Big Data Analysis and Data Science Master
Statistics Project
Mixed Interaction Conditional Gaussian Model for hair chemical analysis

Roberto Gallea

November 2, 2020

Abstract

This report summarizes a study aimed the estimation and evaluation of a Mixed Interaction (MI) Conditional Gaussian (CG) graphical model with purpose of verifying whether it is possible to discriminate people coming from volcanic geological areas from those coming from areas made up of sedimentary rocks, through the analysis of chemical elements collected from hair samples. Such hypothesis verification would reinforce the idea that chemical analysis on hair could lead to identifying areas of environmental risk.

1 Introduction and Overview

Chemical hair analysis is a wide-spread method used in several fields. For example, it is used in forensic toxicology for the detection of many therapeutic drugs and recreational drugs, including cocaine, heroin, benzodiazepines and amphetamines [6], [1]. It has several advantage since it is less invasive than a blood test, and carries time-varying information due to hair's slow-growth. It was also used as a preliminary screening method for the presence of toxic substances deleterious to health after exposures in air, dust, sediment, soil and water, food and toxins in the environment [2].

In this context, chemical hair analysis is applied to detect, using a Mixed Interaction Conditional Gaussian Graphical Model, whether people are coming for volcanic geological areas or not. If such a model could be determined, similar models could be used to assess environmental risks of a region.

The supporting idea is that there exist some elements which people living around Mt. Etna area are more exposed to, mainly due to the volcanic activity. Referring to literature [3], [4], [5], among the elements that could be expected to be found are the following: *Al, Co, Ni, Fe, Si, As* would derive mainly from alteration of the volcanic rocks of Etna; *K, Na, V, Sr, Mo, Cr* would instead indicate a major contribution of volcanic gases. The report is organized as follows: Section 2 provides a description of the used dataset, along with its exploratory analysis. Section 3 illustrates how the model is estimated. Moreover, a benchmark of several variable reduction methods is assessed and results are provided. Finally 4 summarizes the research and outlines further areas of development.

2 Dataset description and analysis

The dataset used is a collection of chemical elements measurements for 417 hair samples, taken from adolescents coming from two geological areas of Sicily island (Italy). The first area is the volcanic region around Mt. Etna, the second area is lithologically constituted by sedimentary rocks. For each sample 116 elements were measured, collected and normalized, along with the information about the region of provenience (*ET*, 274 samples, or *SIC*, 143 samples) and the gender (*male* 197 sample, or *female*, 220 samples). However, since hair characteristics could be considered the same for male and female, such variable is removed in order not to introduce an additional degree of non-necessary complexity.

2.1 Normality test

MI-CG models assume that variables have a normal conditional distribution. Thus, in order to justify the use of such a model, some normality tests are taken on elements variable conditional distributions, w.r.t. to the geological area (*Code* variable). Three types of tests are performed:

- *Direct inspection*: Conditional probability density functions for each chemical variable are plotted and the shape was inspected (Figure 1).
- *Kurtosis*: Kurtosis value (1) (i.e. the degree of tailedness) was extracted for each conditional probability density function. Such values are shown in Figure 2a.

$$Kurt\left[X\right] = E\left[\left(\frac{(X-\mu)^4}{\sigma}\right)\right] \tag{1}$$

- *Negentropy*: Negentropy is the distance of a pdf from normality (2):

$$J(p_x) = S(\phi_x) - S(p_x), \tag{2}$$

  where $S(\phi_x)$ is the differential entropy of the Guaussian density with the same mean and variance as p_x and $S(p_x)$ is the differential entropy of p_x:

$$S(p_x) = -\int p_x(u)\log p_x(u)du, \tag{3}$$

  Negentropy values were extracted for each conditional probability density function. Such values are shown in Figure 2b.

As can be seen from the three figures, almost all of the variables are quasi-normally distributed, and thus can be considered normal for the graphical model estimation.

