entropy

MDPI

Review

A Short Review on Minimum Description Length: An Application to Dimension Reduction in PCA

Vittoria Bruni [1,2,†], Maria Lucia Cardinali [1,†] and Domenico Vitulano [1,2,*,†]

[1] Department of Basic and Applied Sciences for Engineering, Sapienza Rome University, Via Antonio Scarpa 16, 00161 Rome, Italy; vittoria.bruni@uniroma1.it (V.B.); marialucia.cardinali@uniroma1.it (M.L.C.)
[2] Istituto per le Applicazioni del Calcolo, Via dei Taurini 19, 00185 Rome, Italy
* Correspondence: domenico.vitulano@uniroma1.it
† These authors contributed equally to this work.

Abstract: The minimun description length (MDL) is a powerful criterion for model selection that is gaining increasing interest from both theorists and practicioners. It allows for automatic selection of the best model for representing data without having a priori information about them. It simply uses both data and model complexity, selecting the model that provides the least coding length among a predefined set of models. In this paper, we briefly review the basic ideas underlying the MDL criterion and its applications in different fields, with particular reference to the dimension reduction problem. As an example, the role of MDL in the selection of the best principal components in the well known PCA is investigated.

Keywords: minimum description length; principal component analysis; dimension reduction; classification; features extraction

Citation: Bruni, V.; Cardinali, M.L.; Vitulano, D. A Short Review on Minimum Description Length: An Application to Dimension Reduction in PCA. *Entropy* **2022**, *24*, 269. https://doi.org/10.3390/e24020269

Academic Editor: Miguel Rubi

Received: 4 January 2022
Accepted: 10 February 2022
Published: 13 February 2022

1. Introduction

Dimensionality reduction plays a crucial role in the analysis of high-dimensional data. It consists of reducing the number or the dimension of features referred to in a given class of data without losing the capability of being distinctive for that class. It represents a critical issue in classification, as it has been widely proved that classifiers are not able to reach their goal whenever the number of features is too high (too much data) or too small [1,2]. The literature is rich in methods and approaches for reaching this goal; some well-known and popular examples are principal components analysis (PCA) [3], non-negative matrix factorization [4], isomaps [5], t-distributed stochastic neighbor embedding [6], uniform manifold approximation and projection for dimension reduction [7], autoencoders [8], multidimensional scaling (MDS) [9], and so on. Each one is based on one or more criteria to use for dimension reduction. For example, principal components analysis (PCA), which represents one of the most popular and commonly used methodologies, mainly consists of an orthogonal projection of the data onto a lower-dimensional linear space, where the variance of the data is preserved or maximized. As a matter of fact, the dimension reduction problem resembles the sparsity problem, as it requires condensing the peculiarity of the object of interest, which makes it distinguishable from others, into a very small number of features. This comparison/connection is as true as those features that are the coefficients of a given transform: linear or not linear, redundant or not redundant, defined by a single basis or a dictionary. However, the data compaction/compression task is a longstanding, still unsolved, and open problem that has been partially overcome by introducing a dictionary of bases. However, even in this case, it is necessary to define a fast and effective algorithm for the selection of the most significant elements of the dictionary.

The equivalence between dimension reduction , sparsity, and optimal coding tasks, especially in the blind context, is well summarized and conveyed by the minimum description length (MDL) principle, which allows the selection of a good model for approximating

the data with the least complexity [10]. It is based on the concept that good compression means good approximation, which is in agreement with the Kolmogorov complexity.

The MDL principle was formulated about 30 years ago [10,11]. It is mainly based on information theory principles and has inherited several aspects from the Kolmogorov complexity [12]. It has been designed as a statistical inference method [13], where the rationale is that the observed data have to be compressed by the model. Several candidate models can be then compared on the basis of how much they can compress data—by retaining useful information while discarding noise. It turns out that the best model (i.e., the model along with its free parameters) will be the one that gives the shortest code for the data under examination [14,15]. It is worth outlining that MDL changes the perspective of model selection. In fact, it does not assume any 'true' model for the specified data, as do classical probability and Bayesian models. It simply tries to do its best with the set of available candidate models. The basic principle founding MDL is, then, very simple: the simplest model that fits the data well is also the best one.

The simplest formal way to implement MDL is the crude MDL (or two-part code); it selects a model from a set of candidates by minimizing the total cost that is defined as the cost (expressed in terms of bits) required for coding the model plus the number of bits required for coding the data given the model. It is worth observing that the latter is strictly related to the ability of the model to represent the data, and it is often reached by costly models. Hence, the selection of the best model consists of a trade-off between the complexity of the model and good data representation/coding. Unfortunately, the practical construction of the MDL functional is not trivial, nor is its minimization. This is why the literature is rich in proposals that allow researchers to address one or more of these technical issues. In any case, despite the difficulty in designing effective and computable algorithms, many papers demonstrate, both theoretically and empirically, that the information-theoretic minimum message length principle has some advantages over the standard maximum likelihood estimate [16,17]. In addition, as proved in [18,19], a robust, monotonically convergent, and moderately short algorithm for the selection of the optimal two-part MDL code can be defined only by taking advantage of the concept of Kolmogorov complexity.

