An exhaustive variable selection study for linear models of soundscape emotions: rankings and Gibbs analysis

R. San Millán-Castillo*, L. Martino*, E. Morgado*, F. Llorente†
* Dep. of Signal Theory and Communications, Universidad Rey Juan Carlos (URJC), Madrid, Spain.
† Dep. of Statistics, Universidad Carlos III de Madrid (UC3M), Madrid, Spain.

***Abstract*—In the last decade, soundscapes have become one of the most active topics in Acoustics, providing a holistic approach to the acoustic environment, which involves human perception and context. Soundscapes-elicited emotions are central and substantially subtle and unnoticed (compared to speech or music). Currently, soundscape emotion recognition is a very active topic in the literature. We provide an exhaustive variable selection study (i.e., a selection of the soundscapes indicators) to a well-known dataset (emo-soundscapes). We consider linear soundscape emotion models for two soundscapes descriptors: arousal and valence. Several ranking schemes and procedures for selecting the number of variables are applied. We have also performed an alternating optimization scheme for obtaining the best sequences keeping fixed a certain number of features. Furthermore, we have designed a novel technique based on Gibbs sampling, which provides a more complete and clear view of the relevance of each variable. Finally, we have also compared our results with the analysis obtained by the classical methods based on p-values. As a result of our study, we suggest two simple and parsimonious linear models of only 7 and 16 variables (within the 122 possible features) for the two outputs (arousal and valence), respectively. The suggested linear models provide very good and competitive performance, with $R^2 > 0.86$ and $R^2 > 0.63$ (values obtained after a cross-validation procedure), respectively.**

***Index Terms*—Soundscape emotion, variable selection, ranking methods, best sequence search, MCMC algorithms, Gibbs sampling.**

I. Introduction

Environmental noise is one of the most critical risks for population health and well-being. The World Health Organization has recently remarked that noise affects at least 100 million people, only in the European Union [41]. Generally, sound level monitoring and control are the common tools for managing the acoustic environment and sound quality remains dismissed. However, noise abatement is often unavailable or unsuitable in certain scenarios like cities, or does not necessarily result in an approving appraisal of final soundscapes [39]. Hence, "quiet areas" are a new perspective that focuses on the acoustic quality more than on the sound level, and which are being even regulated in the European Union [3].
This vision is limited since it is not accountable for people's experiences in different acoustic environments. Soundscapes provide an alternative and holistic approach to assess human perception, acoustic environments, and context, beyond the concept of noise [33]. Thus, this subjective evaluation depends on physical, psychological, social, and even cultural estimators and their complex interactions.

In the last decade, soundscapes have become one of the most active topics in acoustics. In fact, the number of related research projects and scientific articles grows exponentially [5]. Research requires a sizeable sample of participants in surveys and a considerable amount of locations. These intensive and time-consuming resources may limit the soundscape approach [24]. Soundscape modeling might predict people's perception of acoustic environments at lower expenses [20]. In urban planning and environmental acoustics, the procedure consists of (a) soundscapes recording, (b) calculation of acoustic and psychoacoustic indicators of the signals, (c) collecting other context indicators (e.g. visual information [7]), and (d) ranking of soundscapes audio signals employing emotional descriptors. Finally, the model can be developed.
Soundscape-elicited emotions are substantially different from those related to music or speech because they are more subtle and unnoticed. Thus, soundscape emotion recognition (SER) requires further research to support perception and context descriptors [12], [25]. Additionally to environmental acoustics and urban planning, there is an increasing research interest in SER for certain domains like sound design in films and digital games [22], or sonification in the *Internet of Things (IoT)* [1]. Soundscapes descriptors are identified with perceived emotions and SER becomes a relatively new sub-field of research in affective computing [10]. Rusell's circumplex model can be applied to soundscapes [7], [8], [10], [32] by scaling the perception of soundscapes. Russell's affect representation can be modeled with two main factors: *arousal* represents the eventfulness of the acoustic environment, and *valence* is the pleasantness ratio. Currently, arousal and valence are accepted as the principal and sufficient affective descriptors in research [8], [40], but there are different proposals to enhance soundscapes emotions evaluation with additional or different descriptors [4], and even to include emotion appraisal in procedures of the soundscapes standards [12].

Soundscape modeling has been extensively and recently

reviewed in [20]. Soundscapes indicators (i.e. features), soundscapes descriptors (i.e. outputs), and employed prediction models and their performances are presented. Researchers have been approaching SER from a variety of perspectives, and the results are roughly comparable. However, the published literature shows some trends. Firstly, a large dataset leads to stable and well-performing models. Indicators that include psychoacoustic and perceptual information contribute to improving model performance. Finally, *non-linear models* (NLMs) seem to result in (slightly) more accurate models than *linear models* (LMs). However, NLMs approaches remain complex and challenging for researchers since the model development and the hyperparameters tuning might become demanding. Hence, LMs are usually the preferred choice although they could be often outperformed by NLMs strategies. Some of the predictive LMs provide poor performance ($R^2 = 0.18$) [16], whereas other LMs achieve very good performance ($R^2 = 0.9$) [6]. On the other hand, reported NLMs use sophisticated machine learning techniques such as support vector machines (SVM), artificial neural networks (ANN), or random forests (RF) to name a few. They outperform slightly LMs in terms of prediction, e.g., regarding scores ($R^2 = 0.91$) [18]. Nevertheless, LMs still appear as prevalent in this field, while research with NLMs seems to be just promising, so far.

A wide range of descriptors is modeled by a large array of indicators in a variety of scenarios. Thus, a general framework for comparison seems not to be established. One of the reasons is the scarcity of SER datasets that are publicly available. Emo-soundscapes database (EMO) [10] sets up a free and available dataset for SER comparison from 2017, which is focused on arousal and valence. Thus, other researchers have been exploring EMO as a reference. Firstly, [10] presents a baseline for EMO based on two independent SVMs, in order to to model both arousal and valence, selecting 39 features by a variance threshold. In [2], a comparison of four LM and four NLM is explored and a dimension reduction is performed by a principal components analysis (PCA).[1] Furthermore, the authors in [10] also show that a RF model outperforms the rest of the models with only 25 features. In [1], a fine-tuned RF model with 14 features overcomes the previous RF model, and convolution neural networks (CNNs). Deep learning techniques have been also applied to SER through CNN and 23 simplified mel-frequency cepstral coefficients (MFCC) in [31], and the combination with SVM (Transfer learning) in [11]. Promising results use up to 54 features by heuristic methods despite the limited samples of EMO. Thus, EMO is the selected dataset for our study, because it is a suitable and relevant reference.

The goal of this work is to design simple and interpretable SER linear models. Additionally, this study offers an exhaustive feature selection framework that helps researchers adjust their model errors with the features importance and their relationships. Namely, we provide an exhaustive variable selection study for LM, considering classical and novel methodologies. Hence, the required resources for SER modeling become less laborious and the research community can employ the designed models to improve the knowledge about SER. A wide range of applications such as urban planning, noise monitoring, and sonification production might improve their performance based on SER models. Moreover, variable selection may lead to less significant computing resources. This helps to bring SER models to devices with real-time responsivity, such as the IoT framework. First of all, we divide the methods into two main parts: ranking of variables and selection of the effective number of variables. This approach yields several benefits: (a) allows a better understanding of the different techniques, (b) allows the combination of different ranking and number selection schemes, and (c) produces a more complete view of the variable selection problem. We consider five different ranking methods and also compare the results to the classical ranking method based on p-values [9], [17]. Moreover, we apply the best sequence search (keeping fixed the number of variables). For this purpose, we perform an alternating optimization method that allows us finding easily (at least) local modes. We repeat the procedure for several different runs for obtaining the global mode.

Last but not least, we design a pseudo-target density and a Gibbs sampling scheme which allows us having a complete view of the importance of the variables in terms of prediction error. The results of the Gibbs analysis support and clarify the results obtained previously by the ranking methods and the best sequence search. Some other considerations are only remarked by the Gibbs analysis. This novel technique can be applied for general variable selection purposes (not just for the specific database analyzed). The overall study allows us to propose (a) *parsimonious*, (b) *interpretable*, and (c) *robust* linear models. Namely, we can focus on a few very relevant variables that are highlighted in all the performed analyses. We believe that these variables keep their relevance also in different databases (as also suggested by the cross-validation results). Moreover, focusing on a few variables helps the interpretability of the resulting model.

In summary, we aim at developing well-performing, simple, robust and interpretable SER linear models. In order to achieve this objective, the contributions of the work are the following:

- We apply five different ranking methods to analyze the relevance of the variables in terms of prediction error in arousal and valence.
- Additionally, we apply two more sophisticated methods to find the most relevant variables: 1) the best sequence search; and 2) a Gibbs sampling approach building a suitable target density based on prediction error (a technique developed for the first time in this work but which can be applied for general variable selection purposes).
- Regarding the selection of the effective number of vari-

[1] In order to avoid confusions, it is important to remark that the dimension reduction obtained by a PCA is different from a dimension reduction obtained by applying a variable selection scheme. The dimension reduction by PCA is obtained by suitable linear combinations of variables. These linear combinations can be considered as "new variables" (and/or meta-features). A variable selection technique just selects some of existing variables trying to removing useless redundancy (without creating new features).

ables, we apply several information criteria (such as the well-known AIC and BIC [21]) and a classical $p-$value based approach. The obtained results are also compared with the Gibbs analysis.
- Based on the complete and exhaustive analysis of the previous methods, we offer two different linear models to predict arousal and valence from a very reduced number of variables and with low prediction error.

The rest of the work is organized as follows. Section II-A describes the dataset which is the object of our study. Section II-B presents some background material describing the LM. Section III describes the different techniques that we will use for our analysis: ranking methods, the algorithm for best sequence search, and the Gibbs sampling analysis. Then, Sections IV and V show the results applied to the EMO database for the first output (arousal) and the second output (valence), respectively. Finally, in Section VI, we discuss some conclusions. The detailed results for both outputs are given in the `Supplementary material`.

II. Database and model descriptions

A. Dataset Description

This research explores the EMO that might be considered as the largest publicly available soundscape database with annotations of emotion labels, and the most bench-marked up to now [10]. EMO contains 1213 audio clips which are released under Creative Commons license from the Freesound collaborative audio platform [13]. The Schafer's taxonomy classifies the selected clips into six categories due to their generality and simplicity . The Schafer's categories consider both the identification of the source and the listening context [10], [33] : natural sounds (e.g. birds, wind), human sounds (e.g. laugh, shouts), sounds and society (e.g. party, store), mechanical sounds (e.g. engine, factory), quiet and silence (e.g. quiet park, silent forest), and sounds as indicators (e.g. clock, church bells). EMO consists of 100 audio clips per category within a first subset (i.e. 600 audio clips) and 613 manually mixed sound of two or three categories of the first subset. A crowd-sourcing procedure provided data annotations of perceived arousal and valence, by a ranking-based questionnaire of a two clips pairwise comparison. Eventually, 1182 trusted annotators performed the required tasks with reasonable inter-subject reliability.
Audio clips are monophonic and the sample rate was 44100 Hz, , which is widely considered a high quality standard for audio files. Monophonic recordings are sufficient to evaluate the eventfulness (arousal) and pleasantness (valence) of acoustic environments, among many other soundscapes indicators, according to [42]. EMO has employed both YAAFE [30] and MIRToolbox [19] for the extraction of 122 normalized audio features, applying a 23 ms Hanning window with 50% overlapping. There are three main groups of features of the audio signals:

- **Psychoacoustic features:** They are indexed 4, from 24 to 49, and 113, 114, 117, 118, 119. These features represent perceptual (i.e., subjective) attributes of sounds such as level (i.e., loudness for overall level and MFCC for band-limited levels), spectrum (i.e., sharpness for high frequencies), and temporal and spectrum modulation (i.e., fluctuation).
- **Time-domain features:** They are indexed from 1 to 7, and 22, 23, 52, 115, 116. These features represent the signal dynamics, such as classical estimators based on samples of the audio signal (i.e., energy, entropy of energy, root mean square (RMS), or zero-crossing), the ratio between the magnitude difference at the beginning and the ending of a decay period (i.e., decrease slope), and the percentage of frames showing less energy than the average energy (i.e., low energy).
- **Frequency-domain features:** They are represented by the remainder indexes, i.e., the rest of features. These features represent the shape of the spectrum and the harmonic structure of sounds such as the fundamental frequency of the audio signal (i.e., pitch), the proportion of frequencies that are not multiple of the fundamental frequency (i.e., inharmonicity), the spectral representation based on the 12 equal-tempered pitches of Western music (i.e., chromagrams), the spectral statistical moments (i.e., centroid, the first one; spread, the second one; skewness, the third one), the ratio between the geometric mean and the arithmetic mean (i.e., flatness), the spectral changes between two successive frames (i.e., flux), and the estimation of the amount of high frequency (i.e., roll-off).

For further information, the complete database can be found at https://metacreation.net/emo-soundscapes/.

B. Multiple Linear Regression Model

Let us consider a set of R variables $\mathbf{x} = [x_1, ..., x_R]^\top$ (input vector) and a related variable y (output). In several real-world applications, we observe a dataset of N pairs $\{\mathbf{x}_n, y_n\}_{n=1}^N$. In this work, we consider the case that $R \leq N$. The relationship between inputs and outputs is then studied. A linear observation model is usually used,

$$\begin{aligned} y_n &= \beta_0 + \beta_1 x_{n,1} + \beta_2 x_{n,2} + ...\beta_R x_{n,R} + \epsilon_n, \\ y_n &= \beta_0 + \mathbf{x}_n^\top \boldsymbol{\beta} + \epsilon_n, \end{aligned} \tag{1}$$

where $\boldsymbol{\beta} = [\beta_0, \beta_1, ..., \beta_R]^\top$ is a vector of coefficients and ϵ_n is a Gaussian noise with zero mean and variance σ_ϵ^2, i.e., $\epsilon_n \sim \mathcal{N}(\epsilon|0, \sigma_\epsilon^2)$. Defining the vectors $\mathbf{y} = [y_1, ..., y_N]^\top$ and $\boldsymbol{\epsilon} = [\epsilon_1, ..., \epsilon_N]^\top$, and the $N \times (R+1)$ *design matrix* defined as

$$\mathbf{X} = \begin{bmatrix} 1 & \mathbf{x}_1^\top \\ \vdots & \vdots \\ 1 & \mathbf{x}_N^\top \end{bmatrix},$$

the previous model can be rewritten in the following way,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{2}$$

The least squares (LS) estimator (which coincides with the maximum likelihood estimator, in this setting) is

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}. \tag{3}$$

Hence, the vector of output predictions according to the model is $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}$, and the error vector is

$$\widehat{\mathbf{e}} = \mathbf{y} - \widehat{\mathbf{y}} = \left(\mathbf{I} - \mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\right)\mathbf{y}, \tag{4}$$

where $\widehat{\mathbf{e}} = [\widehat{e}_1, ..., \widehat{e}_N]^\top$, and $\mathbf{I}$ is a $N \times N$ unit (diagonal) matrix. The mean absolute error (MAE) and the mean squared error (MSE) - in one specific realization[2] - are defined as $\texttt{MAE} = \frac{1}{N}\sum_{n=1}^{N}|\widehat{e}_n|$ and $\texttt{MSE} = \frac{1}{N}\sum_{n=1}^{N}\widehat{e}_n^2$, respectively. We also denote as y_1, y_2 the output 1 (arousal) and the output 2 (valence), respectively.

III. Robust variable selection analysis

Variable selection is one of the most important task in machine learning, signal processing and statistics. The main goal of a variable selection approach is to remove the redundancy contained in the data, ideally without any loss of information or, at least, without incurring in a sensible loss of information. Variable selection methods are conceptually formed by two main parts. The first stage consists in *ranking* the variables for their importance, measured with some suitable criterion. In a second stage, based on the previous ranking, a selection of *the number of relevant variables* is performed. This last part can be considered a dimension reduction step, and it is also strongly connected to the model selection problem in nested models (in this case, the model selection problem is often called *order* selection) [21], [36], [37].