2.2 Correlation analysis

The following analysis is aimed to find which variables are most correlated and which ones are not. For this purpose factor analysis (FA) was performed. Factor analysis is a method which models the interrelationships among variables. It focuses primarily on their variance and covariance. Factor analysis assumes that variance can be partitioned into two types of variance, common and unique. The output of factor analysis is a set of *loadings* for each factor. The sum of squared loadings across a factor provides the total *communality* of a variable, i.e. the amount of variance that each variable shares with other ones. For this reason, uncorrelated variables exhibit low communality. By performing a 1-factor FA, variables are sorted according to their communality value. In the studied scenario the 10 most uncorrelated elements resulted to be *Zr, Cd, V, Sr, Co, U, At, As, Cs, Np*, while, the 10 most correlated elements are *Eu, Ge, Er, Nd, Ne, Tl, Th, Ra, Ti, La.* Note that, in the resulting model, one would expect not to see many correlated variables.

3 Methods and tests

After the previous considerations, the model should be estimated using a model selection algorithm. However, a direct estimation could be neither possible nor useful, since the amount of variables is too large. For this reason, a limited subset of variables has to be chosen before proceeding to model selection. All the code used for the method implementation and testing is available as supplemental material at the github repository https://github.com/unipa-bigdata/gallea-statistics-micg-hair.

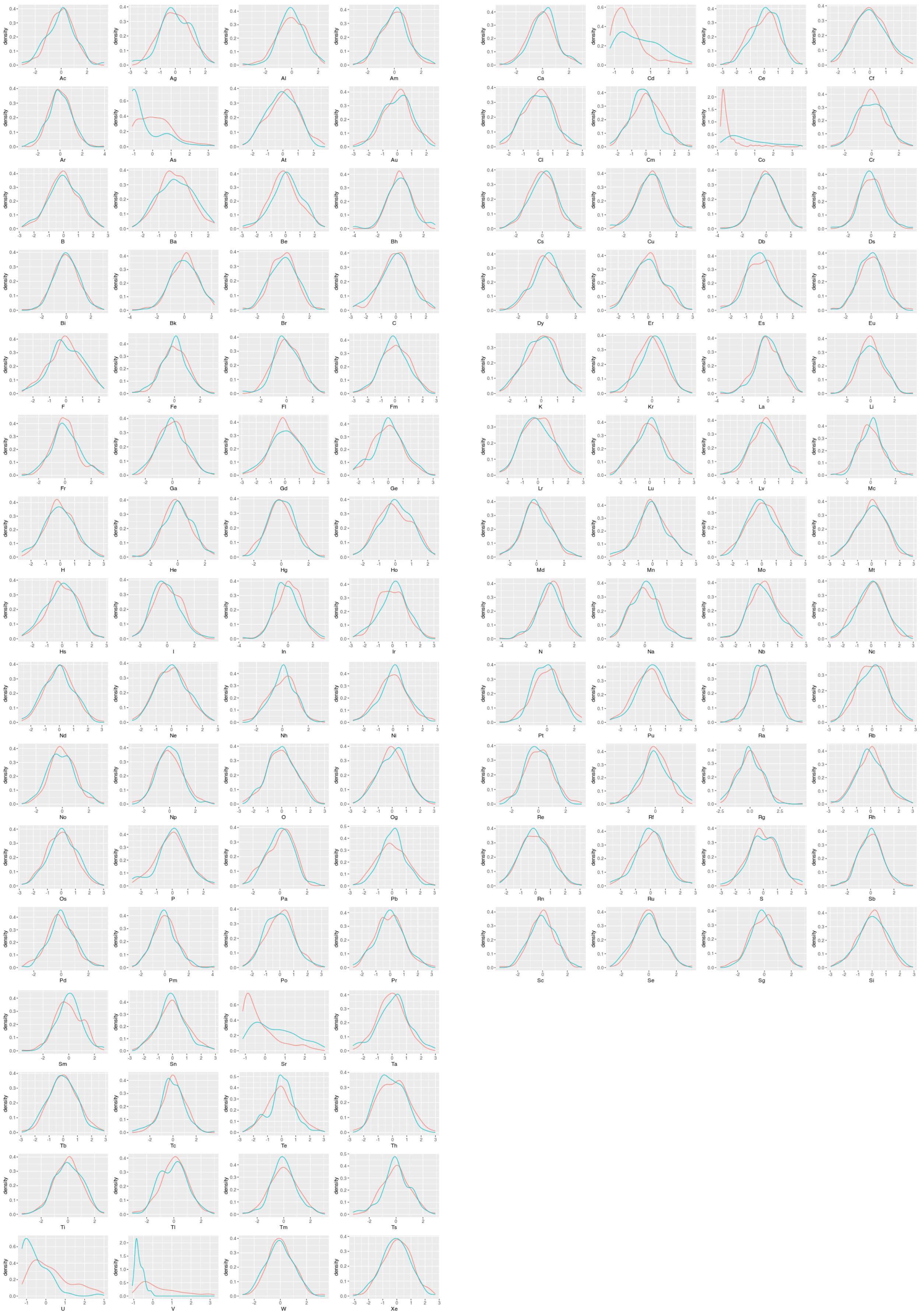
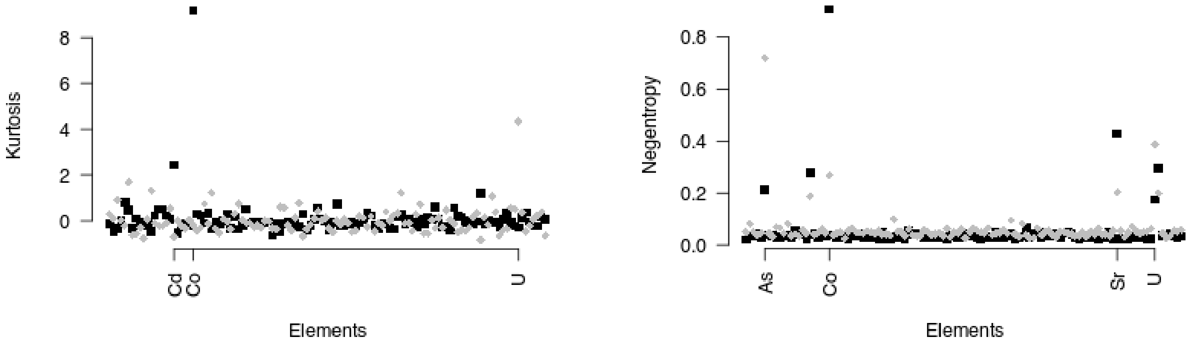