MDL was originally designed for model selection (see, for instance, [13,20–23]) and it has been successively applied in different contexts and for different tasks, such as, for example, picking and tuning the best parameters for a given model [14,15]. However, as mentioned at the beginning of this section, in this paper, we are mainly interested in focusing on the feature reduction problem and, specifically, on how MDL can help with 'automatically' setting the number of components in PCA [24–26]—this represents one of the widely used tools for dimension reduction. To this end, the approximation of the normalised version of the MDL functional proposed in [27] has been studied and applied to some conventional data classification problems.

The remainder of the paper is organized as follows. The next section reviews the theoretical formulation of MDL principle. Section 3 presents a brief overview of some of the main uses of the MDL principle in the field of data processing, with a particular focus on principal component analysis. To show the advantages and potentialities of this framework, some numerical examples concerning MDL-PCA are presented in Section 4, with reference to data classification. Finally, Section 5 draws the conclusions.

2. A Short Review about MDL

The interest of the scientific community has been increasing in recent years, and different versions of MDL have been proposed. In the following, a short overview of the crude MDL (sometimes dubbed 'two-part' code) will be given. However, the normalized maximum likelihood (NML) is actually the most-adopted version of MDL, as it provides an effective solution that is supported by an elegant formalism. Hence, a short review of NLM will be provided. The following explanation is not exhaustive at all. For further reading, an introductory and simple lecture on NML can be found in [13], while a technical summary

can be found in [21,28,29]. Details concerning how NML can be used to select the optimal number of features in a classification problem will be given in the next section. Specifically, the approach in [27] will be presented and discussed in order to evaluate pros and cons of NML as a 'features reduction device'.

As already outlined, the simplest formal way to implement MDL is the crude MDL. Specifically, the simplest model that fits a given data sample $\boldsymbol{x}$ well is also the best one. It turns out that the best MDL model $\bar{\mathcal{M}}$, from a set of candidates $\mathcal{M}^{(1)}, \mathcal{M}^{(2)}, \ldots$ is given by the following minimization:

$$\bar{\mathcal{M}} = \underset{\mathcal{M}=\mathcal{M}^{(1)},\mathcal{M}^{(2)},\ldots}{\operatorname{argmin}} \; l(\mathcal{M}) + l(\boldsymbol{x}|\mathcal{M}), \tag{1}$$

where $l(\mathcal{M})$ is the cost (in terms of bits) for coding the model $\mathcal{M}$, and $l(\boldsymbol{x}|\mathcal{M})$ is the cost required for coding the data $\boldsymbol{x}$ given the model. This minimization is not trivial, as it leads to the best trade-off between two competing requirements: the approximation performance of the model and its cost—measured in terms of bits. The result will then be a suitable balance between (model) complexity and (data through the model) representation.

From Crude MDL to Refined MDL: NML

Crude MDL can be considered the first implementation of the Rissanen philosophy. However, its use is limited in different applications, as it usually needs a suitable weight λ for balancing the two cost terms that define the functional minimized in Equation (1). In fact, as emphasized in [21], "it is more problematic to find a good code for hypotheses $\mathcal{M}$ and often 'intuitively reasonable' codes are used; however, it can happen that the description length $l(\mathcal{M})$ of any fixed point hypothesis M can be very large under one code, but quite short under another", making the procedure somewhat arbitrary. There are several approaches in the literature that use crude MDL in a empirical way by applying a corrective weight to one of the lengths in Equation (1), or by properly selecting the coding procedure [15,30,31]—even with optimal performance. A way of making MDL perform better while still being elegant consists of its refined version, namely, the normalized maximum likelihood (NML). In order to introduce this, some preliminary information theory concepts have to be recalled.

In information theory, it is well known that any message $\boldsymbol{x}$, i.e., a sequence of n symbols $x_1, x_2, \ldots x_n$ belonging to a binary alphabet $H = \{0, 1\}$, can be encoded and compressed in a new message $\boldsymbol{y}$ with m symbols with $m \leq n$ — as matter of fact, there is no constraint on the alphabet: the only request is a finite cardinality, and the binary one allows us to denote messages' lengths in terms of bits. In other words, message $\boldsymbol{x}$ can be compressed, giving the (possibly) shorter message $\boldsymbol{y}$. More formally, we can say that the codelengths of $\boldsymbol{x}$ and $\boldsymbol{y}$ satisfy the following relation, $l(\boldsymbol{x}) \geq l(\boldsymbol{y})$, where l is the codelength function. There are various ways to encode $\boldsymbol{x}$. A property required for the codewords that have to encode $\boldsymbol{x}$ is that they belong to a prefix code: any codeword must not be a prefix of any other. This requirement is necessary to produce a uniquely decodable code that is also instantaneous: any codeword can be decoded once it has been received [12]. This class of (prefix) codes plays a fundamental role, as it regulates the foundations of information theory. On the one hand, there is a sort of equivalence between probability distributions and prefix codes. For any probability density function (pdf) p, there is a corresponding prefix code able to encode $\boldsymbol{x}$ via a code with length $l(\boldsymbol{x}) = -\lceil \log_2 p(\boldsymbol{x}) \rceil$, where $\lceil \cdot \rceil$ denotes the first integer $\geq \cdot$. On the other hand, Shannon's source coding theorem states that this code, with its relative (ideal) length $l(\boldsymbol{x}) - \log p(\boldsymbol{x})$, is optimal for $\boldsymbol{x}$ and then for p. To make the presentation more general, in this paper, log will be used by considering a generic basis. From an Information Theory point of view, $\log_2$ should be used in agreement with the only language any real device understands: the binary one.