Generally, in the literature, these two stages are jointly presented within a unique technique, including the second part as a *stopping rule* in the ranking procedure (e.g., using a *threshold value* and stopping the rank at some position once the threshold value is reached) [15]. Here, we describe separately these two stages: the ranking methods in the next subsection and the selection of the number of features in Section III-D. Hence, we can combine different ranking schemes with different procedures for selecting the number of variables. In this work, we describe five different ranking methods (RMs). The first four ranking procedures are based, in a different way, on the prediction error. To the best of our knowledge, the procedures RM3 and RM4 described below present also some degree of novelty. They allow us to perform a more robust analysis, as discussed in Section IV. As an additional final check on the obtained results, we also apply a classical ranking method based on p-values [9], [17].

In Section III-B, we also describe the best sequence search (keeping fixed the number M of variables in the sequence) and an alternating optimization technique for obtaining the optimal sequence. Furthermore, we introduce a target density based on the prediction error, and employ a Gibbs sampling scheme for drawing samples from it. This analysis allows to have a complete view of the importance of the variables.

[2]The MAE and MSE are theoretically defined as the expectation of the (abosolute or squared) error over different realizations of the data $\mathbf{y}^{(\ell)}$ with $\ell = 1, 2, 3, ...$, where the index ℓ denotes a different realization. Here, we consider a fixed vector of data $\mathbf{y}$ (i.e., the observed data) and compute the error vector in one realization $\widehat{\mathbf{e}} = \mathbf{y} - \widehat{\mathbf{y}}$. Moreover, we average the absolute value (or the squared values) of error components in order to obtain a unique error value.

A. Ranking methods

In this section, we briefly describe the ranking methods (RMs) that we have applied to our dataset. Some of them are well-known techniques, whereas others contain some degree of novelty [15]. We list them below.

RM1 - Forward Selection (FS): *adding variables "forward" minimizing the error*. The method starts searching for the most significant univariable model (in terms of the error in prediction), i.e., considering the linear regression model in Eq. (1) with only one component (namely, one column in the matrix $\mathbf{X}$ in Eq. (2)). We repeat then considering a model with two variables (re-estimating the model for each pair), including (and keeping) the previously selected variable. We iterate the procedure until considering a complete model of R variables. This procedure provides a sequence of included variables that will be the final ranking.

RM2 - Backward Elimination (BE): *removing variables "backward" minimizing the error.* The method starts considering the complete model. Then, we remove the most insignificant variable in terms of the error in prediction, considering models with $R-1$ variables (clearly, removing a different variable and we re-estimate the coefficients for each model). We repeat the procedure considering always smaller models and removing one variable at each iteration. The procedure provides an inverse ranking where the first variable is the worst one and the last is the best one.

RM3: *removing variables maximizing the error.* The method starts again considering the complete model. Then, we remove one variable and compute the error in prediction. Hence, we select *the best variable*, i.e., which (when removed) produces the higher increase of the error in prediction. This variable will be the first in our ranking (the most relevant variable). We repeat the procedure considering the rest of variables.

RM4: *adding variables maximizing the error.* Here, we create a sequence of variables from the worst to the best variable (i.e., increasing their relevance), starting with a univariable model as in RM1, but selecting the worst variable (i.e., the variable which maximizes the prediction error). Then, we consider a model with two variables (keeping the previous select one) and select the second variable which maximizes the prediction error. We repeat the procedure, obtaining a final inverse ranking of the variable, i.e., the last one will be the most relevant variable.

RM5 - based on the correlation coefficient: We compute the Pearson correlation coefficients between one single variable and the output y. Then we rank the features in decreasing order according to the module of correlation coefficients. This procedure is similar to the often so-called *univariable selection* [15].

The joint use of these different ranking procedures allows to perform a robust analysis, obtaining a more complete view of our variable selection problem. Indeed, some ranking methods, although yield sequences far from the smallest possible error, detect relevant variables that appear also by the Gibbs sampling analysis (described below). More specifically,

although we will see that RM1 and RM2 provide the best performance in terms of prediction error, but the results of RM3, RM4 and RM5 reveal other important aspects shown by the rest of our analyses below. Moreover, for completing our view, we will also apply a classical ranking method based on p-values [9], [17], and show the results in Table IV.

B. Best sequence search

Let us define the vector of M different indices

$$\mathbf{v}_M = [k_1, ..., k_M], \qquad M \leq R,$$

where $k_i \in \{1, 2, ..., R\}$ but $k_i \neq k_j$ for $i \neq j$. Considering only the M variables in $\mathbf{v}_M$, we can build a smaller $N \times (M+1)$ design matrix $\mathbf{V}_M$, and consider a smaller $(M+1) \times 1$ vector of coefficients $\widehat{\boldsymbol{\beta}}_M = [\widehat{\beta}_1, ..., \widehat{\beta}_{M+1}]^\top$, which is obtained as

$$\widehat{\boldsymbol{\beta}}_M = \left(\mathbf{V}_M^\top \mathbf{V}_M\right)^{-1} \mathbf{V}_M^\top \mathbf{y}. \tag{5}$$

Moreover, we define the cost function

$$\begin{aligned} C(\mathbf{v}_M) &= ||\mathbf{y} - \widehat{\mathbf{y}}||_p^\alpha, \\ &= ||\mathbf{y} - \mathbf{V}_M \widehat{\boldsymbol{\beta}}_M||_p^\alpha, \end{aligned} \tag{6}$$

where $||\cdot||_p$ is the L_p norm with $p > 0$ and $\alpha > 0$. Note that $C(\mathbf{v}_M)$ is defined in the discrete space of M possible different indices. We desire to find the vector of indices such that

$$\mathbf{v}_M^* = \arg\min_{\mathbf{v}} C(\mathbf{v}_M). \tag{7}$$

Note that an exhaustive search is only possible when M is small (typically it is suitable only for $M \leq 4$). Moreover, a random search in the entire space (as with a simulated annealing approach [23]) can be very costly and to reach the global minimum (or a "good" local minimum) is very difficult. For this reason, we employ an alternating optimization approach that, at least, ensures a fast convergence to a local minimum. Furthermore, we perform the alternating optimization scheme several times (500 runs) with different initializations, and compare the minimum obtained at each run [23]. We finally consider the solution $\widehat{\mathbf{v}}_M$ with the smallest associate cost value $C(\widehat{\mathbf{v}}_M)$, i.e., $\widehat{\mathbf{v}}_M$ is our estimator $\mathbf{v}_M^*$. Below, we describe the alternating optimization method.

Alternating optimization. Choose $M < R+1$, a maximum number of iterations $T_{\texttt{iter}} \geq 1$, and start with $\mathbf{v}_M^{(0)}$.

For $t = 1, ..., T_{\texttt{iter}}$ (or until convergence) repeat:

1) For $j = 1, ..., M$:
   a) Keeping fixed the rest of $M-1$ variable, work only on the j-th variable, i.e., given
   $$\mathbf{b}_j = [k_1^{(t)}, k_2^{(t)}, ...k_{j-1}^{(t)}, k_j, k_{j+1}^{(t-1)} ..., k_M^{(t-1)}],$$
   find
   $$k_j^* = \arg\min_{\mathbf{v}} C(\mathbf{b}_j).$$
   The optimization above can be solved in an exhaustive way since it is a one-dimensional problem.
   b) Set $k_j^{(t)} = k_j^*$, and
   $$\mathbf{b}_{j+1} = [k_1^{(t)}, k_2^{(t)}, ...k_{j-1}^{(t)}, k_j^{(t)}, k_{j+1}, k_{j+2}^{(t-1)} ..., k_M^{(t-1)}].$$
2) Set
   $$\mathbf{v}_M^{(t)} = \mathbf{b}_M, \text{ and } \mathbf{b}_1 = \mathbf{v}_M^{(t)}.$$

C. A Gibbs sampling approach

We generalize the optimization scheme considering a Markov Chain Monte Carlo (MCMC) sampling approach. More specifically, we consider a Gibbs sampler which is the counterpart of the alternating optimization in the Monte Carlo sampling world [23]. The sampling approach (applied in this context) can provide the probability that each variable is contained in the best subset of M elements. Let us recall the vector of M different elements

$$\mathbf{v}_M = [v_1, ..., v_M],$$

where $v_i \in \{1, 2, ..., R\}$ but $v_i \neq v_j$ for $i \neq j$. In this section, we consider the target density

$$p(\mathbf{v}_M) \propto \exp\left(-\eta C(\mathbf{v}_M)\right), \qquad \eta > 0,$$

where $C(\mathbf{v}_M)$ is the cost function previously considered in Eq. (6). The constant η can be used and set to provide a tempering effect [21], [23], [27]. The variables that belong to sequences with smaller errors acquire more value according to $p(\mathbf{v}_M)$. Thus, by drawing samples from $p(\mathbf{v}_M)$, we can obtain the proportion of times that a feature provides a sequence with yields a small error. This is a very important information that helps us to yield a more *robust* analysis, in the sense that we can avoid overfitting at this specific set of data. The overfitting can occur performing the variable selection only considering the best sequence, for instance. Note that this idea can be employed in any problem where a cost function (as function of the parameters of interest) is available.

On the choice of η. We can observe that, as $\eta \to 0$, the density $p(\mathbf{v}_M)$ becomes closer and closer to a delta function around the best sequence $\mathbf{v}_M^*$ in Eq. (7), which is the global minimum of $C(\mathbf{v}_M)$. As $\eta > 0$ grows, more and more local modes appear in $p(\mathbf{v}_M)$. These local modes contain relevant information for our study. As $\eta \to \infty$, the difference among the values of the modes become smaller and smaller, and $p(\mathbf{v}_M)$ tends to a uniform density in support domain. It is important to remark that there is a range of suitable values of η such that the analysis can be performed. These suitable values are all the values of η such that all the possible local modes appear. The interested user can perform some preliminary runs for choosing a proper value of η.

A Gibbs sampling algorithm is a type of Markov Chain Monte Carlo (MCMC) method for drawing samples from general distribution as $p(\mathbf{v}_M)$ above [26], [28], [29]. An MCMC algorithm generates a Markov chain with invariant density exactly the target density, that in our case is $p(\mathbf{v}_M)$. A Gibbs sampler works at each step in a one dimensional space [23], simplifying the multivariate sampling problem drawing from simpler one-dimensional densities. Before describing

the Gibbs sampling method, we have to recall that the j-th *full-conditional density* is

$$p_j(v_j|v_1,...,v_{j-1},v_{j+1},...,v_M) \propto p(\mathbf{v}_M) = p(v_1,...,v_{j-1},v_j,v_{j+1},...,v_M), \quad (8)$$

where all the variables are fixed with the exception of v_j, and the normalizing constant is $p(v_1,...,v_{j-1},v_{j+1},...,v_M)$ that does not depend on v_j. For simplicity, we use the more compact notation

$$p_j(v_j|v_1,...,v_{j-1},v_{j+1},...,v_M) = p_j(v_j|v_{1:j-1},v_{j+1:M}),$$

for denoting the j-th full-conditional density. A detailed description of the Gibbs sampling algorithm is given below.

Gibbs sampler. Choose $M < R$, a maximum number of iterations $T_{\texttt{iter}} \geq 1$, and start with $\mathbf{v}_M^{(0)}$.

For $t = 1,\ldots,T_{\texttt{iter}}$:

1) For $j = 1,\ldots,M$:
   a) Draw $v_j^{(t)} \sim p_j\left(v_j \mid v_{1:j-1}^{(t)}, v_{j+1:M}^{(t-1)}\right)$.
2) Set $\mathbf{v}_M^{(t)} = \left[v_1^{(t)}, v_2^{(t)}, \ldots, v_M^{(t)}\right]$.

D. Selection of the number of relevant variables

Several selection procedures (also denoted as *stopping rules* during the ranking process) can be applied. Clearly, a naive method could be just to set a threshold value (or a percentage) for the prediction error, or by a simple visual inspection of the error curve, i.e., the so-called *elbow method* [36], [38]. In the classical variables selection analysis, practitioners and researchers often consider statistical tests (e.g., F-test and t-test), employed sequentially to decide whether individual variables should be included in the model, and a stopping rule based on p-values [9], [17].
Other approaches rely on the so called *information criteria* methods [21], [37], which are based on the following cost function

$$C(M) = \underbrace{-2\log p(\mathbf{y}|\widehat{\boldsymbol{\beta}}_M)}_{\text{fitting}} + \underbrace{2\xi M}_{\text{penalization}}, \quad (9)$$

where $\xi > 0$ is a constant that specifies the criterion. The first term is a fitting term (based on the maximum likelihood value), whereas the second term is a penalty for the model complexity. The expression (9) encompasses several well-known information criteria proposed in the literature and shown in Table I, which differ for the choice of η [21], [37].

Table I: Different information criterion for model selection (N number of data).

Criterion	Choice of ξ
Bayesian-Schwarz information criterion (BIC) [34]	$(\log N)/2$
Akaike information criterion (AIC) [35]	1
Hannan-Quinn information criterion (HQIC) [14]	$\log(\log(N))$

IV. Results for the output 1 - Arousal

In this section, we describe the results obtained for the first output (arousal). A more complete discussion is provided in the `Supplementary material`.

To carry out the analysis and selection of variables with which we propose our final linear model for the prediction of the arousal, we will follow the next steps:

1) First of all, we apply five ranking techniques (from RM1 to RM5) and analyze the variables that appear classified as the most significant (Section IV-A).
2) We apply the best sequence search (Section IV-B) and the Gibbs sampling analysis (Section IV-C) and analyze the variables that appear as the most relevant according to these methods that start from a fixed number of variables. We compare and discuss the obtained results with the results previously provided by the ranking techniques.
3) We summarize all the previous results by classifying the variables into three levels according to their level of global relevance (Section IV-D).
4) We propose the linear model for the prediction of the arousal including only the most relevant variables according to our analysis and we evaluate their prediction error (Section IV-E).

Note that these same steps are briefly replicated for the second output (valence) throughout Section V.

A. Results of the Ranking Methods

Figure 1 shows the MAE in the estimation of the first output considering models with $M \leq 122$ variables. The variables are ordered according to the different ranking criteria. At each M, we consider the first M variables in each ranking and compute the MAE. Clearly, when $M = 122$ (i.e., we are using all the variables) all the curves reach the same point. The black solid line corresponds to the MAE curves without ordering the variables (i.e., keeping the order in the data matrix). Note that, even in this curve with unordered variables, we can observe the importance of the variables "12-th chromagram standard deviation" (indexed as 112), "loudness mean" (indexed as 113) and "loudness standard deviation" (indexed as 114). Indeed, there is a relevant drop in MAE at the variable 112, the decrease continues with the variable 113, and the derivative seems to be null after the variable 114. Moreover, even in this curve with unordered variables, we can observe in Figure 1(a) that there is already an *elbow* within the interval between the 15-th variable and the 20-th variable [38].
The cyan and blue solid lines correspond to RM1 and RM2 which provides the best results in terms of MAE. Namely, RM1 and RM2 provide two orders of variables which produce the fastest decays in terms of MAE. Figure 1(b) provides the same information of Figure 1(a) but in log-log-scale. Both curves, corresponding to RM1 and RM2, seem to present a clear *elbow* between the 7-th and 8-th ordered variables.
The curves corresponding to RM3, RM4, RM5 are depicted with dashed lines. Although these rankings do not achieve the

best results in this figure, they provide interesting information regarding the importance of the variables, as we discussed below. For instance, note that the error with RM4 has a big drop (reaching the best performance given by RM2) when the third variable is considered, which is feature 1. This feature appears in 8-th position of RM5 but is not considered relevant by RM1, RM2, RM3 and by the best sequence search, as we will see later on. Moreover, the Gibbs analysis confirms its relevance (as we will see below).

The rankings of the variables obtained by RM1, RM2, RM3, RM4, RM5, from the first one to the 20-th one are given in Table II. We highlight with boxes the variables that are within the most important twenty variables in all the ranking methods; these variables are five and are labeled as:

113 ("loudness mean"),
114 ("loudness standard deviation"),
14 ("spread mean"),
13 ("centroid standard deviation"),
and 3 ("decrease slope mean"),

although, variable 3 is never contained within the most important ten variables within the different rankings.

