal the function $G()$ maps increasing values on squared ranks in decreasing order, i.e. $y_i = (m+1-i)^2$. This weighting of ranks has also been used previously in [2] but not in conjunction with our chosen function $F()$. This choice of $F()$ and $G()$ not only map the range of values of $F(X)$ and $G(Y)$ to a normalized range $[1, m^2]$ of values but also map X and Y to triangular distributions concordant with the marginals of X and Y, observed in both, Table 1 and Figure 1.

Finally, the concordance function chosen is given by

$$C(F(X),F(Y)) = \sum_{(F(X)+F(Y)-(m^2+1)>0)} (F(X) + F(Y) - (m^2+1)). \tag{5}$$

This is the sum of the distance of the points on one side of the diagonal to the diagonal itself. Indeed, not taking the negative values in this sum renders our concordance function asymmetric, which introduces the interesting properties of this coefficient. Indeed, looking

at distances from both side of the diagonal would turn equation (5) into a rank correlation (with a new weighting of ranks). Note also that the distance is not the squared difference between ranks (as in Spearman's correlation) but rather a difference between squared ranks.

In other words, a rank correlation ρ measures the deviation from the equality between two rankings (i.e., $F(X)=F(Y)$), whereas the rank minrelation ι is defined to measure the deviation from $F(X)\leq F(Y)$ (but with $F(X)$ and $F(Y)$ being squared rankings).

As a result, if both $F(X)\leq F(Y)$ and $F(Y)\leq F(X)$, then $F(X)=F(Y)$ (correlated), which is analog to the Boolean property stating that if both, $X\rightarrow Y$ and $Y\rightarrow X$ are true, then the two variables are equivalent $X\leftrightarrow Y$. In such a limit case, we have $\iota(X,Y)=\iota(Y,X)=1=\rho(X,Y)$.

4. Relevance Measure

The coefficient $\iota(X,Y)$ ranges between -1 and 1. If $\iota(X,Y)=1$ then we are in the case of Figure 1 suggesting a probabilistic implication ($\hat{p}(X\rightarrow Y)=1$), if $\iota(X,Y)=-1$ then $\iota(X,-Y)=1$ suggesting ($\hat{p}(X\rightarrow -Y)=1$), finally, $\iota=0$ means that the joint distribution is as distant from each of the two triangular joint distributions discussed above (which will be the case when X and Y are independent or if X or Y is constant).

Because there are four different triangular distributions corresponding to the four corners of the (X,Y)-plane, we have $\iota(X,Y)=-\iota(X,-Y)\neq\iota(-X,-Y)=-\iota(-X,Y)$. As a result, in order to have a relevance measure (ranging between 0 and 1) that allows to compare variables in order to predict Y, we define $\iota^*(X,Y)$ as

$$\iota^*(X,Y)=\max|(\iota(X,Y),\iota(-X,Y))|. \quad (6)$$

The rationale is the following: if $\iota^*(X,Y)=1$ then the joint distribution of (X,Y) is a triangular distribution in any of the four possible corners of the (X,Y)-plane. The lower $\iota^*(X,Y)$, the further away from a probabilistic implication pattern is the (X,Y) joint distribution.

5. Experiments

The goal of this section is to show the usefulness of our coefficient $\iota^*(X,Y)$ in data analysis. We first demonstrate its competitiveness on toy examples, then on artificial and real datasets by plugging it into the well-known ranking method [6]. In each experiment, we compare $\iota^*(X,Y)$, equation (6), with (i) $\hat{p}^*(X\rightarrow Y)$ the maximum probabilistic implication using two-classes discretized variables, equation (1), (ii) $\rho_P^2(X,Y)$ the squared Pearson correlation, (iii) $\rho^2(X,Y)$ the squared Spearman's correlation and (iv) $NMI(X,Y)=I(X;Y)/H(X,Y)$ the normalized mutual information using the non-parametrical entropy estimator having the best trade-off accuracy-speed reported in [18] (i.e. the empirical estimate on $\sqrt{m}$ equal frequency binning, m being the number of samples). These various relevance measures can easily be compared because these all return a value between 0 and 1 and have an algorithmic complexity $\leq O(m.log(m))$.

5.1. Multiplication and Addition

Let $X_s=X_i+X_j$ and $X_p=X_i.X_j$ be respectively the sum and the product of two continuous random variables X_i and X_j (uniformly and independently distributed on [0, 1]).

We observe, in Table 3, that when the variables are independent, such as X_i and X_j, all the measures, except the probabilistic implication, agree on a near zero relevance. However, only $\imath^*(X,Y)$ and the probabilistic implication identify X_i as relevant in both cases, the sum and the product.