Now, let us suppose that we have a family $\mathcal{M}$ of distributions $f(\cdot|\boldsymbol{\theta})$ depending on one or more parameters $\boldsymbol{\theta}$. Let us suppose that we have to select among them the model that better fits $\boldsymbol{x}$. Obviously, the best approximation, i.e., the best code for any data $\boldsymbol{x}$, will

be $f(\cdot|\hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate for the data $\boldsymbol{x}$. The codelength of the optimal encoding for $\boldsymbol{x}$, using any distribution of $\mathcal{M}$, is called the *stochastic complexity* of $\boldsymbol{x}$ with respect to $\mathcal{M}$. However, the optimal encoding of $\boldsymbol{x}$, using $f(\cdot|\hat{\boldsymbol{\theta}})$, cannot be used in practice, as its code cannot be specified before data observation. It turns out that an alternative strategy must be found. In particular, the idea is to determine a distribution, and then a code, that performs as well as the family of distributions belonging to $\mathcal{M}$ [20]. MDL suggests introducing the concept of universal distribution p_u, formally defined as:

$$-\log(p_u)(\boldsymbol{x}) \leq -\log(f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_x) + C_n(\mathcal{M}),$$

where $C_n(\mathcal{M})$ behaves as $o(n)$, i.e., $lim_n \frac{C_n(\mathcal{M})}{n} = 0$. The corresponding code will be denoted with universal code. $C_n(\mathcal{M})$ characterizes the universal distribution and represents the maximum difference between the two codes (the universal and the maximum likelihood ones). It is worth observing that more than one universal distribution may exist, and each one is characterized by its own $C_n(\mathcal{M})$; however, apart from a constant, they are almost equivalent and good to approximate $\boldsymbol{x}$.

As already outlined, NML does not assume a 'true' distribution. It simply requires a distribution that is able to fit the data, separating useful information (to preserve) from noise (to discard) [32]. Keeping in mind that, theoretically, the best the model family can do on a dataset $\boldsymbol{x}$ is $-\log f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_x)$, but this is useless from an information theory point of view because it requires prior knowledge of the data $\boldsymbol{x}$; the universal distribution $p_u(\boldsymbol{x})$ allows us to write the additional cost required to encode it, as it follows:

$$-\log p_u(x) + \log f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_x).$$

This quantity is called 'regret' of $p(\boldsymbol{x})$ with respect to $\mathcal{M}$ with the data $\boldsymbol{x}$. It expresses the fact that the universal distribution $p_u(\boldsymbol{x})$ works well, but not 'perfectly', as the original (hypothetical) distribution that originates data $\boldsymbol{x}$ does not exist. Now, the natural question is: which is the worst case scenario?

It can be simply written as a regret, and then:

$$R(p_u||\mathcal{M}) = \max_q E_q\left[\log \frac{f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_x)}{p_u(\boldsymbol{x})}\right] \qquad \forall q.$$

Such a regret should hold for any distribution q, while E_q is the relative expectation. The expression above recalls the Kullback–Leibler divergence [12] that was originally designed for measuring 'the extrabits' required when a message is encoded through a 'wrong' distribution. Hence, the problem can be seen as a minimax one, and it can be formalized as follows:

$$f^{NML} = \arg_p \min_p \max_q E_q\left[\log \frac{f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_x)}{p(\boldsymbol{x})}\right].$$

It is worth providing a few additional details about the equation above.

q stands for any probability distribution that guarantees that $E_q\left[\log \frac{q(\boldsymbol{x})}{f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_x)}\right]$ is finite. f^{NML} is then the NML solution, and it is defined as [20,33]:

$$f^{NML} = \frac{f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_x)}{\int f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_y)\, d\boldsymbol{y}}. \tag{2}$$

It is straightforward to see that $\hat{\boldsymbol{\theta}}_y$ is the ML for the data $\boldsymbol{x}$. Moreover, the denominator integrates over the ML of all possible datasets in a specified context (the one that originated

$\boldsymbol{x}$). The corresponding codelength $-\log\left(f^{NML}(\boldsymbol{x})\right)$ is the stochastic complexity of $\boldsymbol{x}$ with respect to $\mathcal{M}$, that is,

$$sc(\boldsymbol{x}) = -\log f(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_x) + \log \int f(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_y) d\boldsymbol{y}.$$

It is worth outlining that it is not required for p_u and q to belong to the model M as well as f^{NML}—an interesting example is in [13], where the selected distribution does not behave as the one that originated the data.