Figure 1: Chemical elements probability density function conditioned to *Code* variable.



(a)

(b)

Figure 2: Normality measure for chemical elements conditioned probability density functions: Kurtosis (a) and Neg-entropy (b).

3.1 Variables reduction

This is a mandatory step to make the model selection problem tractable. Two approaches are used to choose which variables to discard, the first one is informative, the latter is algorithmic:

- *Synthetic elements are not presents in volcano-related compounds*: thus all of the synthetic elements were discarded from the dataset. These are: *Am, Cm, Bk, Cf, Es, Fm, Md, Ls, No, Rf, Db, Sg, Hs, Mt, Ds, Rg, Nh, Fl, Mc, Lv, Ts, Og, Bh.* Others, such as *Tc, Ra, Sm, Md* are not strictly synthetic but there are few chances they could be related to volcanic effects. These were not discarded but are not expected to be found in the final model.

- *Include variables which exhibit different conditional pdfs*: if conditional *pdf*s are similar for a given variable, they won't affect the variable of intereset, thus they would not be included by any model selection method, so we can drop the most conditionally similar distributed variables. The choice of such *similarity*, denoted as Δ has been addressed in four different ways. This step is crucial for the resulting model selection.

3.1.1 Δ_μ - Difference of means

The first and simplest method is to drop variables which have a low absolute difference in their means.

$$\Delta_\mu = abs(\mu(X_{SIC}) - \mu(X_{ET})) \tag{4}$$

3.1.2 Δ_I - Mutual Information

This methods drops variables having high mutual information values for the respective conditioned *pdf*s. Mutual information is a measure of the informative content of a variable about another. It is similar to correlation, but takes into account non-linear relations.

$$\Delta_I = I(X_{SIC}, X_{ET}) = \sum_{x_{sic} \in X_{SIC}} \sum_{x_{et} \in X_{ET}} p_{(X_{SIC},X_{ET})}(x_{sic}, x_{et}) log \left(\frac{p_{(X_{SIC},X_{ET})}(x_{sic}, x_{et})}{p_{X_{SIC}}(x_{sic}) p_{X_{ET}}(x_{ET})} \right) \tag{5}$$

3.1.3 Δ_{KLD} - Kullback-Leibler Divergence

This methods drops variables having high Kullback-Leibler Divergence (KSD) values for the respective conditioned *pdf*s. KSD could be interpreted as the average difference of the number of bits required for

Variables Reduction metric	Deviance	LogL	BIC	DF-size
Means distance	1383.27	-8487.20	17348.44	89
MI	821.39	-8761.40	17854.62	96
KLD	731.96	-8806.24	17962.39	93
KST	1113.45	-8621.86	17611.74	90

Table 1: Model metrics for the four variable reduction methods.

encoding samples of one variable X using a code optimized for another variable Q rather than one optimized for X. Since such measure is not symmetric, the mean value is used.

$$\Delta_{KSD} = \sum_{x_{sic} in X_{SIC}} p_{X_{SIC}}(x_{sic}) log \left(\frac{p_{X_{SIC}}(x_{sic})}{p_{X_{ET}}(x_{sic})} \right) \tag{6}$$

3.1.4 Δ_{KST} - Kolmogorov-Smirnov Test

This methods drops variables having high Kolmogorov-Smirnov Test (KST). KST is a shape-test that can be used to compare a sample with a reference distribution or, as for the current application, to compare two samples. In the latter case:

$$\Delta_{KST} = sup_x \left| F_{SIC,n}(x) - F_{ET,m}(x) \right|, \tag{7}$$

where $F_{SIC,n}(x)$ and $F_{ET,m}$ are the empirical pdfs of the two samples respectively.

3.2 Model estimation and evaluation

Provided the four variables reduction methods described in the previous section, a model using $n = 15$ variables was estimated using a forward stepwise approach, maximizing Akaike's information criterion (AIC) value, and starting from a total independence model. The resulting models are depicted in Figure 3(a-d). The four graphs show similar results. Indeed, there is a graph community around *Code* node with almost the same elements: *Co, Sr, U, Cd, V, As.* The remaining nodes in the graph are peripheric, grouped in a well-distinct community and weakly connected to the kernel (just one link).