Table 3. Comparisons of respectively, squared Pearson's correlation, squared Spearman's correlation, normalized mutual information, $\hat{p}^*$ (using two-classes discretized variables), $\imath^*$ using rank minrelation, on X_p a sum and of X_s a product of two continuous random variables X_i and X_j (uniformly and independently distributed on [0, 1]). The experiment was repeated 1000 times.

X_i	1000 samples from uniform $\in[0,1]$
X_j	1000 samples from uniform $\in[0,1]$
X_s	$= X_i + X_j$
X_p	$= X_i.X_j$

X	X_i	X_i	X_i
Y	X_s	X_p	X_j
$\rho_P^2(X,Y)$	0.50	0.43	0
$\rho^2(X,Y)$	0.49	0.44	0
$NMI(X,Y)$	0.13	0.15	0.08
$\hat{p}^*(X \to Y)$	0.88	0.92	0.76
$\imath^*(X,Y)$	0.94	0.99	0.04

5.2. Multiplication Together with Linear Dependencies

Let us consider a noisy linear dependencies ($G = A + \varepsilon$) and non-noisy multivariate product ($A = B.C.D$), that can both be used to predict a target variable A (see Table 4). Which of G or B,C,D are better ranked by the various criteria, in order to predict A (where $B.C.D$ are uniformly and independently distributed and ε normally distributed)?

Both $NMI(X,Y)$ and $\hat{p}^*(X \to Y)$ seem to have a small spread between relevant and irrelevant values. Interestingly correlations and mutual information would rank variable G first whereas $\imath^*(X,Y)$ would rank higher B,C,D.

Table 4. Comparisons of respectively, squared Pearson's correlation, squared Spearman's correlation, normalized mutual information, max probabilistic implication using two-classes discretized variables, $\imath^*$ using rank minrelation, averaged over 1000 repetitions.

B,C,D	$= unif(0,1)$
A	$= B \times C \times D$
ε	$= N(0,15)$
G	$= A + \varepsilon$

X	A	A	A
Y	B,C,D	G	ε
$\rho_P^2(X,Y)$	0.24	0.48	0
$\rho^2(X,Y)$	0.28	0.34	0
$NMI(X,Y)$	0.11	0.12	0.08
$\hat{p}^*(X \to Y)$	0.92	0.92	0.83
$\imath^*(X,Y)$	0.88	0.79	0.05

5.3. Artificial Dataset

The question that arise at this point is: *is* $\imath^*(X,Y)$ *able to discriminate between relevant and irrelevant variables better than* $\rho^2(X,Y)$ *or* $NMI(X,Y)$*?* In order to answer this question, we make use of a synthetically generated dataset where relevant and irrelevant variables are known. In this experiment, we compare the performances of variables ranking

using the various criteria. We consider a ranking strategy superior if the average position of the relevant variables using that criterion is lower than for the others. The rationale being that a better selection criterion should return a lower average position (i.e. relevant variables should be ranked first).

As artificial datasets, we adopt the 10 datasets (i.e. KO1...KO5, MF1...MF5) of 100 variables and 100 samples coming from the DREAM4 challenge [14] where the goal was to identify predictor variables for each variables (i.e. a network inference task). However, in this case, we focus only on the few variables per dataset that have more than 10 predictors. This minimal number of predictors ensures some stability in the results reported. Indeed, if one predictor variable happens to be badly ranked, there are at least 9 others that could compensate (if the criterion is indeed superior). The number of target variables (i.e. ranking tasks), wins and losses of each criterion for each dataset are reported in Table 5.

Table 5. Wins/losses of $\iota^*(X,Y)$ vs other information measures in ranking strategies on target variables having more than 10 predictors in the 10 datasets of the DREAM4 competition. Column 1 indicates the dataset, column 2 indicates the number of variables having more than 10 predictors in that dataset and columns 3, 4, 5 and 6 reports the wins and losses of the two ranking methods on those target variables. A method wins if the average position of the predictors in the ranking is lower than for the other method. Bold notations are used when $\iota^*(X,Y)$ outperform the other criterion.

W/L of ι^*	vs $\rho_P^2(X,Y)$	vs $\rho^2(X,Y)$	vs $NMI(X,Y)$	vs $\hat{p}^*(X \to Y)$	#targets
KO1	**4/1**	**3/2**	1/4	**5/0**	5
KO2	**5/2**	2/5	**5/2**	**6/1**	7
KO3	**4/1**	**5/0**	**5/0**	**5/0**	5
KO4	**4/2**	3/3	**4/2**	**5/1**	6
KO5	3/6	4/5	4/5	**7/2**	9
MF1	0/5	**4/1**	**3/2**	**5/0**	5
MF2	**4/3**	**4/3**	**5/2**	**7/0**	7
MF3	**3/2**	**4/1**	**4/1**	**5/0**	5
MF4	3/3	3/3	**5/1**	**6/0**	6
MF5	**8/1**	**7/2**	**7/2**	**7/2**	9
Tot	**38/26**	**39/25**	**43/21**	**58/6**	64

We observe that ι^* exhibits results that are more than 50% better than its competitors on a task that consists in identifying the known set of predictors of target variables of ten artificial datasets. We observe that Pearson's correlation is the second best relevance measure coherently with the conclusions reached in [14].