The stochastic complexity is composed of two terms: the first one quantifies how much the model M approximates the data $\boldsymbol{x}$, while the second one is a measure of the complexity of M. The latter is interesting, as it describes 'how many data' can be well fitted by the model M: as many data can be fit by M as the model M is complex [34].

Finally, an equivalent definition of complexity is provided by the minimum of the worst-case expected regret—details can be found in [33]. Additional aspects concerning MDL, such as asymptotic approximations to NML, and its relation to Bayesian statistics and MDL predictive inference, are out of the scope of this contribution—a deeper but simpler reading concerning these topics can be found in [13].

3. MDL Applications: A Review

Despite the difficulty of its practical application and implementation, MDL has been widely used in different fields by introducing different kinds of approximation, technical tricks, bounds, and so on, for practically using and adapting it to each context. Rissanen himself accurately studied the problem of MDL-based denoising and clustering. In the first case, noise is considered the incompressible part of the data [35]; as a result, MDL can provide the best threshold value whenever denoising is performed in the wavelet domain. In the second one [36], optimal clustering provides the best compression, i.e., the lowest coding cost for each cluster. In this section, we briefly describe some examples of the variety of applications and uses of MDL by grouping them with respect to the main purpose of the specific application they refer to.

Most of papers concerning MDL mainly use it according to its general and original meaning, i.e., the compression and learning model. In this context, it is worth mentioning some recent studies which provide new practical MDL-based ways to compute tight compression bounds in deep-learning models. In particular, in [37], it has been observed that prequential coding yields much better codelengths than variational inference, correlating better with the test set performance — we remind that in the prequential coding, a model with default values is used to encode the first few data; then, the model is trained on these few encoded data; the partially trained model is used to encode the next data; then, the model is retrained on all data encoded so far, and so on. On the contrary, in [38], an MDL-based strategy is used for determining a parameter-free stopping criterion for semi-supervised learning in time series classification, while in [39], the problem of model change tracking and detection has been addressed and studied in both data-compression and hypothesis-testing scenarios. In the first case, an upper bound for the minimax regret for model changes has been found; in the second one, error probabilities for the MDL change test have been derived, and they rely on the information-theoretic complexity, i.e., the complexity of the model class or the model itself and the α-divergence. In a more recent paper [40], the same author introduced the descriptive dimension that characterizes the performance of the MDL-based learning and change detection. In the context of machine learning, MDL has been used for preventing overfitting [41], especially in the case of little available training data. In particular, it has been used for ensuring that there is less information in the weights than in the output vectors of the training cases; to this aim, the model cost is the number of bits it takes to describe the weights, and the cost of the data given the model is the number of bits it takes to describe the discrepancy between the correct output and the output of the neural network on each training case. Very recently, in [42], the neural network training process has been seen as a model selection problem, and the

model complexity of its layers has been computed as the optimal universal code length by means of a normalized maximum likelihood formulation. This kind of approach offers a new tool for analyzing and understanding neural networks while speeding up the training phase and increasing the sensitivity to imbalanced data. More generally, model selection theory allows for an information-theoretic analysis of deep neural networks through the information bottleneck principle [43,44].

With regard to fitting/regression, in [45], MDL is used to successfully reduce the number of false positives in best-fitting-based gene regulatory networks that govern specific cellular behavior and processes. In particular, it has been proved that MDL-based filtering strategies can be computationally less burdensome than using the MDL algorithm alone; in fact, the computation of data-coding length is more complex than calculating the error estimate of the best-fit algorithm, and the computational complexity increases dramatically as the sample size increases. In the same application context, MDL is used for finding the optimal threshold that defines the regulatory relationships between genes [46]. In a different context, and using a different strategy, MDL is used for determining the number of modes in non stationary and highly oscillating signals [31], while in [47], MDL allows for unsupervised spectral unmixing of spectrally interfering gas components of unknown nature and number.