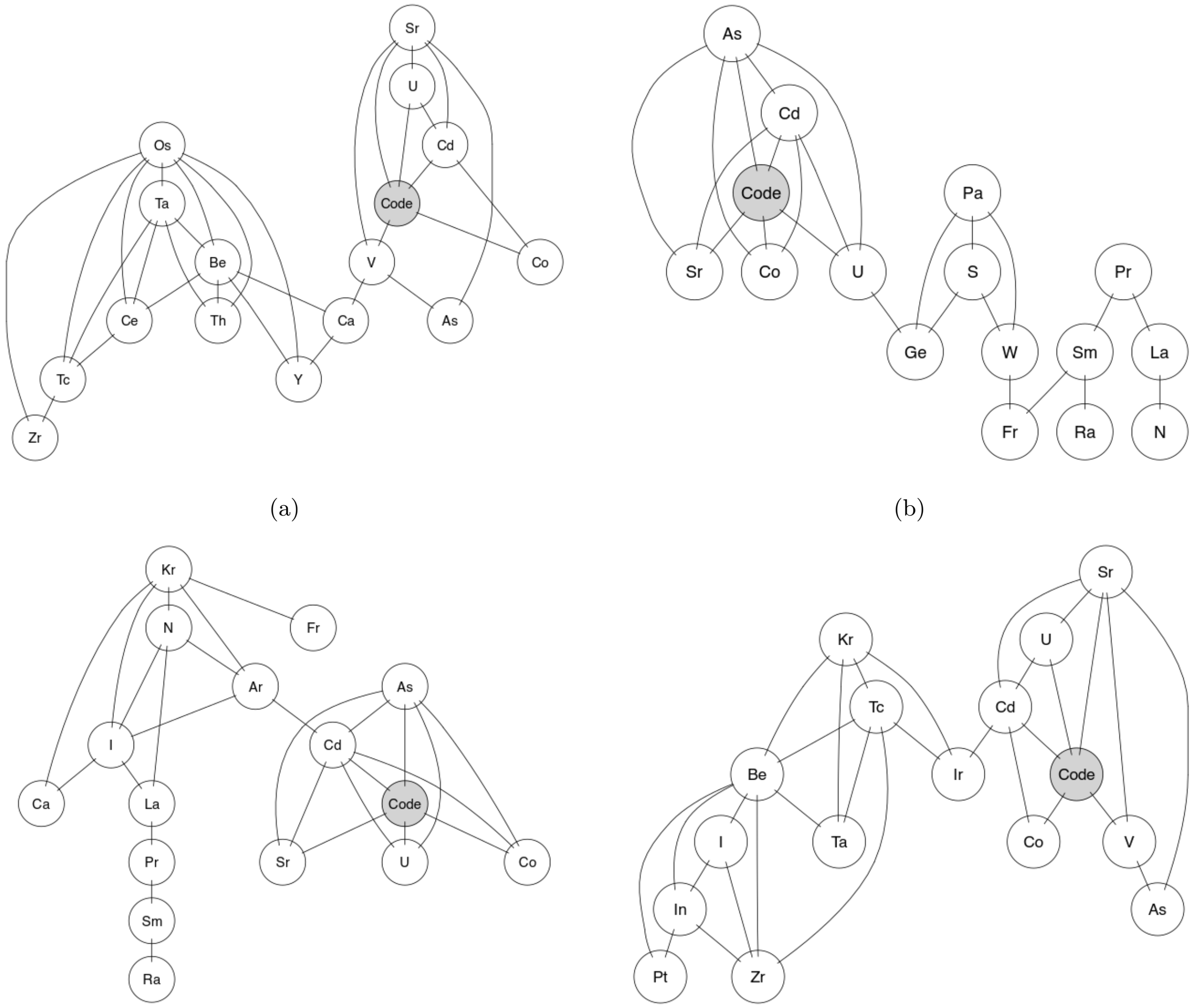
Some performance metrics are extracted from the four models and are shown in Table 1. The algorithmic results are compatible with the *a-priori* knowledge, indeed *CO, Sr* and *V* are among the expected elements, due to either volcanic rock composition or released gases.

3.3 Additional benchmarks

In order to validate how the variable reductions methods perform from a more general perspective, some benchmarks are taken:

- *Metrics as a function of number of variables*: for each variable reduction method, a model for $n = 2 \ldots 15$ variables was estimated and Deviance, LogL, BIC and DF-size were measured. Figure 4(a-d) show the results. The plots show how all the measure except Deviance are similar without regard to the method used. Instead, MI provides much better results for Deviance w.r.t. other methods, especially with a low number of variables. However, KLD compares to MI for higher number of variables. This is not surprising since MI computation comprises evaluation of KLD.