B. Results of the best sequence search

In Table III, we can observe the best sequences for $M = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$, obtained with the alternative optimization procedure (after 10^3 independent runs with different random initializations). See also the decrease of the error in Figure 2(a). All the best sequences are given just in ascending order of the labels. Indeed, unlike with the Gibbs approach, we cannot discriminate among the variables within a best sequence.
In Table III, each new entry in the best sequence (as M grows - with respect to the previous shorter sequence), is highlighted with a box. We remark especially the best sequence with $M = 7$, i.e.,

4, ("maximum fluctuation"),
8 ("roll-off mean"),
14 ("spread mean"),
56 ("pitch standard deviation"), (10)
113 ("loudness mean"),
114 ("loudness standard deviation"),
and 115 ("energy mean"),
(shown here in ascending order of the labels).

They exactly coincide with the ranking given by RM2 of the first seven most important variables, i.e.,

$$113, 14, 8, 114, 115, 56, \text{ and } 4, \tag{11}$$

shown here in decreasing order of importance by RM2.

C. Gibbs sampling analysis

In the Gibbs sampling analysis, the sequences of length M, $\mathbf{v}_M = [v_1, ..., v_M]$ (with $v_i \in \{1, 2, ..., R+1\}$, $v_i \neq v_j$ for $i \neq j$), are weighted according to error $C(\mathbf{v}_M)$ or, more specifically, according to

$$p(\mathbf{v}_M) \propto \exp\left(-\eta C(\mathbf{v}_M)\right), \qquad \eta > 0. \tag{12}$$

In our simulation, we set $\eta = 100$. Moreover, for defining $C(\cdot)$, we consider the L_1 norm (and $\alpha = 1$). The variables that belongs to sequences with smaller errors acquire more weight/importance. In some sense, the Gibbs analysis provides the probability that a feature provides a sequence of small error. In Figure 3, we show the results of Gibbs sampling analysis for $M \in \{2, 6, 10, 20\}$. The dashed line depicts the equiprobability (uniform) distribution with probability $1/122$. Clearly, probabilities bigger than $1/122$ denote the most important variables.
We can observe that the results are coherent with the previous results above. For instance, the variable 113 is clear the most important and also the features 114, 4, 115, 56, 8 are quite relevant. The Gibbs analysis also confirms that the feature 39 seems relevant for small M but, as M grows, the importance of this feature disappears.
However, by the Gibbs analysis, we can obtain more interesting information. The features 4 and 56 are quite relevant even from small values of M, confirming also the results of the best sequence search. The features 14 seems to have a relevance very similar to 114. As we have also previously stated, the Gibbs analysis clearly shows that the variable 115 is the fourth most important feature (as we expect after a care look of the rankings). Surprisingly, the feature 1 seems to be equally relevant than the feature 115: we provide an explanation below. The variable 13 seems also to be some relevance by the Gibbs analysis specially as M grows. However, as we expected for the previous study, its relevance is moderate.
Furthermore, we can also observe the importance of other variables whose relevance was not clear from the previous analysis above. This is the case of the following features:

1, ("RMS mean"),
20, ("flatness mean"),
50, ("flux mean"),
16, ("skewness mean"),
and 22 ("entropy mean").

The feature 1 deserves some specific comments (see `Supplementary Material`). The feature 20 is in the fourth position of RM3, in the 10-th position of RM4, and in the 6-th position of RM5. The variable 50 appears also in the fifth position of RM4, and in the 10-th position of RM5. The variable 16 appears in the best sequences for $M \geq 9$ and in 9-th position of RM1. The feature 22 has not been detected by the previous studies: it does not appear either in the rankings, or in the best sequence search (at least for $M \leq 12$ as in Table III).

D. Summary for the output 1

Here, we classify the features into four different classes: *very relevant* (Level 1), *relevant* (Level 2), and *relevant but maybe only for the specific dataset* (Level 3), and the rest of variables belong to the class *non-relevant* (Level 4).

Level 1. After all the studies, we can assert at least 7 variables are very relevant:

$$113, 114, 14, 115, 4, 8, \text{ and } 56, \quad \text{(ordered by Gibbs analysis)},$$

which are shown in decreasing order of importance considering the Gibbs sampling analysis. This is also the best sequence for $M = 7$, as shown in Table III and includes also the first 7 elements in the ranking obtained by RM2 but with a different order,

$$113, 14, 8, 114, 115, 56, \text{ and } 4 \quad \text{(ordered by RM2)}.$$

Level 2. Other important variables are

$$1, 22, 20, 50, 13, 16, \text{ and } 3 \quad \text{(ordered by Gibbs analysis)},$$

but the features 1 ("RMS mean") and 50 ("flux mean") are highly correlated to the variable 115 ("energy mean"), as shown by Figure 2(b) and as we could intuitively expect. The variable 13 ("centroid standard deviation") appears in some of the first twelve best sequences. However, the Gibbs analysis reveals (that in terms of robustness) other variables such as 20 ("flatness mean") and 22 ("entropy mean") are more or a similar relevance than 13. The feature 20 is particularly important in RM3 (4-th position) and RM5 (6-th position). The feature 16 ("skewness mean") does not appears in the best first twenty variables in the rankings, but appears in the best sequences permanently for $M \geq 9$. In the Gibbs analysis, the feature 16 acquires some relevance as M grows. Finally, the feature 3 ("decrease slope mean") does not appear in the first twelve best sequences and it does not seems relevant by the Gibbs analysis. However, it appears within the first twenty more relevant variables in *all* the ranking methods.[3]

Level 3. Other possibly important variables, which appear in the best sequence search and in the rankings RM1 and RM2, are

43 ("7th MFCC stand. deviation")
and 107 ("7th chromagram stand. deviation").

The variable 43 appears in 8-th position in RM2 and 13-th position in RM1. Moreover, it appears in the best sequences for $M \geq 8$. The feature 43 appears in 11-th position in RM2 and 14-th position in RM1. Moreover, it appears in the best sequences for $M \geq 10$. On the other hand, the Gibbs analysis does not associate any particular relevance to these variables.

[3]More surprisingly, for the output 2 - valence -, only three features are included within the first twenty more relevant variables in *all* the ranking methods: they are the variables 113, 114 and again 3. Namely the feature 3 (as 113 and 114) is included within the first twenty more relevant variables in *all* the ranking methods for both outputs.

E. Selection of the number of variables and suggested model for the output 1

First of all, we have applied different information criteria, such as the AIC and BIC, shown in Table I [21]. The more adequate results have been provided the Bayesian information criterion (BIC) which suggests to use 17 variables whereas AIC suggests the use of 40 variables. We have also tried the classical analysis based on p-values which suggests 71 variables [9], [17]. The results of the corresponding ranking method is given in Table IV. The first most relevant 7 variables are again 113, 14, 8, 4, 56, 115 and 114, i.e., the *very relevant* features that we have found after our analysis.

However, after our exhaustive study, we believe that a more parsimonious model can be suggested. The most parsimonious LM after all the consideration in our study, is the model which includes at least the seven *very relevant* variables (described above),

$$113, 114, 14, 115, 4, 8, \text{ and } 56. \tag{13}$$

More precisely, the suggested linear model for the output 1 is

$$\begin{aligned} y_1 = &- 0.5293 + 3.6494\ x_{113} + 1.8080\ x_{114} + \\ &- 1.5534\ x_{14} - 3.8491\ x_{115} + \\ &+ 1.5056\ x_4 + 1.1714\ x_8 - 0.3450\ x_{56}, \end{aligned} \tag{14}$$

obtaining a MAE of 0.1593, MSE of 0.0432 (i.e, RMSE of 0.2078), and $R^2 = 0.8703$. Considering a Monte Carlo cross-validation procedure (with 80% of the data in the train-set and the rest of 20% of data in the test-set, chosen randomly in each $2 \cdot 10^4$ independent runs), we obtain MAE of 0.1611, MSE of 0.0450 (i.e, RMSE of 0.2118), and $R^2 = 0.8641$. Namely, we have a very slight increase of MAE and MSE (or a slight decrease of R^2), proving the robustness of our proposed model.

V. Results for the output 2 - Valence

In this section, we analyze briefly the results the output 2 of the dataset (valence). The decreases of the error for the RMs are shown in Figure 4. A complete discussion is provided in the `Supplementary Material`. Here, we also point out the the variables which seem relevant for both outputs (1 and 2) and some features just relevant for output 2.

A. Ranking, best sequences and Gibbs analysis

The most important features for output 2 are

114 ("loudness standard deviation"),
113 ("loudness mean"),
14 ("spread mean"),
and 3 ("decrease slope mean").

They are also relevant for the output 1 (as we can see in the main body of the work). The variable 114 seems to increase its relevance with respect the output 1, whereas the variable 3 is much more relevant for the output 2. The importance of these features is confirmed by the Gibbs analysis in Fig. 5, specially for the feature 14. Indeed, the case of feature 14 is very interesting and reveals also the importance of the Gibbs

analysis. The variable 14 does not seem relevant following RM1 and RM2 (which provides the sequences with the smaller errors) and does not appears in the best sequences in Table VII. However, it is the third more important variables for RM3, RM4 and RM5, and it is the third more relevant variable for the Gibbs analysis (see Fig. 6). Furthermore, it acquires more relevance as M grows (see again Fig. 6). The variables

1 ("RMS mean"),
and 115 ("energy mean"),

(which are highly correlated, also with the feature 50; see the main body of the paper for further details) appear relevant for the second output as well. The variable 1 is contained in the first twenty more important features in RM2, RM4 and RM5. Moreover, the variable 1 appears in the best sequences playing the role of the variable 115, i.e., when the feature 115 does not appear in those sequences. Namely, due to their correlation in the rankings and in the best sequences, the presence of one of them (1 or 115) avoids the presence of the other one. The feature 115 appears in the best sequences almost in a stable way for $M \geq 3$. The Gibbs analysis confirms the relevance of both 1 and 115, and they seem even more relevant than the variable 3. Furthermore, the following variables

50, ("flux mean"),
20, ("flatness mean"),
13, ("centroid standard deviation"),
and 8 ("roll-off mean"),

are also important for the output 2. The variable 50 seems relevant bit is highly correlated with the features 1 and 115. The variables above have certain relevance also for the output 1. Now, we discuss some features that seem to have importance only for the output 2 (i.e., valence). A careful look to the results reveals that the following features

88 ("inharmonicity standard deviation"),
31 ("8th MFCC mean"),
40 ("4th MFCC standard deviation"),
42 ("6th MFCC standard deviation"),
52 ("Low Energy"),
79 ("11th chromagram center stand. deviation"),
109 ("9th chromagram stand. deviation"),
and 110 ("10th chromagram standard deviation"),

are relevant, and appear in the best sequences for the output 2. Moreover, the feature 88 is the second more important in RM1 and appears in the 13-th position of RM3. The importance of 88 is confirmed (and is even more clear) by the Gibbs sampling analysis shown in Figure 5. The feature 31 seems to be relevant for RM1, RM5 and the Gibbs analysis. Moreover, it appears in the best sequences for $M \geq 11$. The variable 40 is within the first twenty most important variables of RM1 and RM2. It also appears in the ranking based on p-values in the 10-th position (see Table VI). Moreover, it starts to appear in the best sequences for $M \geq 10$. The importance of the feature 40 increases as M grows, following the Gibbs analysis. The feature 42 seems to have also the same importance of the variable 40 for the Gibbs analysis, and appears in the 16-th position of the GR and in the 15-th position of RM1. The variable 52 takes the 9-th position in RM2, appears in the best sequences for $M \geq 10$ and seems relevant according to the Gibbs analysis. The feature 79 appears within the first twenty more important variables in RM1, RM2 and RM3. From the Gibbs analysis, it seems clear that its importance grows with M. In the best sequences, the feature 79 appears in a stable way for all the best sequences with $M \geq 8$. The feature 110 is contained within the first twenty more important variables in RM1, RM2 and RM3. Following the Gibbs analysis, the feature 109 is similar or more relevant than 110. It also appears in the best sequence with $M = 7$ and in the 6-th position of the classical ranking based on p-values. The feature 50 seems relevant by the Gibbs analysis but it is very correlated to 1 and 115.

B. Selection of the number of variables and suggested model for the output 2

From the results, we can notice that the output 2 (valence) is less linear correlated with the variables $\mathbf{x}$, compared with the first output (arousal). The BIC suggests the use of 22 variables, whereas the AIC suggests the use of 68 variables. The classical p-values method suggests the use of 83 variables. However, all the considerations in our study above, we believe that a more parsimonious model can be proposed. In our opinion, The most parsimonious linear model that we can suggest is the model which includes at least the six *very relevant* variables which are (see the considerations above)

114 ("loudness standard deviation"),
113 ("loudness mean"),
14 ("spread mean"),
88 ("inharmonicity standard deviation"),
115 ("energy mean"),
3 ("decrease slope mean"),
(ordered by the Gibbs analysis),

and the ten *relevant* variables,

8, ("roll-off mean"),
20, ("flatness mean"),
79 ("11th chromagram center stand. deviation"),
4, ("maximum fluctuation"),
109 ("9th chromagram stand. deviation"),
110 ("10th chromagram standard deviation"),
40, ("4th MFCC standard deviation"),
31, ("8th MFCC mean"),
42, ("6th MFCC standard deviation"),
and 52 ("Low Energy"),
(ordered by the Gibbs analysis),

where we have included the feature 4 due to the Gibbs analysis. It appears also in the first twenty positions of RM3,

RM4 and RM5: in the 12-th position, in the 18-th position, and in the 15-th position, respectively. The variable 72 has been excluded due to the Gibbs analysis, as well. More precisely, the suggested model for the output 2 is

$$\begin{aligned} y_2 =& 0.2831 - 2.4741\ x_{114} - 1.0919\ x_{113} + 0.8070\ x_{14} + \\ & 0.2538\ x_{88} + 2.8482\ x_{115} - 0.6448 x_3 + \\ & - 1.4867\ x_8 + 1.1290\ x_{20} - 0.2003\ x_{79} + \\ & - 0.7192\ x_4 + 0.5182\ x_{109} + 0.1642\ x_{110} + \\ & 0.3312\ x_{40} + 0.2978\ x_{31} + 0.4621\ x_{42} - 0.5342\ x_{52}, \end{aligned} \tag{15}$$

obtaining a MAE of 0.2799, MSE of 0.1182 (i.e, an RMSE of 0.3437), and an $R^2 = 0.6452$. Considering a Monte Carlo cross-validation procedure (with 80% of the data in the train-set and the rest of 20% of data in the test-set, chosen randomly in each $2 \cdot 10^4$ independent runs), we obtain MAE of 0.2849, MSE of 0.1233 (i.e, RMSE of 0.3509), and $R^2 = 0.6311$. As for the output 1, we have a very slight increase of MAE and MSE (and a slight decrease of R^2), proving the robustness of our proposed model.

VI. Final discussion and conclusions

From the previous section, we observe that within the most important features for both outputs, 1 and 2, are

$$114, 113, 14, 115, \text{ and } 3, \text{ (ordered by Gibbs analysis).}$$

Therefore, we can conclude that the psychoacoustic features 113, 114 ("loudness mean" and "loudness standard deviation", respectively), the frequency-domain feature 14 ("spread mean"), and the time-domain feature 115 ("energy mean") are the most relevant variables in our study. They have been included in both suggested models. The frequency-domain feature 3 ("decrease slope mean"), although has not been included in the suggested model for the output 1, appears within the first twenty more relevant variables in all the rankings, for both outputs. The relevant features reveal the importance of the psychoacoustic indicators in SER. However, time and frequency-domain features have been also included into the suggested models. These results are in line with other studies which also highlight that subjective perception and time-dynamics of the signals (jointly embedded in indicators) lead to better model scores [1], [20]. The valence model provides worse performance than arousal one, even involving more features. This outcome agrees with the literature and it seems to be due to the prevalence of neutral annotations of valence in some soundscape categories [11].