A pioneering paper concerning MDL-based clustering is [48], where a simple MDL cost functional is used to search the tree for a level of clustering with a minimum description length. In [36], the MDL principle is used for data clustering based on the assumption that a good clustering is such that it allows efficient compression when the data are encoded together with the cluster labels. It is worth stressing that, based on the observation that an efficient compression is possible only by discovering the underlying regularities that are common to all the members of a group, this approach also implicitly defines a similarity metric between the data items. Formally, the global code length criterion to be optimized is defined by using the intuitively appealing universal normalized maximum likelihood code, which has been shown to produce optimal compression rates in an explicitly defined manner—the local independence of the model has to be assumed to get a computable algorithm. Ref. [49] presents a study concerning the use of MDL, specifically, the normalized maximum likelihood (NML) version, in the dynamic model selection. The aim is to track changes of clustering structures so that the sum of the data's code-length and clustering changes' code-length is minimized. The study is restricted to the Gaussian mixture model for representing the data, and it has been shown that the proposed method is able to detect cluster changes significantly more accurately than the Akaike information criterion (AIC)-based methods [50] and Bayesian information criterion (BIC) [51]-based methods—an application to market analysis is proposed. In [52], MDL is used for IoT applications. Specifically, a hierarchical clustering is applied for grouping datasets received from sensor nodes: if any pairs of received datasets can be compressed by the MDL principle, they are combined into one cluster.

MDL based strategies are successfully applied for solving the dimension/features reduction problem. Among them, it is worth mentioning the one recently presented in [27], where MDL has been used for the selection of the number of components for the PCA method. Since it is not trivial to practically define MDL, a linear regression model has been used as the bound for its normalized version. In the same context, MDL-based matrix factorization has been proposed in [53], where the objective function is designed through an MDL-based formulation to guide the formation of the matrices defining the model, allowing an automatic and natural trade-off between accuracy and model complexity. In [54], the problem of finding the appropriate feature functions and number of moments is formulated as a model selection problem. MDL is then used for solving it, and it has also been shown that it generalizes the minimax entropy principle. The method has been successfully applied to the gene selection problem to decide on the quantization level and number of moments for each gene; however, the extension to problems involving larger datasets requires more efficient approximations to calculate the complexity.

As further examples, MDL can also be properly used for selecting: (*i*) the least number of image points from which image quality is assessed, in agreement with the human visual system information coding approach, as in [30]; (*ii*) features, which are selected adaptively during online learning, based on their usefulness for improving the objective function, as in [55] ; (*iii*) points on shapes defined as curves to allow for shape recognition, as in [56]; (*iv*) a characteristic subset of patterns on labeled graphs with complex shapes and that are representative of the data, as in [57]. Interesting MDL applications are also the ones that directly work in the wavelet domain and that further take adavantage of the data decorrelation and compaction properties of the transform. For example, in [58], the MDL principle is used for preventing over- or underfitting problems in detrending near-infrared spectroscopy (NIRS) data for neuroimaging applications; in [59], the same principle is used for wavelet-based compression and, in particular, for the selection of the best wavelet and threshold, while in [60], the soft-thresholding-based denoising problem is considered. Finally, in [61], the noisy and original data are properly separated by determining their histogram and retaining the coefficients belonging to specific bins—the optimal set of bins is found by minimizing the sum of the two code lengths for the denoised signal and the noise.

Finally, with regard to the computational cost required by the implementation of an MDL-based method, several efforts have been made in the literature. As a representative example, we mention the method presented in [62], where a computationally feasible algorithm for computing the NML (normalized maximum likelihood) criterion for tree-structured Bayesian networks has been proposed. In particular, the exponential time, required for building Bayesian trees and forests, has been reduced to a polynomial law—in this way, the advantages offered by the information-theoretic normalized maximum likelihood (NML) criterion in Bayesian network structure learning are preserved and easily exploited.

NML for Dimension Reduction in PCA

As mentioned in the Introduction, in this paper, we focus on MDL-based feature reduction and, in particular, on the 'automatic' selection of the number of components in PCA (principal component analysis). The standard measure of quality of a given principal component is the proportion of total variance that it accounts for. As a result, very often, the desidered percentage is fixed and the number of components is derived. However, the number of components often depends on the specific task, and setting the optimal percentage of variance to retain is sometimes user-dependent. However, as the problem is crucial, different methods and criteria have been proposed in the literature. A possible classification of those methods refers to the methodological approach [63,64], i.e.:

- ad-hoc rules, as, for example, the Cattel's scree test [65] and the indicator function [66];
- statistical tests, such as Bartlett's test [67] and the Malinowski's F-test [68];
- computational criteria, such as cross-validation (CV) [69], bootstrapping and permutation, such as Horn's parallel analysis [70], and SVD-based methods [71].

However, it has been shown that each selection method performs differently in real cases, depending on the task. In addition, most of them require a certain computational burden—see [69] for a complete review.

An interesting approach that combines NML and PCA is contained in [27], where an elegant formulation for solving this problem is proposed.