- *Robustness to noise*: To assess robustness to noise, the following experiment is taken: after determining the subset of N_{var} variables, N_{synth} synthetic noise variables with normal distribution are generated and variables reduction is performed again. The robustness metric is given by the minimum number of variables that must taken before all of the original subset is recovered. For clarification, consider the following example: $N_{var} = 3$ and $N_{synth} = 2$, variables are *Bo*, *Sr* and V and the procedure injects X_1, X_2. If variable reduction sorts variables as *Bo*, X_1, V, *Sr*, X_2, robustness measure is 4, because the initial subset is recovered after taking the first 4 variables. Hence, lower values correspond to better



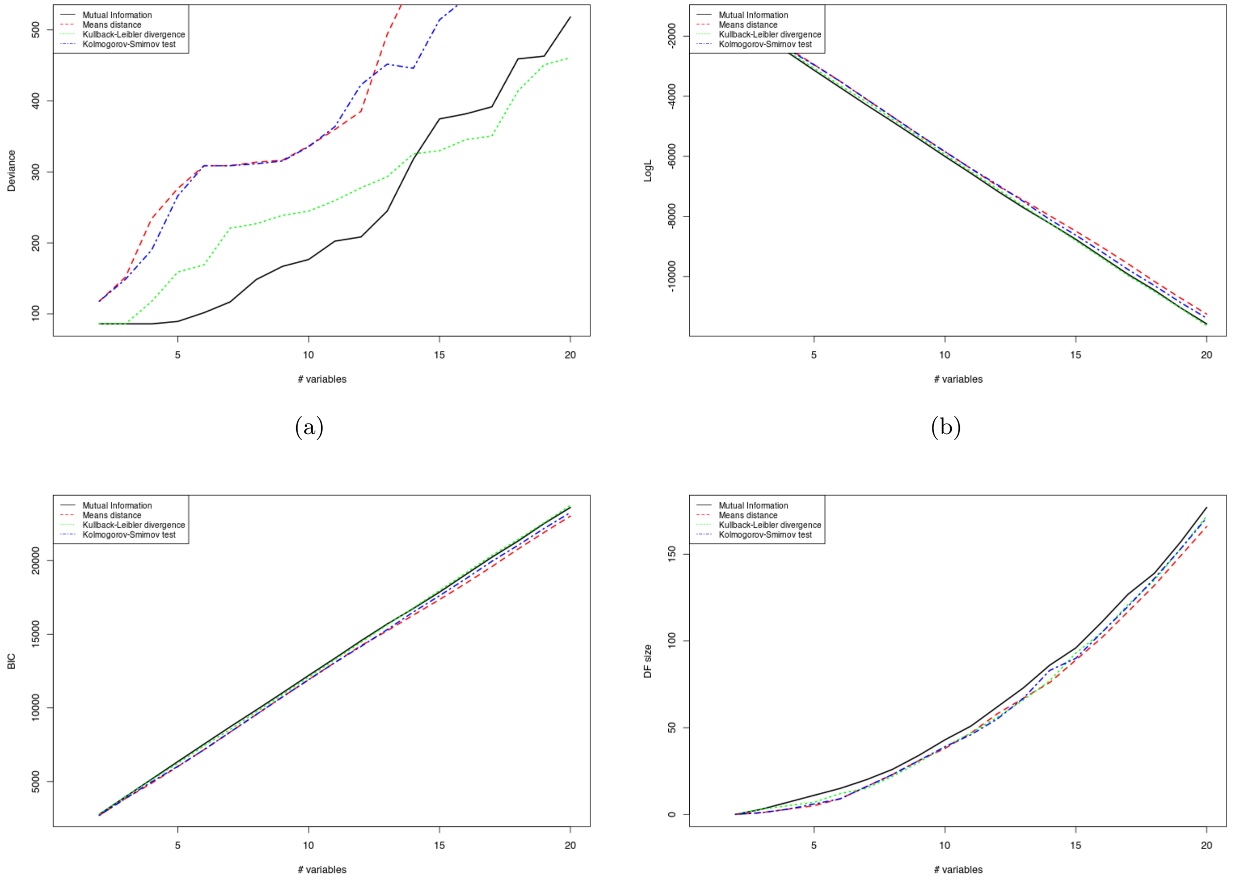
(a)

(b)

(c)

(d)

Figure 3: Models estimated using different variable reduction methods: Difference of means (a), Mutual Information (b), Kullback-Leibler Divergence (c), Kolmogorov-Smirnov Test (d).