We remark that the suggested LMs are very cheap and parsimonious models (including only 7 variables for arousal and 16 for valence, over the 122 possible features) and provide quite high R^2 coefficients ($R^2 = 0.8703$ for arousal and $R^2 = 0.6452$ for valence), and small MSEs (0.045 for arousal and 0.118 for valence), compared with the results previously obtained in the literature, even using non-linear models and including more variables. Indeed, our results are competitive with respect to the EMO baseline that employs a non-linear SVR and many more features (exactly 39 variables), both in arousal (with an MSE of 0.048) and valence (with an MSE of 0.128) [10]. In [2], the authors suggest also linear models with EMO obtaining worse results: specifically MSE ≈ 0.090 for arousal and MSE ≈ 0.160 for valence) using also more features in their models (exactly 25). Recent studies with EMO have shown that more sophisticated nonlinear models (such as RF) can reach good scores with 15 features for arousal (MSE ≈ 0.050) and 14 features for valence (MSE ≈ 0.140). Finally, other authors using other complex nonlinear models, such as CNNs and data augmentation techniques, obtain slightly better metrics (MSE ≈ 0.035 for arousal, and MSE ≈ 0.078 for valence), but also including substantially more variables in their models: from 23 up to 54 features [11], [31]. All these considerations confirm the quality of our suggested models.

Due to the exhaustive study that we have performed, we believe that the suggested LMs are robust and allow good prediction in different databases. Thus, the obtained parsimonious models can help the design of SER methods, and its practical applications by remarking the most relevant features. As future research lines, we plan to extend our variable selection study (including the proposed Gibbs analysis) for nonlinear models, and then judge if this non-linearity is strongly required with the EMO dataset, since the proposed LMs provides already very good performance. Furthermore, we plan to design novel schemes for selecting automatically a reasonable number of variables, when the priority is to obtain the simplest (and hence cheapest) possible model. Indeed, at least with these soundscapes data, the current benchmark techniques often seem to widely overestimate the adequate number of relevant variables.

Acknowledgments

The authors acknowledge support by the Agencia Estatal de Investigación AEI (project SPGRAPH, ref. num. PID2019-105032GB-I00), by Young Researchers R&D Project with ref. num. F861 (AUTO-BA-GRAPH) funded by Community of Madrid and Rey Juan Carlos University and F. Llorente acknowledges support by Spanish government via grant FPU19/00815.

References

[1] Faranak Abri, Luis Felipe Gutiérrez, Prerit Datta, David RW Sears, Akbar Siami Namin, and Keith S Jones. A comparative analysis of modeling and predicting perceived and induced emotions in sonification. Electronics, 10(20):2519, 2021.

[2] Faranak Abri, Luis Felipe Gutiérrez, Akbar Siami Namin, David RW Sears, and Keith S Jones. Predicting emotions perceived from sounds. In 2020 IEEE International Conference on Big Data (Big Data), pages 2057–2064. IEEE, 2020.

[3] European Environmental Agency. Good practice guide on quiet areas. Technical report no. 4, 2014.

[4] Francesco Aletta, Jian Kang, and Östen Axelsson. Soundscape descriptors and a conceptual framework for developing predictive soundscape models. Landscape and Urban Planning, 149:65–74, 2016.

[5] Francesco Aletta and Jieling Xiao. What are the current priorities and challenges for (urban) soundscape research? Challenges, 9(1):16, 2018.

[6] Pierre Aumond, Arnaud Can, Bert De Coensel, Dick Botteldooren, Carlos Ribeiro, and Catherine Lavandier. Modeling soundscape pleasantness using perceptual assessments and acoustic measurements along paths in urban context. Acta Acustica united with Acustica, 103(3):430–443, 2017.

[7] Östen Axelsson, Mats E Nilsson, and Birgitta Berglund. A principal components model of soundscape perception. The Journal of the Acoustical Society of America, 128(5):2836–2846, 2010.

[8] William J Davies, Neil S Bruce, and Jesse E Murphy. Soundscape reproduction and synthesis. Acta Acustica United with Acustica, 100(2):285–292, 2014.

[9] M.A. Efroymson. Multiple regression analysis. Mathematical methods for digital computers, pages 191–203, 1960.

[10] Jianyu Fan, Miles Thorogood, and Philippe Pasquier. Emo-soundscapes: A dataset for soundscape emotion recognition. In 2017 Seventh international conference on affective computing and intelligent interaction (ACII), pages 196–201. IEEE, 2017.

[11] Jianyu Fan, Fred Tung, William Li, and Philippe Pasquier. Soundscape emotion recognition via deep learning. Proceedings of the Sound and Music Computing, 2018.

[12] André Fiebig, Pamela Jordan, and Cleopatra Christina Moshona. Assessments of acoustic environments by emotions–the application of emotion theory in soundscape. Frontiers in Psychology, 11:3261, 2020.

[13] Eduardo Fonseca, Jordi Pons Puig, Xavier Favory, Frederic Font Corbera, Dmitry Bogdanov, Andres Ferraro, Sergio Oramas, Alastair Porter, and Xavier Serra. Freesound datasets: a platform for the creation of open audio datasets. In Hu X, Cunningham SJ, Turnbull D, Duan Z, editors. Proceedings of the 18th ISMIR Conference; 2017 oct 23-27; Suzhou, China.[Canada]: International Society for Music Information Retrieval; 2017. p. 486-93. International Society for Music Information Retrieval (ISMIR), 2017.

[14] E. J. Hannan and B. G. Quinn. The determination of the order of an autoregression. Journal of the Royal Statistical Society. Series B (Methodological), 41(2):190–195, 1979.

[15] G. Heinze, C. Wallisch, and D. Dunkler. Variable selection - a review and recommendations for the practicing statistician. Biometrical journal, 60(3):431–449, 2018.

[16] Karmele Herranz-Pascual, Igone García, Itziar Aspuru, Itxasne Díez, and Álvaro Santander. Progress in the understanding of soundscape: objective variables and objectifiable criteria that predict acoustic comfort in urban places. Noise Mapping, 3(1), 2016.

[17] R. R. Hocking. The analysis and selection of variables in linear regression. Biometrics, pages 1–49, 1976.

[18] Xinchen Hong, Yu Jiang, Shuting Wu, Linying Zhang, and Siren Lan. Study on evaluation model of soundscape in urban park based on radial basis function neural network: A case study of shiba park and kamogawa park, japan. In IOP Conference Series: Earth and Environmental Science, volume 300, page 032036. IOP Publishing, 2019.

[19] Olivier Lartillot, Petri Toiviainen, and Tuomas Eerola. A matlab toolbox for music information retrieval. In Data analysis, machine learning and applications, pages 261–268. Springer, 2008.

[20] Matteo Lionello, Francesco Aletta, and Jian Kang. A systematic review of prediction models for the experience of urban soundscapes. Applied Acoustics, 170:107479, 2020.

[21] F. Llorente, L. Martino, D. Delgado, and J. Lopez-Santiago. Marginal likelihood computation for model selection and hypothesis testing: an extensive review. arXiv:2005.08334, 2020.

[22] Phil Lopes, Antonios Liapis, and Georgios N Yannakakis. Modelling affect for horror soundscapes. IEEE Transactions on Affective Computing, 10(2):209–222, 2017.

[31] Stavros Ntalampiras. Emotional quantification of soundscapes by learning between samples. Multimedia Tools and Applications, 79(41):30387–30395, 2020.

[23] D. Luengo, L. Martino, M. Bugallo, V. Elvira, and Sarkka S. A survey of monte carlo methods for parameter estimation. EURASIP J. Adv. Signal Process., 25:1–62, 2020.

[24] Peter Lundén and Malin Hurtig. On urban soundscape mapping: A computer can predict the outcome of soundscape assessments. In INTER-NOISE and NOISE-CON Congress and Conference Proceedings, volume 253, pages 2017–2024. Institute of Noise Control Engineering, 2016.

[25] Weiyi Ma and William Forde Thompson. Human emotions track changes in the acoustic environment. Proceedings of the National Academy of Sciences, 112(47):14563–14568, 2015.

[26] L. Martino, V. Elvira, and G. Camps-Valls. The recycling Gibbs sampler for efficient learning. Digital Signal Processing, 74:1–13, 2018.

[27] L. Martino, F. Llorente, E. Curbelo, Javier Lopez-Santiago, and J. Miguez. Automatic tempered posterior distributions for bayesian inversion problems. Mathematics, 9(7), 2021.

[28] L. Martino, J. Read, and D. Luengo. Independent doubly adaptive rejection Metropolis sampling within Gibbs sampling. IEEE Transactions on Signal Processing, 63(12):3123–3138, June 2015.

[29] L. Martino, H. Yang, D. Luengo, J. Kanniainen, and J. Corander. A fast universal self-tuned sampler within Gibbs sampling. Digital Signal Processing, 47:68 – 83, 2015.

[30] Benoit Mathieu, Slim Essid, Thomas Fillon, Jacques Prado, and Gaël Richard. Yaafe, an easy to use and efficient audio feature extraction software. In ISMIR, pages 441–446. Citeseer, 2010.

[32] James A Russell. A circumplex model of affect. Journal of personality and social psychology, 39(6):1161, 1980.

[33] R Murray Schafer. The soundscape: Our sonic environment and the tuning of the world. Simon and Schuster, 1993.

[34] G. Schwarz et al. Estimating the dimension of a model. The annals of statistics, 6(2):461–464, 1978.

[35] D.J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van der Linde. Bayesian measures of model complexity and fit. J. R. Stat. Soc. B, 64:583–616, 2002.

[36] P Stoica and Y Selén. Cross-validation rules for order estimation. Digital Signal Processing, 14:355–371, 2004.

[37] P Stoica and Y Selén. Model-order selection: a review of information criterion rules. IEEE Signal Processing Magazine, pages 36–47, 2004.

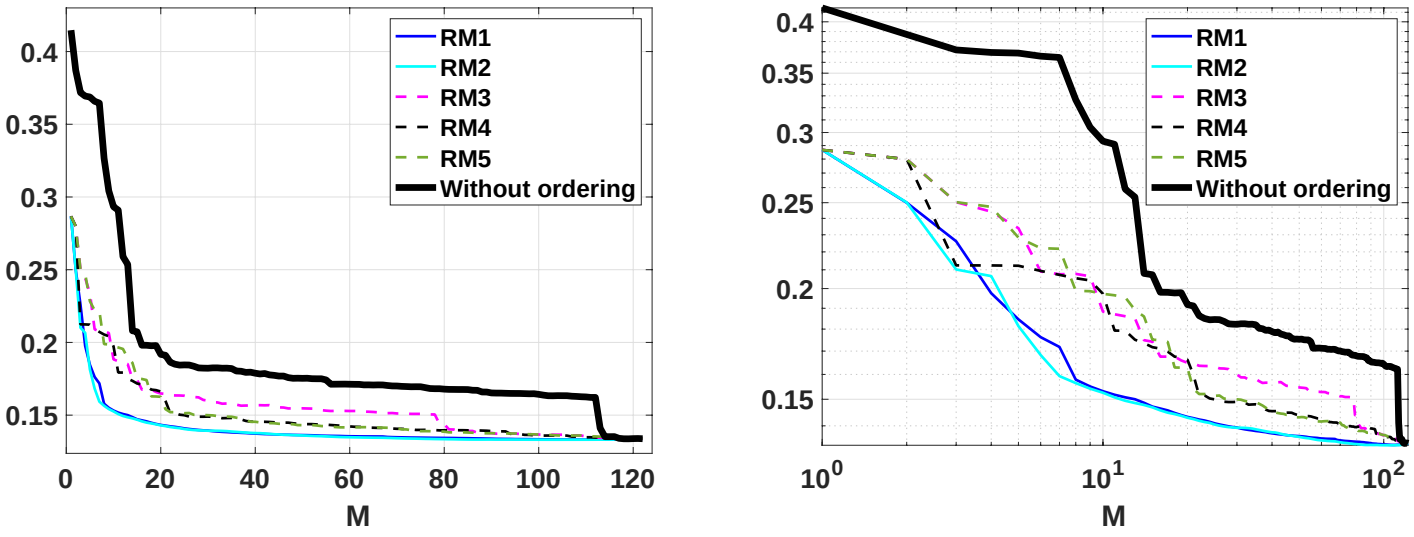
[38] R. L. Thorndike. Who belongs in the family? Psychometrika, 3:267–276, 1953.

[39] Kirsten A-M Van den Bosch, David Welch, and Tjeerd C Andringa. The evolution of soundscape appraisal through enactive cognition. Frontiers in psychology, 9:1129, 2018.

[40] Daniel Västfjäll, Mendel Kleiner, and Tommy Gärling. Affective reactions to interior aircraft sounds. Acta Acustica united with Acustica, 89(4):693–701, 2003.

[41] World Health Organization. Environmental Noise Guidelines for the European Region. WHO Regional Office for Europe, 1st edition, 2018.

[42] Chunyang Xu and Jian Kang. Soundscape evaluation: Binaural or monaural? The Journal of the Acoustical Society of America, 145(5):3208–3217, 2019.



(a) MAE versus M, ordering the variables.

(b) The same of Figure (a) but in log-log scale.

Figure 1: **(Output 1)** MAE versus M obtained ordering the variables according to the different rankings. At each M, we consider the first M variables in each ranking and compute the MAE. Clearly, when $M = 122$ (i.e., we are using all the variables) all the curves reach the same point. The black solid line corresponds to the MAE curves without ordering the variables.

Table II: **Results of the ranking methods - Output 1**

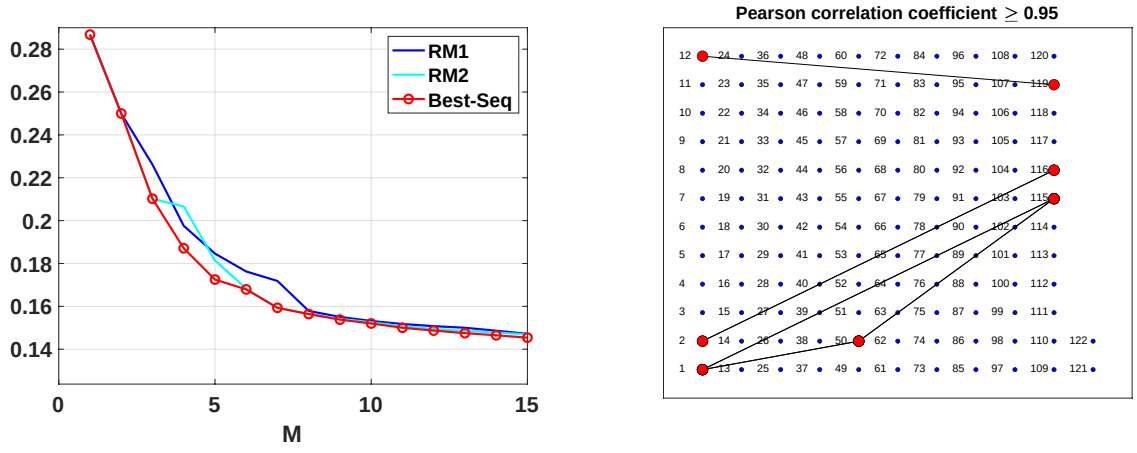
Meth.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
RM1	113	39	14	8	4	56	115	114	16	13	37	77	43	107	55	3	11	88	38	48
RM2	113	14	8	114	115	56	4	43	18	13	107	55	11	88	77	38	3	59	81	78
RM3	113	114	14	20	18	119	13	12	9	8	21	3	15	56	88	4	7	65	38	39
RM4	113	114	1	115	50	116	2	51	14	20	119	12	8	13	9	21	117	18	15	3
RM5	113	114	14	116	2	20	51	1	115	50	13	9	3	21	4	122	29	56	31	23

Table III: **Best Sequences - Output 1**

M	**Labels of the features in the best sequence**											
1	113											
2	39	113										
3	8	14	113									
4	4	8	14	113								
5	10	56	113	114	115							
6	10	13	56	113	114	115						
7	4	8	14	56	113	114	115					
8	4	8	14	43	56	113	114	115				
9	4	8	14	16	43	56	113	114	115			
10	4	8	14	16	43	56	107	113	114	115		
11	4	8	13	14	16	43	56	107	113	114	115	
12	4	8	13	14	16	43	55	56	107	113	114	115