Let us suppose that $\boldsymbol{X}$ is an $n \times m$ matrix, containing the data or the corresponding features. The PCA of $\boldsymbol{X}$ consists of the following minimization:

$$\arg_{\boldsymbol{W},\boldsymbol{Z}:rank(W)=rank(Z)} \min ||\boldsymbol{X} - \boldsymbol{W}\boldsymbol{Z}^T||_F^2, \tag{3}$$

where $\boldsymbol{W}$ and $\boldsymbol{Z}$ are two matrices whose sizes, respectively, are $n \times k$ and $k \times m$, and whose rank is equal to k, while $||\cdot||_F$ denotes the Frobenius norm [72]. The following theorem [25,73] holds that:

Theorem 1 (Eckart–Young–Mirsk). *Let $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T$ be the SVD (singular value decomposition) of $\boldsymbol{X}$, with $\boldsymbol{\Lambda} = diag(\lambda_1, \ldots, \lambda_m)$, while $\boldsymbol{U}$ and $\boldsymbol{V}$ are unitary. Let $\boldsymbol{U}_k$ and $\boldsymbol{V}_k$ be the 'reduced versions' of $\boldsymbol{U}$ and $\boldsymbol{V}$, i.e., containing their first k columns, then:*

$$||\boldsymbol{X} - \boldsymbol{W}\boldsymbol{Z}^T||_F^2 \geq ||\boldsymbol{X} - \boldsymbol{U}_k diag(\lambda_1, \ldots, \lambda_k)\boldsymbol{V}_k||_F^2 = \sum_{i=k+1}^{m} \lambda_i^2, \tag{4}$$

with $\boldsymbol{W} = \boldsymbol{U}_k diag(\lambda_1, \ldots, \lambda_k)$; $\boldsymbol{Z} = \boldsymbol{V}_k$.

This theorem shows that any 'selection/reduction' component leads to a loss of information, and it also quantifies this loss. It turns out that, in principle, Equation (4) could be combined with the NML solution in Equation (2) in order to get the formal stochastic complexity of the PCA-based reduction of $\mathbf{X}$ to k components. Unfortunately, the evaluation of the denominator in Equation (2) is not trivial, as it depends on the eigenvalues of arbitrary matrices. The approach presented in [27] suggests a way to address this issue by adopting the NML of linear regression that takes advantage of quantized versions of the unitary matrices $\boldsymbol{V}_k$. The main trick of this approach consists in considering the generative form of PCA, i.e.,

$$\boldsymbol{X} = \boldsymbol{W}_k \boldsymbol{V}_k^T + \eta,$$

where $\eta \sim \mathcal{N}(0, \tau \boldsymbol{I}_k)$ is the error that is supposed to be normally distributed, and by considering a perturbation of the matrix V_k as follows:

$$V_k^\epsilon = V_k + \epsilon E_k,$$

where ϵ is the quantization bin size for the values of the unitary matrix V_k (whose elements belong to the range $(-1, 1)$), with $\epsilon \leq \frac{1}{m}$ and $|E_k| \leq \frac{1}{2}$, and by writing the corresponding NML—see [27] for the technical details. It turns out that the problem resembles the linear regression one, where the elements of the unitary matrix $\boldsymbol{V}$ are suitably quantized using the quantization parameter ϵ. This way leads to the following result, in agreement with [35]:

Theorem 2. *Let $sc(\boldsymbol{X}; k)$ be the stochastic complexity of the PCA-based reduction of $\boldsymbol{X}$ to k components, then:*

$$\begin{aligned} sc(\boldsymbol{X}; k) \simeq \quad & (nm - kn) \log\left(\textstyle\sum_{i=k+1}^{m} \lambda_i^2\right) + nk \log(||\boldsymbol{X}^T\boldsymbol{X}||_F^2) + \\ & + (mn - kn - 1) \log\left(\tfrac{mn}{mn-kn}\right) - (nk + 1) \log(nk) + \Delta s, \end{aligned} \tag{5}$$

with $0 \leq \Delta s \leq mk \log(2/(m\epsilon))$, $n \times m$ as the dimension of $\boldsymbol{X}$, and ϵ as the quantization bin, such that $\epsilon < \frac{1}{m}$.

It is worth observing that the first term in the second member of Equation (5) represents the code length of the part of the data that adds no further information about the optimal model, i.e., the information that can be neglected; the remaining terms define the length of a code from which the optimal model, which is defined by the ML parameters and that belongs to the subclass of quantized loading matrices of rank k, can be decoded. As a result, the optimal number of principal components is the value of k that minimizes the second member of Equation (5), and the latter only depends on quantities that are known or that can be computed directly from the data.

4. Experimental Results

To better evaluate pros and cons of the theoretical results presented in the previous section, three numerical experiments are presented, referring to two very different datasets. The first dataset is the hyperspectral image Indian Pines [74], captured through the AVIRIS sensor at the Indian test site of North-Western Indiana; each spectrum contains the spectral information of 220 bands in the 0.4–2.5 μm wavelength region, and it is classified in one of

the 16 (+ background) identified classes (such as farmland, forest, highway, housing); each image is composed of 145 × 145 pixels. In particular, the corrected Indian Pines dataset has been downloaded, in which the number of spectral bands is reduced to 200 by removing the ones covering the region of water absorption. The second dataset consists in 162 ECG recordings from the PhysioNet database [75] obtained from three groups of people with cardiac arrhythmia (96 records), congestive heart failure (30 records), and normal sinus rhythms (36 records). For comparative studies, two standard methods for the selection of the number of components to be retained have been considered. The first one refers to the percentage of variance that the components are required to retain; in particular, 90%, 95%, and 99% of the variance of the original data have been considered; the second one refers to the Bartlett test [67]. All tests have been performed implementing a Matlab code (release 2021) on a Intel(R) Core(TM) i7-1065G7 CPU 1.30GHz-1.50 GHz workstation with RAM equal to 16GB.