(a)

(b)

(c)

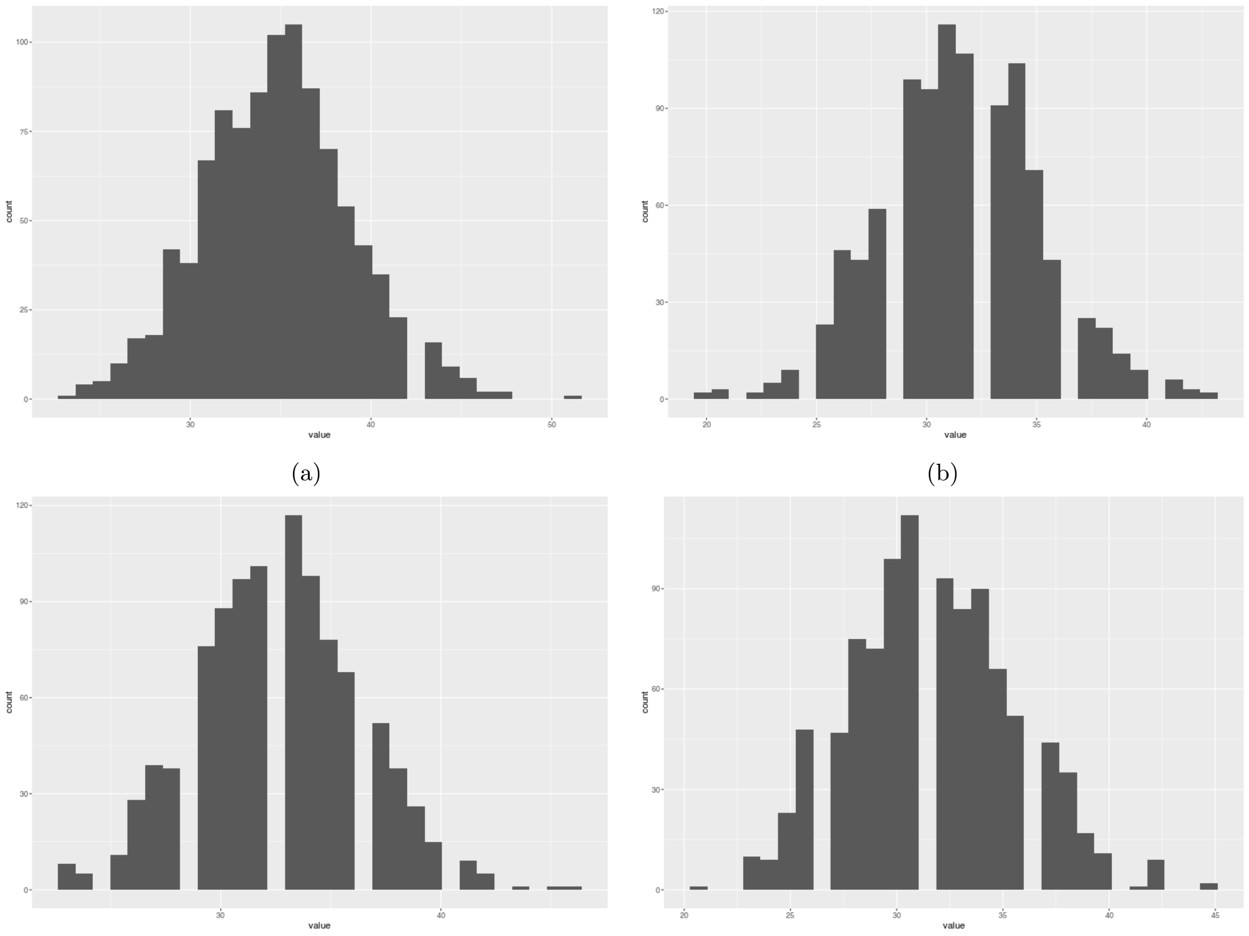
(d)