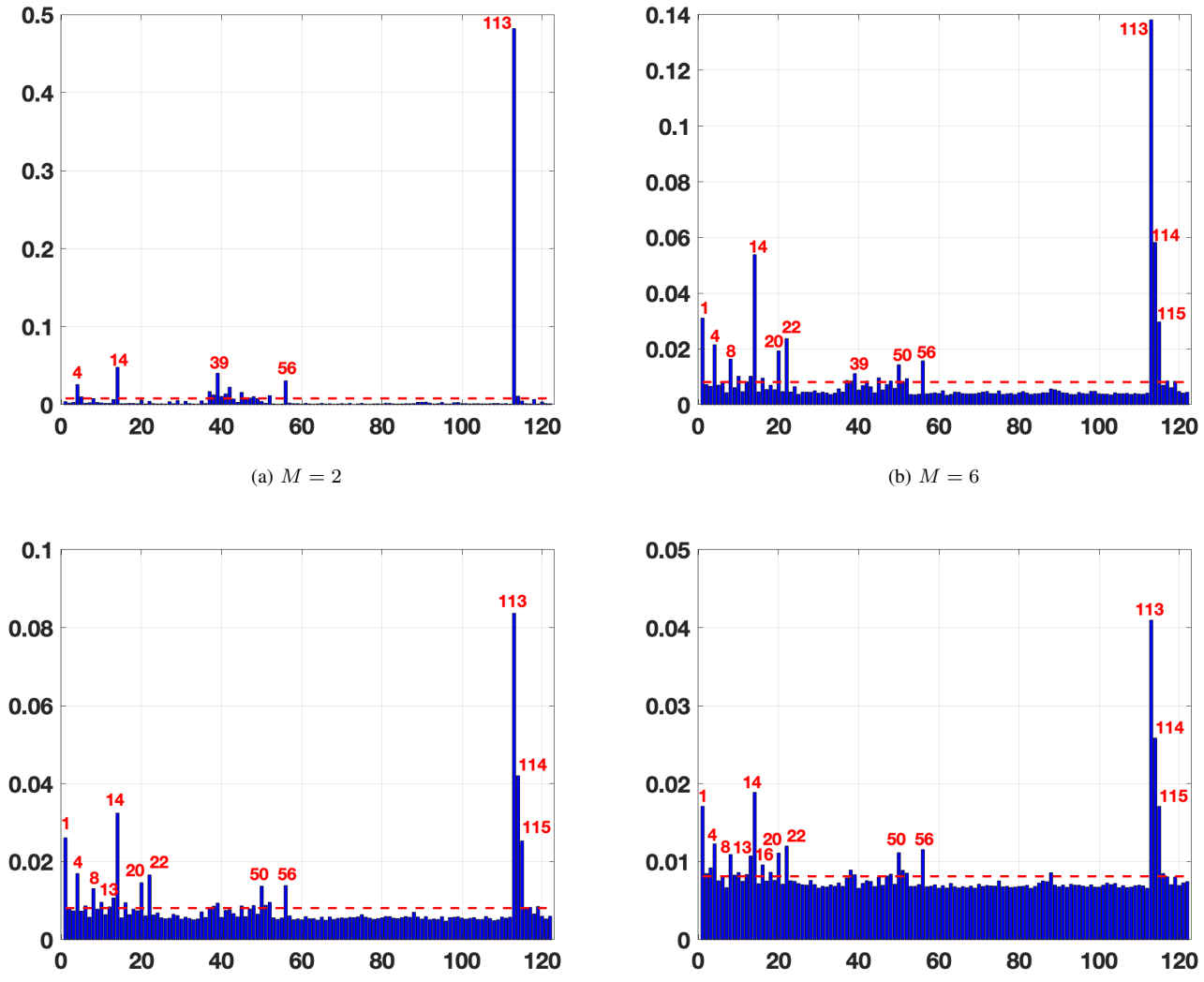
Table IV: **Results of RM based on p-values - Output 1**

Ranking Method	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
RM based on p-values	113	14	8	4	56	115	114	16	43	75	55	13	37	122	2	107	11	40	3	38



(a) (b)

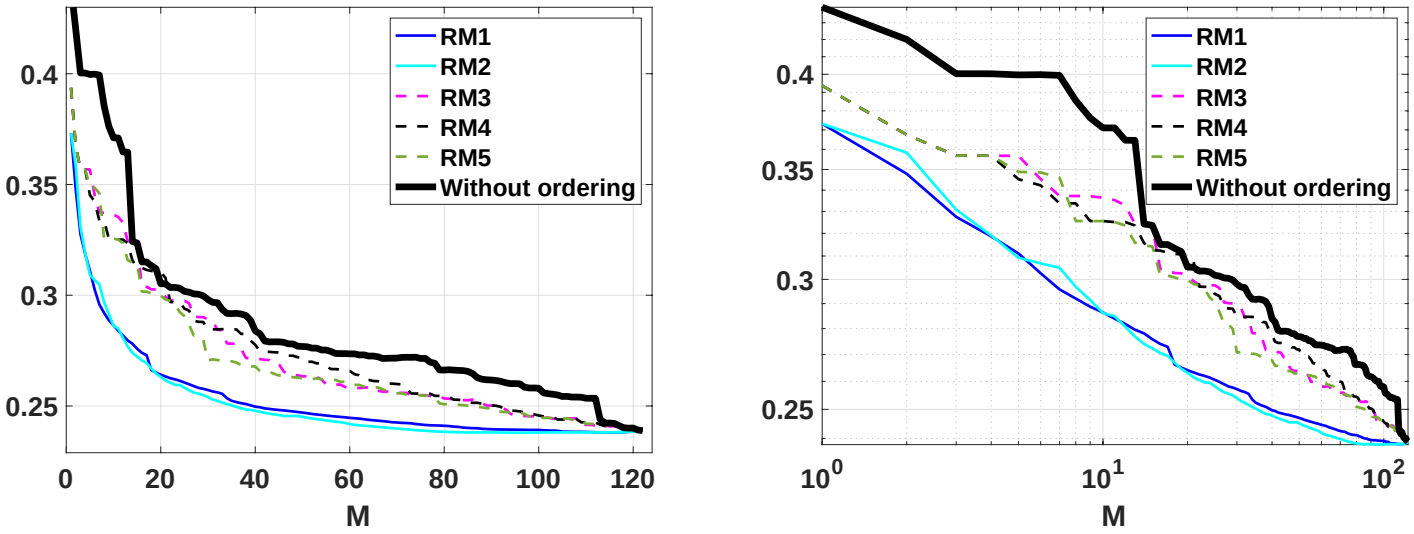
Figure 2: **(Output 1) (a)** MAE versus M obtained ordering the variables according to RM1, RM2 and the best sequence search. **(b)** The connections among the variables with the Pearson correlation coefficient ρ such that $|\rho| \geq 0.95$.



(a) $M = 2$ (b) $M = 6$

(c) $M = 10$ (d) $M = 20$

Figure 3: **(Output 1)** Results in terms of probabilities obtained by a Gibbs sampling analysis. The dashed line depicts the equiprobability (uniform) distribution with probability $1/122$.



(a) MAE versus M, ordering the variables.

(b) The same of Figure (a) but in log-log scale.

Figure 4: **(Output 2)** MAE versus M obtained ordering the variables according to the different rankings. At each M, we consider the first M variables in each ranking and compute the MAE. Clearly, when $M = 122$ (i.e., we are using all the variables) all the curves reach the same point. The black solid line corresponds to the MAE curves without ordering the variables.

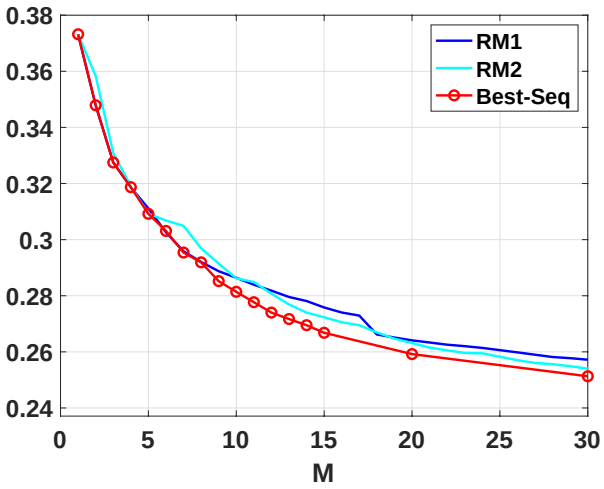


Figure 5: **(Output 2)** MAE versus M obtained ordering the variables according to RM1, RM2 and the best sequence search.

Table V: **Results of the ranking methods - Output 2**

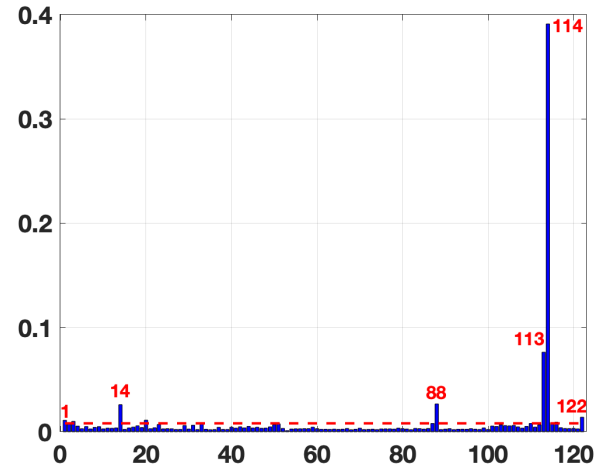
Meth.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
RM1	114	88	115	3	42	5	113	109	64	31	59	15	9	121	13	36	79	75	40	110
RM2	114	1	113	110	3	79	94	62	52	72	12	14	40	15	10	101	91	92	29	67
RM3	113	114	14	20	18	3	119	9	13	65	94	4	88	86	79	110	101	121	34	42
RM4	113	114	14	20	50	51	2	116	1	115	18	9	13	3	119	19	17	4	12	34
RM5	113	114	14	116	2	20	51	1	115	50	13	9	3	21	4	122	29	56	31	23

Table VI: **Results of RM based on p-values - Output 2**

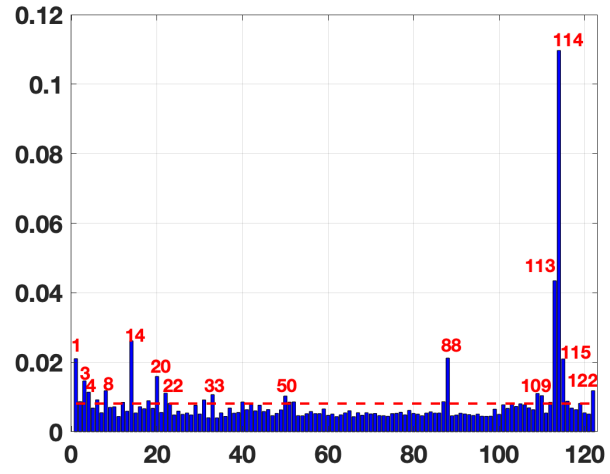
Ranking Method	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
RM based on p-values	114	88	115	3	113	109	5	25	65	40	33	37	4	122	68	107	36	2	51	47

Table VII: **Best Sequences - Output 2**

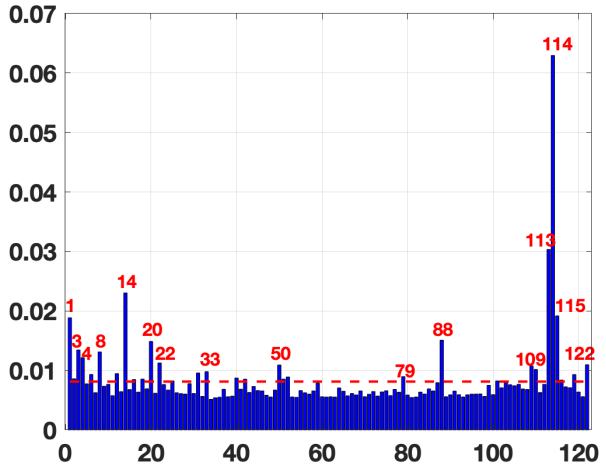
M	**Labels of the features in the best sequence**											
1	114											
2	[88]	114										
3	88	114	[115]									
4	[3]	88	114	115								
5	3	[104]	[113]	114	115							
6	[33]	[40]	113	114	115	[119]						
7	[3]	[5]	[88]	[109]	113	114	115					
8	3	5	[72]	[79]	[110]	113	114	115				
9	3	5	72	79	[88]	110	113	114	115			
10	3	5	[40]	72	79	88	[112]	113	114	115		
11	3	[31]	40	[52]	72	79	[91]	[110]	113	114	115	
12	[1]	3	31	40	52	72	79	[88]	91	110	113	114



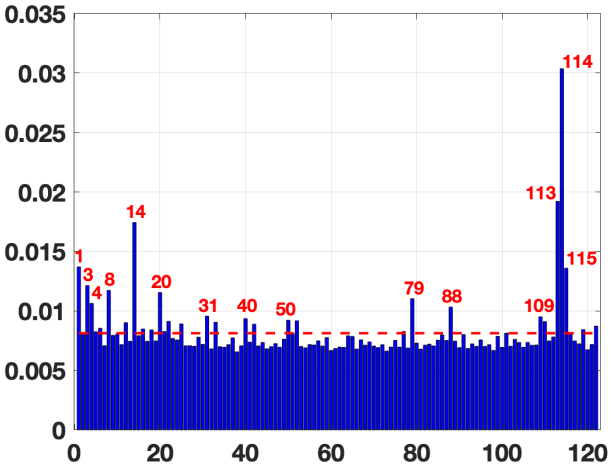
(a) $M = 2$



(b) $M = 6$



(c) $M = 10$



(d) $M = 20$

Figure 6: **(Output 2)** Results in terms of probabilities obtained by a Gibbs sampling analysis. The dashed line depicts the uniform discrete distribution with probability $1/122$.

Supplementary Material of "An exhaustive variable selection study for linear models of soundscape emotions: rankings and Gibbs analysis"

R. San Millán-Castillo*, L. Martino*, E. Morgado*, F. Llorente†
* Dep. de Teoría de la Señal y Comunicaciones, Universidad Rey Juan Carlos (URJC), Madrid, Spain.
† Dep. de Estadística, Universidad Carlos III de Madrid (UC3M), Leganés (Madrid), Spain.

I. Results for the output1 - Arousal

We provide a detailed description of the results obtained for the first output "arousal", and for the second output "valence".

A. Results of the Ranking Methods

Figure 1 shows the MAE in the estimation of the first output considering models with $M \leq 122$ variables. The variables are ordered according to the different ranking criteria. At each M, we consider the first M variables in each ranking and compute the MAE. Clearly, when $M = 122$ (i.e., we are using all the variables) all the curves reach the same point. The black solid line corresponds to the MAE curves without ordering the variables (i.e., keeping the order in the data matrix). Note that, even in this curve with unordered variables, we can observe the importance of the variables "12-th chromagram standard deviation" (indexed as 112), "loudness mean" (indexed as 113) and "loudness standard deviation" (indexed as 114). Indeed, there is a relevant drop in MAE at the variable 112, the decrease continues with the variable 113, and the derivative seems to be null after the variable 114. Moreover, even in this curve with unordered variables, we can observe in Figure 1(a) that there is already an *elbow* within the interval between the 15-th variable and the 20-th variable [**?**].

The cyan and blue solid lines correspond to RM1 and RM2 which provides the best results in terms of MAE. Namely, RM1 and RM2 provide two orders of variables which produce the fastest decays in terms of MAE. Figure 1(b) provides the same information of Figure 1(a) but in log-log-scale. Both curves, corresponding to RM1 and RM2, seem to present a clear *elbow* between the 7-th and 8-th ordered variables.

The curves corresponding to RM3, RM4, RM5 are depicted with dashed lines. Although these rankings do not achieve the best results in this figure, they provide interesting information regarding the importance of the variables, as we discussed below. For instance, note that the error with RM4 has a big drop (reaching the best performance given by RM2) when the third variable is considered, which is feature 1. This feature appears in 8-th position of RM5 but is not considered relevant by RM1, RM2, RM3 and by the best sequence search, as we will see later on. Moreover, the Gibbs analysis confirms its relevance (as we will see below).