For the sake of clarity, we split the formula of $sc(\boldsymbol{X};k)$ in Equation (5) into the following terms:

1. $a = (nm - kn)\log\left(\sum_{i=k+1}^{min(m,n)}[\lambda_i^2]\right)$;
2. $b = nk\log\left(||X^TX||_F^2\right)$;
3. $c = (mn - kn - 1)\log\left(\frac{mn}{mn-kn}\right)$;
4. $d = -(nk+1)\ln(nk)$;
5. $e = mk\log\left(\frac{2}{m\epsilon}\right)$, i.e., the upper bound of Δs.

The value of the quantization parameter ϵ has been selected using the theoretical results concerning high resolution quantizers [76]. In this context, the distortion is minimized with a uniform scalar quantization, which means that the distortion has to be significantly less than the variance of the signal to quantize [77]. That is why ϵ has been selected two or three orders of magnitude less than the variance of the matrix $\boldsymbol{V}$ of the SVD decomposition of original data matrix $\boldsymbol{X}$.

TEST 1 The first test is carried out on the hyperspectral dataset and follows the numerical experiments presented in [27]. A set composed of N signals randomly picked from N different classes ($N \leq 16$) plus P random linear combinations of them corrupted by Gaussian noise has been considered—the weights of the linear combination are extracted from a normal distribution of non-negative values with variance $\sigma^2 = 1$, while the Gaussian noise is zero mean, with standard deviation equal to $\sigma = 0.001$. The goal is to find the number of the original signals N.
Each column of the resulting matrix $\boldsymbol{X}$ is a signal so that the dimension of $\boldsymbol{X}$ is $n \times m$, with $n = 200$ being the number of spectral bands and $m = (N + P)$ being the total number of signals. In agreement with [27], the following two configurations have been considered: *(i)* $N = 5$ and $P = 25$, *(ii)* $N = 10$ and $P = 20$. In both cases, the number of independent components N is correctly identified. Figure 1 depicts the behaviour of $sc(\boldsymbol{X};k)$ with respect to k. As can be observed, the estimated stochastic complexity clearly presents a minimum in correspondence to $k = N$. It is worth noting that the local relative minimum shown by the two curves is caused by the term a, which depends on the singular values. The quantization step ϵ has been set equal to 10^{-8}. However, it is worth noting that, in this case, the choice of the quantization parameter is not crucial, since the contribution of the term e to the general trend of $sc(\boldsymbol{X};k)$ is negligible when compared with the contribution of the term b. The computing time required for performing the test has been less than 0.066 s.

TEST 2 The second test refers to ECG data. Here, the same number of signals is randomly selected from the three classes, and the aim is to identify the number of classes.
It is worth observing that, in this case, the dimension of the data matrix $\boldsymbol{X}$ is such that $m \leq 90$, while $n = 65536$. As a consequence of this imbalance, the combined effect of the terms a and d for not-normalized data, and of the term d in the case of data normalized w.r.t the (euclidean) norm of the signal with a maximum norm,

leads to a trivial absolute minimum corresponding to $k = m$, independently of the choice of the quantization step ϵ—resulting in a not-consistent estimation of the cost of the model. This results in the conclusion that the formula in Equation (5) generally fails in any case for which the length of the signals n far outweighs their number m. In Figure 2, the shape of $sc(\mathbf{X}, k)$ is depicted for both not-normalized and normalized data, $\epsilon = 10^{-8}$, and $m = 90$ (30 recordings from each class). Similar plots are obtained for $m = 60$ and $m = 30$. The computing time required for performing the test has been less than 0.135 s. More consistent results are obtained by sampling the analyzed signals; however, sampling may cause the loss of some distinctive features for the signal belonging to the different classes, resulting in the estimation of a smaller number of independent classes, as is shown in Figure 3. In this case, a NML depending on both the number of components and the sampling step would be preferable.