5.4. Real Datasets

In the previous task, the variables to be selected were known in advance. It is usually not the case in real datasets. In order to compare ranking strategies on real data, we evaluate the prediction accuracy of different learning algorithms (i.e. linear model, random forest and radial SVM) using as input variables the best ranked ones using our various criteria. We assume here that a better criterion leads to a better ranking of variables which in turn leads to better prediction performances of a model built on these top ranked variables. We carried out an experimental session based on four regression datasets publicly available [21]. For computational reasons, we have limited the number of samples per

dataset to 500 (randomly sampled). The name of the datasets together with the number of variables and number of samples are reported in Table 6.

In order to eliminate a possible variable selection bias, each dataset is first divided into two equal parts, one for ranking variables and one for evaluating those rankings. The evaluation of a ranking method is given by the mean squared error returned by a 10-fold cross-validation of a linear regression (R lm function), a SVM with radial kernel (R package e1071) and a random forest (R package randomForest). In order to avoid the bias related to the size of the feature set, we average the performance over all the feature sets size (that range from 2 to 10 for each dataset), as done in [3]. Table 7 reports the statistically significant wins and losses on the four datasets for each of the three learning algorithms. Non-statistically-significant results are considered as tails.

Table 6. Regression datasets, together with their number of variables n and number of samples m.

dataset	name	n	m	dataset	name	n	m
1	Ailerons	35	500	3	Triazines	58	186
2	Pol	26	500	4	Wisconsin	32	194

Table 7. Statistically significant wins/tails/losses, using 10-fold-cross-validated mean squared error returned by a linear regression, a random forest and a radial SVM averaged over subset sizes ranging from 2 to 10. For each of the four datasets, a ranking is returned using $\iota^*(X,Y)$ and compared to the ranking provided by each other criterion.

W/T/L of ι^*	vs $\rho_P^2(X,Y)$	vs $\rho^2(X,Y)$	vs $NMI(X,Y)$	vs $\hat{p}^*(X \to Y)$
LIN	**2/2/0**	**1/3/0**	1/2/1	**2/2/0**
RFOREST	2/0/2	**2/1/1**	2/0/2	**4/0/0**
SVM	**2/2/0**	1/2/1	1/2/1	**3/1/0**
Total	**6/4/2**	**4/6/2**	4/4/4	**9/3/0**

We observe here that $\iota^*(X,Y)$ outperform by far $\hat{p}^*(X \to Y)$. It also outperforms both correlations, and reach similar results than the normalized mutual information criterion.

6. Conclusions

The goal of this paper has been to introduce a new coefficient called a rank minrelation, meant to capture probabilistic implication patterns. The $\iota(X,Y)$ coefficient is ranging between -1 and 1 through 0 when variables are independent. We deem that our analysis on toy examples, our experiments both on synthetic data and real data shows clearly that this new coefficient offer an interesting way of capturing a multivariate dependency at the bivariate level. Our new coefficient has been shown competitive with four of the most used information measures, namely empirical mutual information, Spearman's and Pearson's correlations and the maximal probabilistic implication (based on discretized variables). Those measures exhibit good trade-offs interpretability-accuracy-speed. Indeed, all these measures have an algorithmic $\leq O(m.\log(m))$, including our new measure. It is likely that heavier information measures, for example, kernel-based or copulas-based, would not be as fast. However, future experiments will address this question. Also, new selection strategies that use more efficiently the directionality of minrelations

should be compared to more efficient variable selection strategies than the simple ranking method.

Many important theoretical questions are left untreated in this paper such as the asymptotic properties of the coefficient, an analysis of the convergence of the coefficient w.r.t. the targeted triangular joint distribution, or even the implicit assumptions made on the probability distributions of X and Y. However, further research will address these questions.

Finally, the connection between probabilistic implication and causality has not been discussed either. This omission has been made deliberately because we feel that this delicate subject require an entire article by itself. However, we refer the reader to [1, 10, 20] for additional reading on this topic.