The rankings of the variables obtained by RM1, RM2, RM3, RM4, RM5, from the first one to the 20-th one are given in Table I. We highlight with boxes the variables that are within the most important twenty variables in all the ranking methods; these variables are five and are labeled as:

113 ("loudness mean"),
114 ("loudness standard deviation"),
14 ("spread mean"),
13 ("centroid standard deviation"),
and 3 ("decrease slope mean"),

although, variable 3 is never contained within the most important ten variables within the different rankings. This is also remarked by Table II where we construct a *global ranking* (GR) considering the positions of each variable in the intermediate rankings RM1, RM2, RM3, RM4 and RM5. For building the GR, we consider the intermediate rankings with equal value and assign to each of the first variables a score of 122, to the second variable a score of 121, to the third variable a score of 120 and so on until assigning to the last variable a score of 1. The global score is obtained by summing all the intermediate scores. In Table II, we show a normalized score dividing the actual score by the maximum possible global score, that is $122 \cdot 5 = 610$. This maximum global score is achieved by the variable indexed as 113 (indeed, it has a normalized score of 1). By the GR appears even more clear that

$$113, 114, \text{ and } 14,$$

are the three most important features (shown in decreasing order of importance) for the first output "arousal". Moreover, from Tables I-II, we can see that the variables

4 ("maximum fluctuation"),
56 ("pitch standard deviation"),
8 ("roll-off mean"),
115 ("energy mean"),

deserve a special mention as well, since they are out from the first best 20 variables only in one ranking. The variable 4 is out of the first 20 best variables only in RM4 (where is in position 21). The feature 56 is out of the first 20 best variables only in RM4 (where is in position 22). The feature 8 is out of the first 20 best variables only in RM5 (where is in position 21). The variable 115 is within the first 10 best variables, except for the RM4 where takes the position 86. Hence, removing RM4, the 115-th variable seems even more relevant that 3-th variable and even better than 8-th variable (it would have the position 4 in the GR, do not considering RM4). Furthermore, other variables deserve a mention, for instance,

$$\begin{aligned} &1 \text{ (“RMS mean”),} \\ &20 \text{ (“flatness mean”),} \\ &\text{and } 18 \text{ (“kurtosis mean”).} \end{aligned}$$

As we previously stated, the variable 1 does not appear in the first best 20 variables on the GR, but is the third variable by RM4 and is in position 8 by RM5. Moreover, the variable 1 appears as a particularly relevant in the Gibbs sampling study (see below). The variable 20 appears in position 4 and 6 in the ranking by RM3 and RM5, respectively. Again, as variable 1, the feature 20 seems to have some relevance as shown by the Gibbs analysis below. Finally, the variable 18 appears in position 9 and 5 in the ranking by RM2 and RM3, respectively.

B. Results of the best sequence search

In Table III, we can observe the best sequences for $M = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$, obtained with the alternative optimization procedure (after 10^3 independent runs with different random initializations). All the best sequences are given just in ascending order of the labels. Indeed, unlike with the Gibbs approach, we cannot discriminate among the variables within a best sequence.

In Table III, each new entry in the best sequence (as M grows - with respect to the previous shorter sequence), is highlighted with a box. We remark especially the best sequence with $M = 7$, i.e.,

$$\begin{aligned} &4, \text{ (“maximum fluctuation”),} \\ &8 \text{ (“roll-off mean”),} \\ &14 \text{ (“spread mean”),} \\ &56 \text{ (“pitch standard deviation”),} \\ &113 \text{ (“loudness mean”),} \\ &114 \text{ (“loudness standard deviation”),} \\ &\text{and } 115 \text{ (“energy mean”),} \\ &\text{(shown here in ascending order of the labels).} \end{aligned} \tag{1}$$

They exactly coincide with the ranking given by RM2 of the first seven most important variables, i.e.,

$$\begin{aligned} &113, 14, 8, 114, \boxed{115}, 56, \text{ and } 4, \\ &\text{(shown here in decreasing order of importance by RM2),} \end{aligned} \tag{2}$$

where we have highlighted the feature 115 for comparing below with the results of the GR. In the best sequence of 8 features in Table III contains the first most important variables of RM2. Note also that the sequence of the seven most relevant features obtained by the GR, in Table II, is

$$\begin{aligned} &113, 114, 14, 8, \boxed{13}, 4, \text{ and } 56, \\ &\text{(shown in decreasing order of importance by GR),} \end{aligned} \tag{3}$$

where we can find all the elements of the best sequence in Eq. (1) with exception of the feature 115 (“energy mean”), replaced by the feature 13 (which represents the “centroid standard deviation”). However, as we have already remarked above, the variable 115 is penalized by the its bad position in RM3. Removing the ranking of RM3, the feature 115 would be in the fourth position of the GR. Note also that the feature 115 appears in all the best sequences for $M \geq 5$. This confirms that a more fair position for the feature 115 in a GR would be around fourth and fifth position. The variable 13 appears, in a *stable way*, for the best sequences with $M \geq 11$ (its first appearance is for $M = 6$, but then disappears until $M = 11$). The fact that the feature 115 is more important than the feature 13 is also confirmed the Gibbs analysis (see below).

From Table III, we can also observe that the variables 56 (“pitch standard deviation”) and 4 (“maximum fluctuation”) seem to be even more relevant than the feature 13, appearing permanently in all the best sequences with $M \geq 5$ and $M \geq 7$, respectively (the first appearance of the variable 4 is for $M = 4$). The features 8 (“roll-off mean”) and 14 (“spread mean”) appear firstly with $M = 3, 4$ but permanently only with $M \geq 7$. The variable 113 appears for all the best sequences with $M \geq 1$, confirming that is the most important feature. The feature 114 appears permanently with $M \geq 4$. In best sequence with $M = 2$, we have the variable 39 (that represents the “3rd MFCC standard deviation”): this feature appears in the 16-th position of GR, and the second position in RM1. However, the variable 39 is not included in any of the best sequences with $M > 2$.

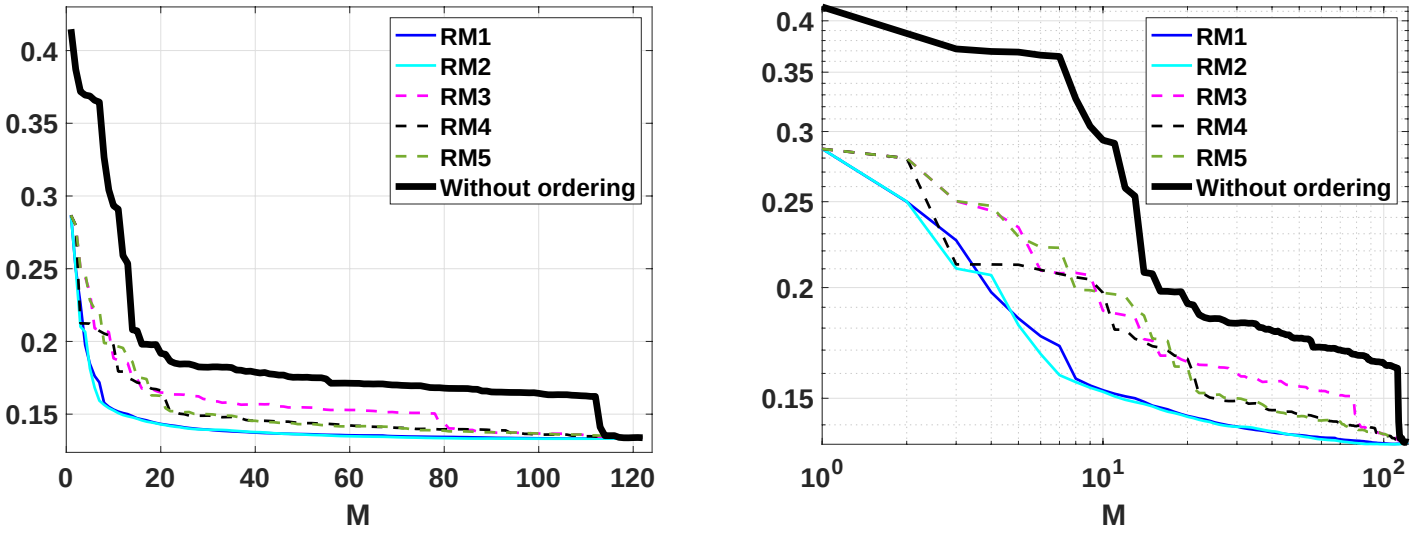
It is important to observe that the results obtained with different methodologies are coherent, and are also verified and clarified (in some cases) by the Gibbs analysis, as we will see below. Finally, in Figure 2(a), we can see the decay of the MAE as M grows according to the best sequences. We can see that RM1 and RM2 provide MAEs very close to the MAE of the best sequences.

C. Gibbs sampling analysis

In the Gibbs sampling analysis, the sequences of length M, $\mathbf{v}_M = [v_1, ..., v_M]$ (with $v_i \in \{1, 2, ..., R+1\}$, $v_i \neq v_j$ for $i \neq j$), are weighted according to error $C(\mathbf{v}_M)$ or, more specifically, according to

$$p(\mathbf{v}_M) \propto \exp\left(-\eta C(\mathbf{v}_M)\right), \qquad \eta > 0. \tag{4}$$

In our simulation, we set $\eta = 100$. Moreover, for defining $C(\cdot)$, we consider the L_1 norm (and $\alpha = 1$). The variables that belongs to sequences with smaller errors acquire more weight/importance. In some sense, the Gibbs analysis provides the probability that a feature provides a sequence of small error. In Figure 3, we show the results of Gibbs sampling analysis for $M \in \{2, 6, 10, 20\}$. The dashed line depicts the



(a) MAE versus M, ordering the variables.

(b) The same of Figure (a) but in log-log scale.

Figure 1: MAE versus M obtained ordering the variables according to the different rankings **(for the output 1 - arousal)**. At each M, we consider the first M variables in each ranking and compute the MAE. Clearly, when $M = 122$ (i.e., we are using all the variables) all the curves reach the same point. The black solid line corresponds to the MAE curves without ordering the variables.

Table I: **Results of the ranking methods - output 1.**

Meth.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
RM1	113	39	14	8	4	56	115	114	16	13	37	77	43	107	55	3	11	88	38	48
RM2	113	14	8	114	115	56	4	43	18	13	107	55	11	88	77	38	3	59	81	78
RM3	113	114	14	20	18	119	13	12	9	8	21	3	15	56	88	4	7	65	38	39
RM4	113	114	1	115	50	116	2	51	14	20	119	12	8	13	9	21	117	18	15	3
RM5	113	114	14	116	2	20	51	1	115	50	13	9	3	21	4	122	29	56	31	23

Table II: **Possible GR - output 1.**

GR:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Variable:	113	114	14	8	13	4	56	3	88	15	115	38	23	20	21	39	18	37	48	117
RM1:	1	8	3	4	10	5	6	16	18	25	7	19	24	30	58	2	88	11	20	31
RM2:	1	4	2	3	10	7	6	17	14	24	5	16	23	75	28	56	9	57	21	25
RM3:	1	2	3	10	7	16	14	12	15	13	82	19	24	4	11	20	5	29	43	21
RM4:	1	2	9	13	14	21	22	20	23	19	4	24	25	10	16	28	18	26	39	17
RM5:	1	2	3	21	11	15	18	13	23	25	9	31	20	6	14	27	22	33	38	81
N. Score:	1	0.98	0.97	0.92	0.92	0.90	0.90	0.88	0.85	0.83	0.83	0.83	0.82	0.80	0.80	0.79	0.77	0.75	0.74	0.72

equiprobability (uniform) distribution with probability $1/122$. Clearly, probabilities bigger than $1/122$ denote the most important variables.

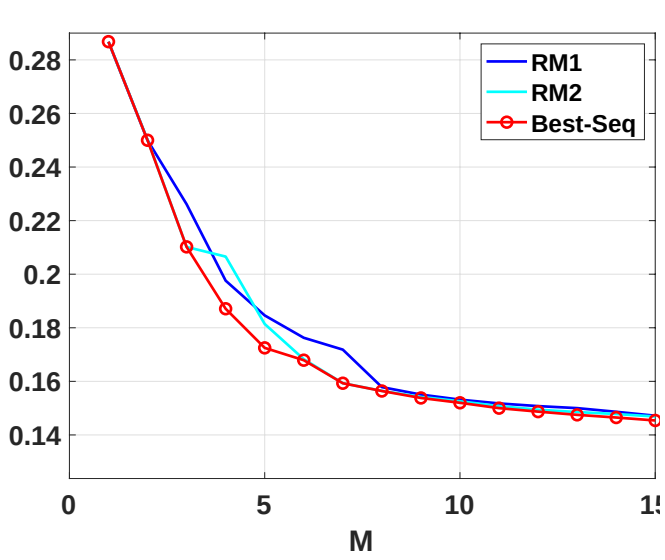
We can observe that the results are coherent with the previous results above. For instance, the variable 113 is clear the most important and also the features 114, 4, 115, 56, 8 are quite relevant. The Gibbs analysis also confirms that the feature 39 seems relevant for small M but, as M grows, the importance of this feature disappears.

However, by the Gibbs analysis, we can obtain more interesting information. The features 4 and 56 are quite relevant even from small values of M, confirming also the results of the best sequence search. The features 14 seems to have a relevance very similar to 114 (confirming the results of the GR). As we have also previously stated, the Gibbs analysis clearly shows that the variable 115 is the fourth most important feature (as we expect after a care look of the rankings). Surprisingly, the feature 1 seems to be equally relevant than the feature 115: we provide an explanation below. The variable 13 seems also to be some relevance by the Gibbs analysis specially as M grows. However, as we expected for the previous study, its relevance is moderate.

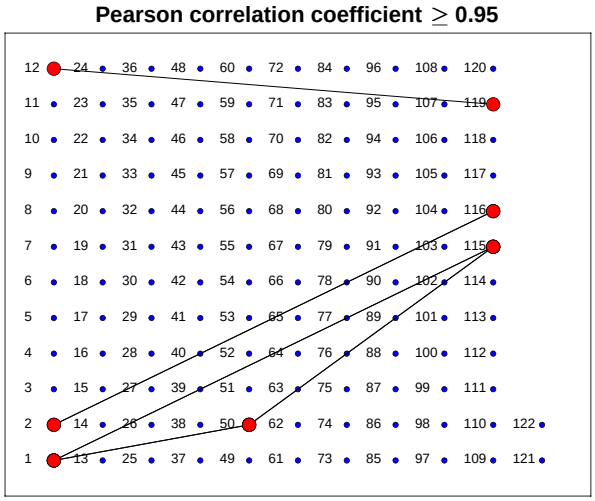
Furthermore, we can also observe the importance of other variables whose relevance was not clear from the previous

Table III: **Best Sequences - Output 1**

M	Labels of the features in the best sequence											
1	113											
2	39	113										
3	8	14	113									
4	4	8	14	113								
5	10	56	113	114	115							
6	10	13	56	113	114	115						
7	4	8	14	56	113	114	115					
8	4	8	14	43	56	113	114	115				
9	4	8	14	16	43	56	113	114	115			
10	4	8	14	16	43	56	107	113	114	115		
11	4	8	13	14	16	43	56	107	113	114	115	
12	4	8	13	14	16	43	55	56	107	113	114	115



(a)



(b)

Figure 2: **(a)** MAE versus M obtained ordering the variables according to RM1, RM2 and the best sequence search **(Output 1)**. **(b)** The connections among the variables with the Pearson correlation coefficient ρ such that $|\rho| \geq 0.95$.

analysis above. This is the case of the following features:

1, ("RMS mean"),
20, ("flatness mean"),
50, ("flux mean"),
16, ("skewness mean"),
and 22 ("entropy mean").

The feature 1 deserves some specific comments (see below). The feature 20 is in 14-th position of the GR, in the fourth position of RM3, in the 10-th position of RM4, and in the 6-th position of RM5. The variable 50 appears also in the fifth position of RM4, and in the 10-th position of RM5. The variable 16 appears in the best sequences for $M \geq 9$ and in 9-th position of RM1. The feature 22 has not been detected by the previous studies: it does not appear either in the rankings, or in the best sequence search (at least for $M \leq 12$ as in Table III).