Figure 4: Plot of metrics as a function of the number of variables used for model estimation: Deviance (a), LogL (b), BIC (c), DF-size (d).

robustness. The range of such values is $[N_{var}, N_{var} + N_{synth}]$. The operation is repeated 1000 times with $N_{var} = 15$ and $N_{synth} = 100$. Robustness measures for each simulation are then binned in a histogram. Results shown in Figure 5 tell that MI is the most robust metric, since, in average, requires to take $\sim$ 31 variables to recover the original subset, while MD requires $\sim$ 35, KLD $\sim$ 33 and KSD $\sim$ 32. Gaps in the histogram are due to the fact that some of the real variables (which are fixed), due to their importance are always chosen before others, given the number of variables to include in the model. Indeed, the presence of such gaps indicates better discriminant power of the metric.

4 Summary and Conclusions

A Mixed-Interaction Gaussian Graphical model estimation was applied to the problem of detecting the people living in two different geological areas from the samples collected by a chemical hair analysis. After a dataset introduction, description and analysis, the number of variables has been reduced using priori information on the elements and four different algorithmic methods: mean difference (MD), Mutual Information (MI), Kullback-Leibler Divergence (KLD) and Kolmogorov-Smirnov Test (KST). The resulting model was evaluated and assessed in the four cases. Moreover, two benchmarks have applied to test the performance and the robustness of the variable reduction methods, which shows that MI is generally better than other methods.



(a)

(b)

(c)

(d)

Figure 5: Robustness measure histogram for MD (a), MI (b), KLD (c) and KST (d).

Further work can be done to improve the model translating a deeper geological knowledge, this will require the involvement of domain-specific expertise. Moreover the real discriminative ability of the model should be extensively tested ”in the wild” to assess its effectiveness.

References

[1] Marie Balikova. “Hair analysis for drugs of abuse. Plausibility of interpretation”. In: *Biomedical papers of the Medical Faculty of the University Palacký, Olomouc, Czechoslovakia* 149 (Jan. 2006), pp. 199–207. DOI: 10.5507/bp.2005.026.

[2] Donald W. Black et al. “Multiple Chemical Sensitivity Syndrome: Symptom Prevalence and Risk Factors in a Military Population”. In: *Archives of Internal Medicine* 160.8 (Apr. 2000), pp. 1169–1176. ISSN: 0003-9926. DOI: 10.1001/archinte.160.8.1169. eprint: https://jamanetwork.com/journals/jamainternalmedicine/articlepdf/485289/ioi90228.pdf. URL: https://doi.org/10.1001/archinte.160.8.1169.

[3] SALVATORE GIAMMANCO et al. “MAJOR AND TRACE ELEMENTS GEOCHEMISTRY IN THE GROUND WATERS OF A VOLCANIC AREA: MOUNT ETNA (SICILY, ITALY)”. In: *Water Research* 32.1 (1998), pp. 19–30. ISSN: 0043-1354. DOI: https://doi.org/10.1016/S0043-1354(97)00198-X. URL: http://www.sciencedirect.com/science/article/pii/S004313549700198X.

[4] E. A. Mathez. “Influence of degassing on oxidation states of basaltic magmas”. In: 310.5976 (Aug. 1984), pp. 371–375. DOI: 10.1038/310371a0.

[5] Silvio Mollo et al. *Cooling history of a dike as revealed by mineral chemistry: a case study from Mt. Etna volcano.* June 2011.

[6] Michael J. Welch et al. “Hair Analysis for Drugs of Abuse: Evaluation of Analytical Methods, Environmental Issues, and Development of Reference Materials*”. In: *Journal of Analytical Toxicology* 17.7 (Nov. 1993), pp. 389–398. ISSN: 0146-4760. DOI: 10.1093/jat/17.7.389. eprint: https://academic.oup.com/jat/article-pdf/17/7/389/2068756/17-7-389.pdf. URL: https://doi.org/10.1093/jat/17.7.389.