References

1. Baral, C. and Hunsaker, M. (2007). Using the probabilistic logic programming p-log for causal and counterfactual reasoning and non-naive conditioning. *IJCAI 07*, 243-249.
2. Blest, D. C. (2000). Theory and methods: Rank correlation-an alternative measure. *Australian and New Zealand Journal of Statistics*, 42(1), 101-111.
3. Bontempi, G. and Meyer, P. E. (2010). Causal filter selection in microarray data. *Proceedings of the 27th International Conference on Machine Learning (ICML 10)*, 95-102.
4. Butte, A. J. and Kohane, I. S. (2000). Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements. *Pacific Symposium on Biocomputing*, 5, 418-429.
5. Sahoo, D., Dill, D. L., Gentles, A. J., Tibshirani R. and Plevritis, S. K. (2008). Boolean implication networks derived from large scale, whole genome microarray datasets. *Genome Biology*, 9(10), R157.
6. Duch, W., Winiarski, T., Biesiada, J. and Kachel, A. (2003). Feature selection and ranking filters. *International Conference on Artificial Neural Networks (ICANN) and International Conference on Neural Information Processing (ICONIP)*, 251-254.
7. Faith, J., Hayete, B., Thaden, J., Mogno, I., Wierzbowski, J., Cottarel, G., Kasif, S., Collins, J. and Gardner, T. (2007). Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. *PLoS Biology*, 5(1), 54-66.
8. Gras, R., Kuntz, P. and Briand, H. (2001). Les fondements de l'analyse statistique implicative et quelques prolongements pour la fouille de données. *Mathématiques et Sciences Humaines. Mathematics and Social Sciences*, 154-155, 9-29.
9. Haenni, R. (2005). Towards a unifying theory of logical and probabilistic reasoning. *ISIPTA 05, 4th International Symposium on Imprecise Probabilities and Their Applications*, 5, 193-202.
10. Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press.
11. Josang, A. (2007). Probabilistic logic under uncertainty. *Proceedings of the Thirteenth Australian Symposium on Theory of Computing*, 65, 101-110. Australian Computer Society, Inc.
12. Liu, J. and Desmarais, M. (1997). A method learning implication networks from empirical data: Algorithm and monte-carlo simulation-based validation. *IEEE Transactions on Knowledge and Data Engineering*, 9(6), 990–1004.

13. Mango, A. (1997). Rank correlation coefficients: A new approach. *Computational Statistics and Data Analysis on the Eve of the 21st Century. Proceedings of the Second World Congress of the IASC*, 29, 471-476.

14. Marbach, D., Schaffter, T., Mattiussi, C. and Floreano, D. (2009). Generating realistic in silico gene networks for performance assessment of reverse engineering methods. *Journal of Computational Biology*, 16(2), 229-239.

15. Margolin, A. A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R. D. and Califano, A. (2006). ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7, 1-15.

16. Meyer, P. E., Kontos, K., Lafitte, F. and Bontempi, G. (2007). Information-theoretic inference of large transcriptional regulatory networks. *EURASIP Journal on Bioinformatics and Systems Biology*, Special Issue on Information-Theoretic Methods for Bioinformatics, DOI: 10.1155/2007/79879.

17. Nelsen, R. B. (2001). Kendall tau metric. *Encyclopedia of Mathematics*, 3, 226-227.

18. Olsen, C., Meyer, P. E. and Bontempi, G. (2009). On the impact of entropy estimation on transcriptional regulatory network inference based on mutual information. *EURASIP Journal on Bioinformatics and Systems Biology*, DOI: 10.1155/2009/308959.

19. Peng, H., Long, F. and Ding, C. (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8), 1226-1238.

20. Raedt, L. D. and Kersting, K. (2003). Probabilistic logic learning. *ACM SIGKDD Explorations Newsletter*, 5(1), 31-48.

21. Torgo, L. http://www.liaad.up.pt/ltorgo/regression/datasets.html.

22. Yu, L. and Liu, H. (2004). Efficient feature selection via analysis of relevance and redundancy. *Journal of Machine Learning Research*, 5, 1205-1224.

Author's Biography:

Patrick Emmanuel Meyer received the Electromechanical Engineering degree in 2003 and the Ph.D. degree in Sciences (statistical machine learning), in 2008, from the Universite Libre de Bruxelles (ULB, Belgium). After postdoctoral research in computational biology at the Computer Science and Artificial Intelligence Laboratory of the Massachusetts Institute of Technology (CSAIL, MIT, USA), at the BROAD Institute of MIT and Harvard (USA) and at the FNRS (Belgium), he became, in 2014, Professor in Bioinformatics at the Universite de Liege (ULg, Belgium). Among other scientific productions, he co-authored, the open-source R and Bioconductor package, Mutual Information NETworks and the Drosophila modENCODE paper published in Science. His interests cover statistical machine learning, information theory and systems biology.