About the feature 1. Recall that the variable 1 ("RMS mean") does not appear in the first best 20 variables on the GR, but is the third most important variable by RM4 and is in position 8-th by RM5. It is important to note that the feature 1 seems very important after Gibbs sampling study. The variable 1 seems to have equal importance than the feature 115, obtaining virtually the same probabilities. A correlation study between pairs of variable reveals that the features 1 and 115 have a very high Pearson correlation coefficient (greater than 0.95), as shown in Figure 2(b). Moreover, Figure 2(b) shows there is a strong correlations among the features 1, 50 and 115 (forming a triangle in the figure).

Comparison RM2 and BS with $M = 12$. The best sequence

for $M = 12$ is

4 (“maximum fluctuation”),
8 (“roll-off mean”),
13 (“centroid standard deviation”),
14 (“spread mean”),
$\boxed{16}$ (“skewness mean”),
43 (“7th MFCC stand. deviation”),
55 (“pitch mean”),
56 (“pitch standard deviation”),
107 (“7th chromagram stand. deviation”),
113 (“loudness mean”),
114 (“loudness standard deviation”),
and 115 (“energy mean”),

and the first better variables for RM2 are

$113, 14, 8, 114, 115, 56, 4, 43, \boxed{18}, 13, 107,$ and 55,
(ordered by RM2).

They differ only for the variables 16 (“skewness mean”) and 18 (“kurtosis mean”). After the Gibbs analysis, the feature 16 seems slightly more relevant than 18. See, e.g., Fig. 3(d). However, following the Gibbs analysis, other variables could provide a more robust results, for instance, the features 20 (“flatness mean”), 22 (“entropy mean”) or 50 (“flux mean”).

D. Summary for the output 1 - Arousal

Here, we classify the features into four different classes: *very relevant* (Level 1), *relevant* (Level 2), and *relevant but maybe only for the specific dataset* (Level 3), and the rest of variables belong to the class *non-relevant* (Level 4).

Level 1. After all the studies, we can assert at least 7 variables are very relevant:

$113, 114, 14, 115, 4, 8,$ and $56,$ (ordered by Gibbs analysis),

which are shown in decreasing order of importance considering the Gibbs sampling analysis. This is also the best sequence for $M = 7$, as shown in Table III and includes also the first 7 elements in the ranking obtained by RM2 but with a different order,

$113, 14, 8, 114, 115, 56,$ and 4 (ordered by RM2).

With the exception of the feature 115, the rest of variables are also contained in the first 7 positions of the GR in Table II, that is

$113, 114, 14, 8, 13$ (instead of 115), $4,$ and 56,
(ordered by GR).

Level 2. Other important variables are

$1, 22, 20, 50, 13, 16,$ and 3 (ordered by Gibbs analysis),

but the features 1 (“RMS mean”) and 50 (“flux mean”) are highly correlated to the variable 115 (“energy mean”), as shown by Figure 2(b) and as we could intuitively expect. The variable 13 (“centroid standard deviation”) is the fifth in GR and appears in some of the first twelve best sequences. However, the Gibbs analysis reveals (that in terms of robustness) other variables such as 20 (“flatness mean”) and 22 (“entropy mean”) are more or a similar relevance than 13. The feature 20 is the 14-th variable in the GR, and is particularly important in RM3 (4-th position) and RM5 (6-th position). The feature 16 (“skewness mean”) does not appears in the best first twenty variables in the rankings, but appears in the best sequences permanently for $M \geq 9$. In the Gibbs analysis, the feature 16 acquires some relevance as M grows. Finally, the feature 3 (“decrease slope mean”) does not appear in the first twelve best sequences and it does not seems relevant by the Gibbs analysis. However, it appears in 8-th position of GR and within the first twenty more relevant variables in *all* the ranking methods.[1]

Level 3. Other possibly important variables, which appear in the best sequence search and in the rankings RM1 and RM2, are

43 (“7th MFCC stand. deviation”)
and 107 (“7th chromagram stand. deviation”).

The variable 43 appears in 8-th position in RM2 and 13-th position in RM1. Moreover, it appears in the best sequences for $M \geq 8$. The feature 43 appears in 11-th position in RM2 and 14-th position in RM1. Moreover, it appears in the best sequences for $M \geq 10$. On the other hand, the Gibbs analysis does not associate any particular relevance to these variables.

E. Selection of the number of variables for the output 1

First of all, we have applied different information criteria, such as the AIC and BIC, shown in Table **??** [**?**]. The more adequate results have been provided the Bayesian information criterion (BIC) which suggests to use 17 variables whereas AIC suggests the use of 40 variables. We have also tried the classical analysis based on p-values which suggests 71 variables [**?**], [**?**]. The results of the corresponding ranking method is given in Table IV. The first most relevant 7 variables are again 113, 14, 8, 4, 56, 115 and 114, i.e., the *very relevant* features that we have found after our analysis.

However, after our exhaustive study, we believe that a more parsimonious model can be suggested. The most parsimonious LM after all the consideration in our study, is the model which includes at least the seven *very relevant* variables (described above),

$$113, 114, 14, 115, 4, 8, \text{ and } 56. \tag{5}$$

More precisely, the suggested linear model is

$$\begin{aligned} y_1 = &- 0.5293 + 3.6494\ x_{113} + 1.8080\ x_{114} + \\ &- 1.5534\ x_{14} - 3.8491\ x_{115} + \\ &+ 1.5056\ x_4 + 1.1714\ x_8 - 0.3450\ x_{56}, \end{aligned} \tag{6}$$

[1]More surprisingly, for the output 2 - valence -, only three features are included within the first twenty more relevant variables in *all* the ranking methods: they are the variables 113, 114 and again 3. Namely the feature 3 (as 113 and 114) is included within the first twenty more relevant variables in *all* the ranking methods for both outputs.

obtaining a MAE of 0.1593, MSE of 0.0432 (i.e, RMSE of 0.2078), and $R^2 = 0.8703$. Considering a Monte Carlo cross-validation procedure (with 80% of the data in the train-set and the rest of 20% of data in the test-set, chosen randomly in each $2 \cdot 10^4$ independent runs), we obtain MAE of 0.1611, MSE of 0.0450 (i.e, RMSE of 0.2118), and $R^2 = 0.8641$. Namely, we have a very slight increase of MAE and MSE (or a slight decrease of R^2), proving the robustness of our proposed model.

II. Results for the output2 - Valence

In this section, we analyze the the output 2 of the dataset, i.e., valence. From the results that we will provide in this section (by all the figures and tables below), we can note that the output 2 (valence) is less linear correlated with the variables $\mathbf{x}$, compared with th first output (arousal). Here we mainly discuss the variables which seem relevant for both outputs (1 and 2) and some features just relevant for output 2. The most important features for output 2 are

114 ("loudness standard deviation"),
113 ("loudness mean"),
14 ("spread mean"),
and 3 ("decrease slope mean").

They are also relevant for the output 1 (as we can see in the main body of the work). The variable 114 seems to increase its relevance with respect the output 1, whereas the variable 3 is much more relevant for the output 2. The importance of these features is confirmed by the Gibbs analysis in Fig. 5, specially for the feature 14. Indeed, the case of feature 14 is very interesting and reveals also the importance of the Gibbs analysis. The variable 14 does not seem relevant following RM1 and RM2 (which provides the sequences with the smaller errors) and does not appears in the best sequences in Table VII. However, it is the third more important variables for RM3, RM4 and RM5, and it is the third more relevant variable for the Gibbs analysis (see Fig. 6). Furthermore, it acquires more relevance as M grows (see again Fig. 6). The variables

1 ("RMS mean"),
and 115 ("energy mean"),

(which are highly correlated, also with the feature 50; see the main body of the paper for further details) appear relevant for the second output as well. The variable 1 is contained in the first twenty more important features in RM2, RM4 and RM5. Moreover, the variable 1 appears in the best sequences playing the role of the variable 115, i.e., when the feature 115 does not appear in those sequences. Namely, due to their correlation in the rankings and in the best sequences, the presence of one of them (1 or 115) avoids the presence of the other one. The feature 115 appears in the best sequences almost in a stable way for $M \geq 3$. The Gibbs analysis confirms the relevance of both 1 and 115, and they seem even more relevant than the variable 3. Furthermore, the following variables

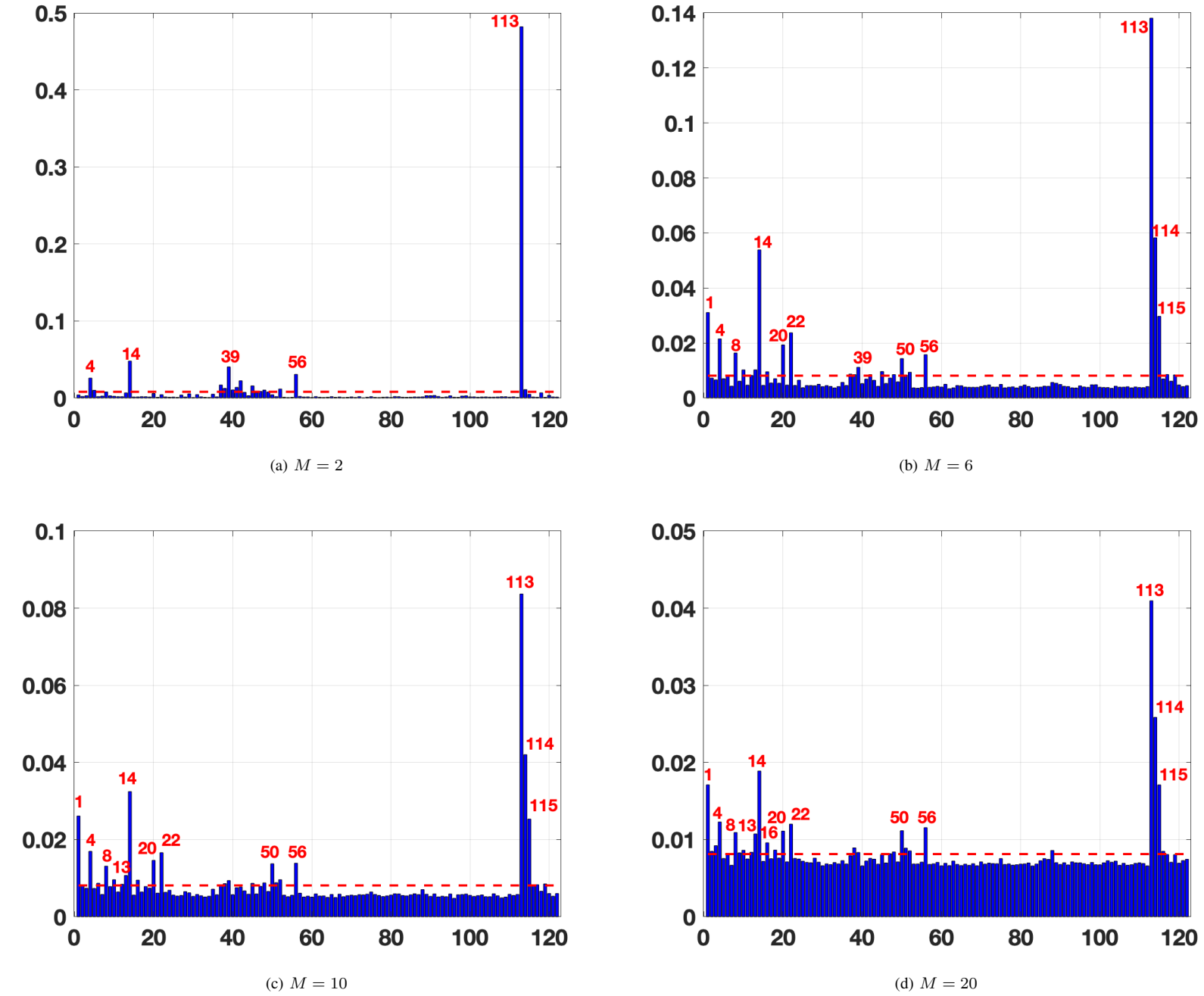
50, ("flux mean"),
20, ("flatness mean"),
13, ("centroid standard deviation"),
and 8 ("roll-off mean"),

are also important for the output 2. The variable 20 is in the 7-th position of the GR, the variable 50 is of the 8-th position in the GR (but is highly correlated with the features 1 and 115), the variable 13 is of the 9-th position of the GR and the feature 8 is in the 14-th position of the GR.
The variables above have certain relevance also for the output 1. Below, we discuss some features that seem to have importance only for the output 2 (i.e., valence).

Variables relevant mainly for the output2 (not for output 1). A careful look to the results reveals that the following features

88 ("inharmonicity standard deviation"),
31 ("8th MFCC mean"),
40 ("4th MFCC standard deviation"),
42 ("6th MFCC standard deviation"),
52 ("Low Energy"),
79 ("11th chromagram center stand. deviation"),
109 ("9th chromagram stand. deviation"),
and 110 ("10th chromagram standard deviation"),

are relevant, and appear in the best sequences for the output 2. Moreover, the feature 88 is the second more important in RM1 and appears in the 13-th position of RM3 (and in the 18-th position of the GR). The importance of 88 is confirmed (and is even more clear) by the Gibbs sampling analysis shown in Figure 5. The feature 31 seems to be relevant for RM1, RM5 and the Gibbs analysis. Moreover, it appears in the best sequences for $M \geq 11$. The variable 40 is within the first twenty most important variables of RM1 and RM2, and takes the 11-th position in the GR. It also appears in the ranking based on p-values in the 10-th position (see Table VIII). Moreover, it starts to appear in the best sequences for $M \geq 10$. The importance of the feature 40 increases as M grows, following the Gibbs analysis. The feature 42 seems to have also the same importance of the variable 40 for the Gibbs analysis, and appears in the 16-th position of the GR and in the 15-th position of RM1. The variable 52 takes the 9-th position in RM2, appears in the best sequences for $M \geq 10$ and seems relevant according to the Gibbs analysis. The feature 79 appears within the first twenty more important variables in RM1, RM2 and RM3. From the Gibbs analysis, it seems clear that its importance grows with M. In the best sequences, the feature 79 appears in a stable way for all the best sequences with $M \geq 8$. The feature 110 is contained within the first twenty more important variables in RM1, RM2 and RM3. It is in the 10-th position of the GR. Following the Gibbs analysis, the feature 109 is similar or



(a) $M = 2$

(b) $M = 6$

(c) $M = 10$

(d) $M = 20$

Figure 3: Results in terms of probabilities obtained by a Gibbs sampling analysis for the Output 1. The dashed line depicts the equiprobability (uniform) distribution with probability $1/122$.

Table IV: **Results of RM based on p-values for the output 1**

Ranking Method	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
RM based on p-values	113	14	8	4	56	115	114	16	43	75	55	13	37	122	2	107	11	40	3	38

more relevant than 110. It also appears in the best sequence with $M = 7$ and in the 6-th position of the classical ranking based on p-values. The feature 50 seems relevant by the Gibbs analysis but it is very correlated to 1 and 115 (see the main body of this work for further details).

On the possible consensus of RM1, RM2 and best sequences. Figure 4 shows the MAE obtained by ordering the variables with the different ranking. Figure 5 depicts the MAE curve associated to the best sequences and compared with the curves of RM1 and RM2. Again, as with output 1, these three curves are very close. However, it is important to

remark that the corresponding sequences of variables ordered by RM1, RM2 and according to the best sequences *are very different*. They agree, in the first twelve more important feature, just for

114 ("loudness standard deviation"),
113 ("loudness mean"),
1 ("RMS mean") - or 115 ("energy mean"),
and 3 ("decrease slope mean").

Considering only RM1 and the best sequences, there is also a consensus on the features

88 ("inharmonicity standard deviation"),
and 109 ("9th chromagram stand. deviation").

Considering only RM2 and the best sequences, there is also a consensus on the variables

79 ("11th chromagram center stand. deviation"),
and 72 ("4th chromagram center stand. deviation").

On the consensus of RM1, the ranking based on p-values, and best sequences. The results of the classical ranking method based on p-values is given on Table VIII. The stopping rule of this classical analysis suggests to use 83 variables. We can observe that the first five more relevant variables

114 ("loudness standard deviation"),
88 ("inharmonicity standard deviation"),
115 ("energy mean"),
3 ("decrease slope mean"),
113 ("loudness mean"),
(ordered for their relevance),

are also important for RM1. Note that the first four variables above are in the same position and the same order in RM1 and in Table VIII. Moreover, the first four features form the best sequence with $M = 4$. The variable 113 appears in the best sequence with $M = 5$.

III. Proposed model for the output 2

For the output 2, the BIC suggests the use of 22 variables, whereas the AIC suggests the use of 68 variables. The classical p-values method suggests the use of 83 variables. However, all the considerations in our study above, we believe that a more parsimonious model can be proposed. In our opinion, The most parsimonious linear model that we can suggest is the model which includes at least the six *very relevant* variables which are (see the considerations above)

114 ("loudness standard deviation"),
113 ("loudness mean"),
14 ("spread mean"),
88 ("inharmonicity standard deviation"),
115 ("energy mean"),
3 ("decrease slope mean"),
(ordered by the Gibbs analysis),

and the ten *relevant* variables,

8, ("roll-off mean"),
20, ("flatness mean"),
79 ("11th chromagram center stand. deviation"),
$\boxed{4}$, ("maximum fluctuation"),
109 ("9th chromagram stand. deviation"),
110 ("10th chromagram standard deviation"),
40, ("4th MFCC standard deviation"),
31, ("8th MFCC mean"),
42, ("6th MFCC standard deviation"),
and 52 ("Low Energy"),
(ordered by the Gibbs analysis),

where we have included the feature 4 due to the Gibbs analysis. It appears also in the first twenty positions of RM3, RM4 and RM5: in the 12-th position, in the 18-th position, and in the 15-th position, respectively. The variable 72 has been excluded due to the Gibbs analysis, as well. More precisely, the suggested model is

$$\begin{aligned} y_2 =& 0.2831 - 2.4741\ x_{114} - 1.0919\ x_{113} + 0.8070\ x_{14} + \\ & 0.2538\ x_{88} + 2.8482\ x_{115} - 0.6448 x_3 + \\ & - 1.4867\ x_8 + 1.1290\ x_{20} - 0.2003\ x_{79} + \\ & - 0.7192\ x_4 + 0.5182\ x_{109} + 0.1642\ x_{110} + \\ & 0.3312\ x_{40} + 0.2978\ x_{31} + 0.4621\ x_{42} - 0.5342\ x_{52}, \end{aligned} \tag{7}$$

obtaining a MAE of 0.2799, MSE of 0.1182 (i.e, an RMSE of 0.3437), and an $R^2 = 0.6452$. Considering a Monte Carlo cross-validation procedure (with 80% of the data in the train-set and the rest of 20% of data in the test-set, chosen randomly in each $2 \cdot 10^4$ independent runs), we obtain MAE of 0.2849, MSE of 0.1233 (i.e, RMSE of 0.3509), and $R^2 = 0.6311$. As for the output 1, we have a very slight increase of MAE and MSE (and a slight decrease of R^2), proving the robustness of our proposed model.

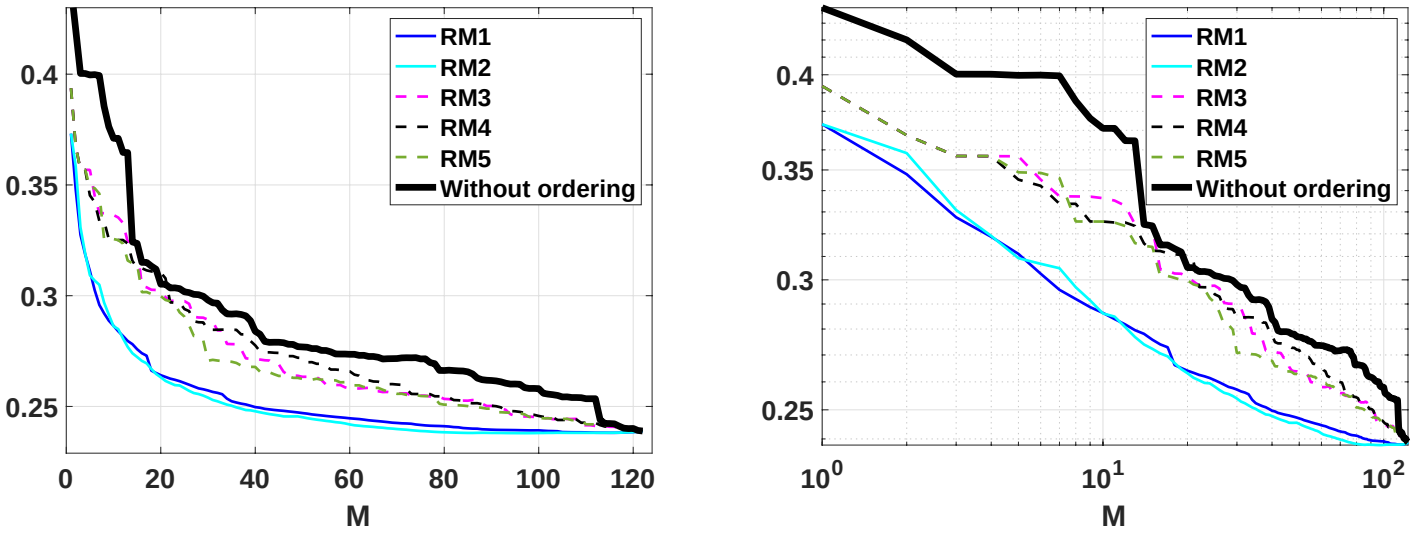
References

[1] M.A. Efroymson. Multiple regression analysis. *Mathematical methods for digital computers*, pages 191–203, 1960.

[2] R. R. Hocking. The analysis and selection of variables in linear regression. *Biometrics*, pages 1–49, 1976.

[3] F. Llorente, L. Martino, D. Delgado, and J. Lopez-Santiago. Marginal likelihood computation for model selection and hypothesis testing: an extensive review. *arXiv:2005.08334*, 2020.

[4] R. L. Thorndike. Who belongs in the family? *Psychometrika*, 3:267–276, 1953.



(a) MAE versus M, ordering the variables.

(b) The same of Figure (a) but in log-log scale.

Figure 4: MAE versus M obtained ordering the variables according to the different rankings **(for Output 2)**. At each M, we consider the first M variables in each ranking and compute the MAE. Clearly, when $M = 122$ (i.e., we are using all the variables) all the curves reach the same point. The black solid line corresponds to the MAE curves without ordering the variables.

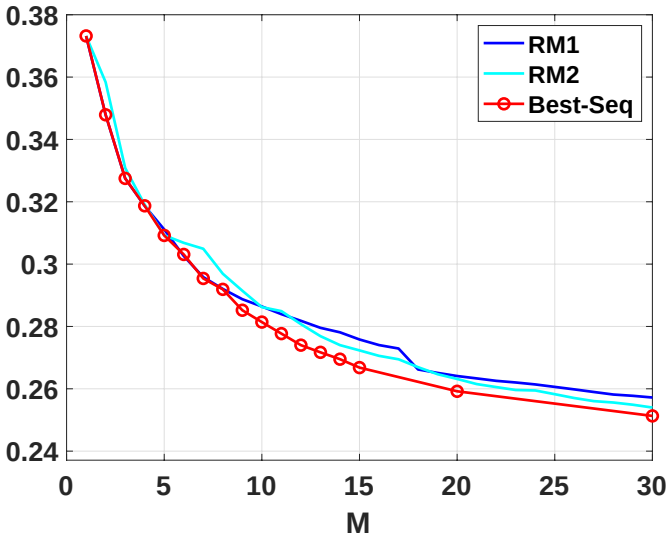


Figure 5: MAE versus M obtained ordering the variables according to RM1, RM2 and the best sequence search **(Output 2)**.

Table V: **Results of the ranking methods - Output 2.**

Meth.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
RM1	114	88	115	3	42	5	113	109	64	31	59	15	9	121	13	36	79	75	40	110
RM2	114	1	113	110	3	79	94	62	52	72	12	14	40	15	10	101	91	92	29	67
RM3	113	114	14	20	18	3	119	9	13	65	94	4	88	86	79	110	101	121	34	42
RM4	113	114	14	20	50	51	2	116	1	115	18	9	13	3	119	19	17	4	12	34
RM5	113	114	14	116	2	20	51	1	115	50	13	9	3	21	4	122	29	56	31	23

Table VI: **Possible global ranking (GR) - Output 2.**

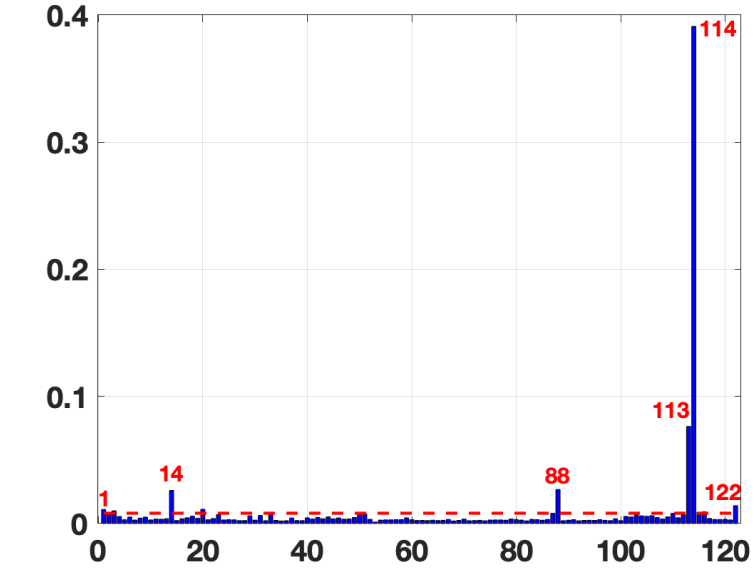
GR:	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Variable:	114	113	3	14	9	12	20	50	13	110	40	122	34	8	101	42	29	88	2	15
RM1:	1	7	4	33	13	34	39	58	15	20	19	43	31	40	27	5	50	2	46	12
RM2:	1	3	5	12	30	11	67	32	91	4	13	27	22	68	16	51	19	81	42	14
RM3:	2	1	6	3	8	22	4	34	9	16	21	42	19	27	17	20	41	13	89	60
RM4:	2	1	14	3	12	19	4	5	13	26	73	42	20	21	22	78	57	67	7	82
RM5:	2	1	13	3	12	28	6	10	11	98	41	16	78	21	101	30	17	23	5	25
N. Score:	0.99	0.98	0.94	0.92	0.88	0.82	0.81	0.78	0.78	0.74	0.73	0.73	0.73	0.72	0.71	0.71	0.71	0.70	0.70	0.69

Table VII: **Best Sequences - Output 2**

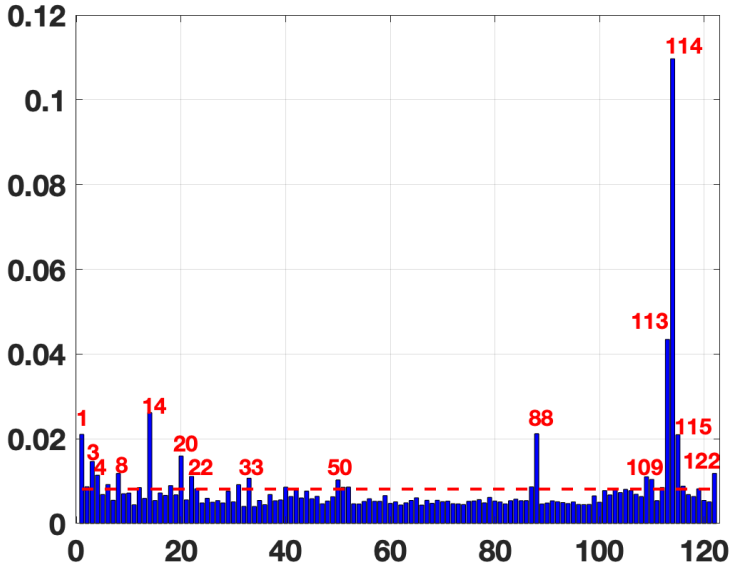
M	**Labels of the features in the best sequence**											
1	114											
2	88	114										
3	88	114	115									
4	3	88	114	115								
5	3	104	113	114	115							
6	33	40	113	114	115	119						
7	3	5	88	109	113	114	115					
8	3	5	72	79	110	113	114	115				
9	3	5	72	79	88	110	113	114	115			
10	3	5	40	72	79	88	112	113	114	115		
11	3	31	40	52	72	79	91	110	113	114	115	
12	1	3	31	40	52	72	79	88	91	110	113	114

Table VIII: **Results of RM based on p-values - output 2.**

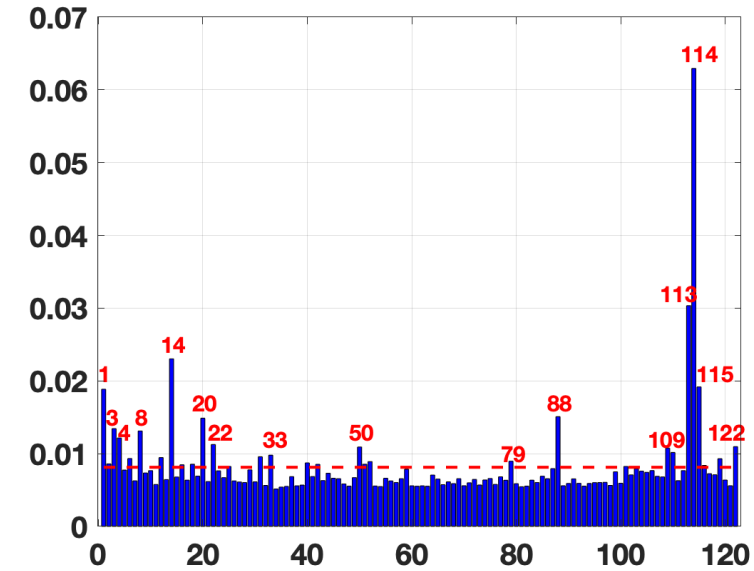
Ranking Method	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
RM based on p-values	114	88	115	3	113	109	5	25	65	40	33	37	4	122	68	107	36	2	51	47



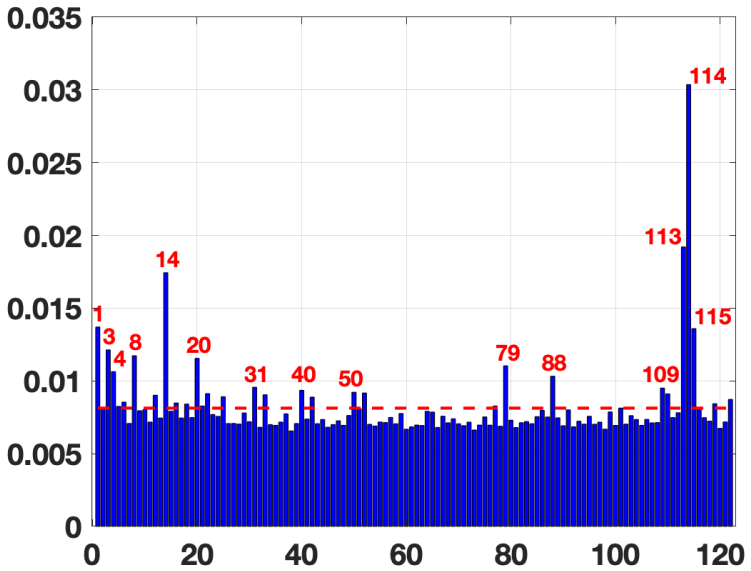
(a) $M = 2$



(b) $M = 6$



(c) $M = 10$



(d) $M = 20$

Figure 6: Results in terms of probabilities obtained by a Gibbs sampling analysis for the output 2 (valence). The dashed line depicts the uniform discrete distribution with probability $1/122$.