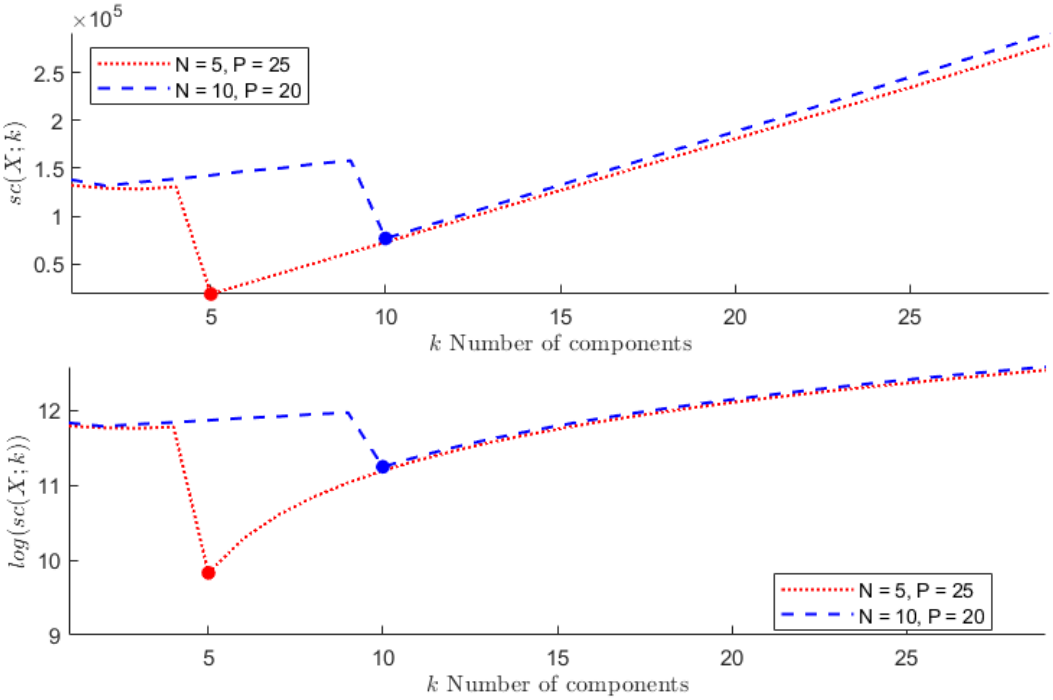

Figure 1. TEST 1. (**Top**) Red dotted line: $sc(\mathbf{X}; k)$ versus the number of components k for $N = 5$, $P = 25$; the minimum is correctly attained at $k = 5$. Blue dashed line: $sc(\mathbf{X}; k)$ versus the number of components k for $N = 10$, $P = 20$; the minimum is correctly attained at $k = 10$. (**Bottom**) The same plot where $log(sc(\mathbf{X}; k))$ has been considered to improve its readability in correspondence to the minimum value.

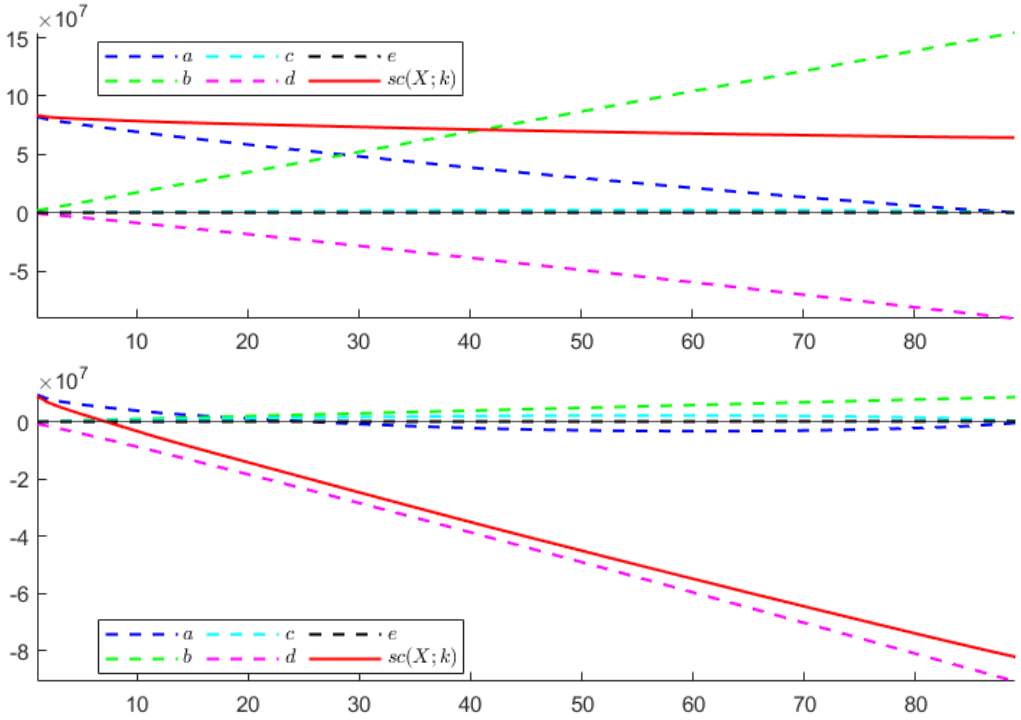

Figure 2. TEST 2. Plot of $sc(\mathbf{X};k)$ and its components versus the number of components k. (**Top**): non-normalized data. (**Bottom**): normalized data.

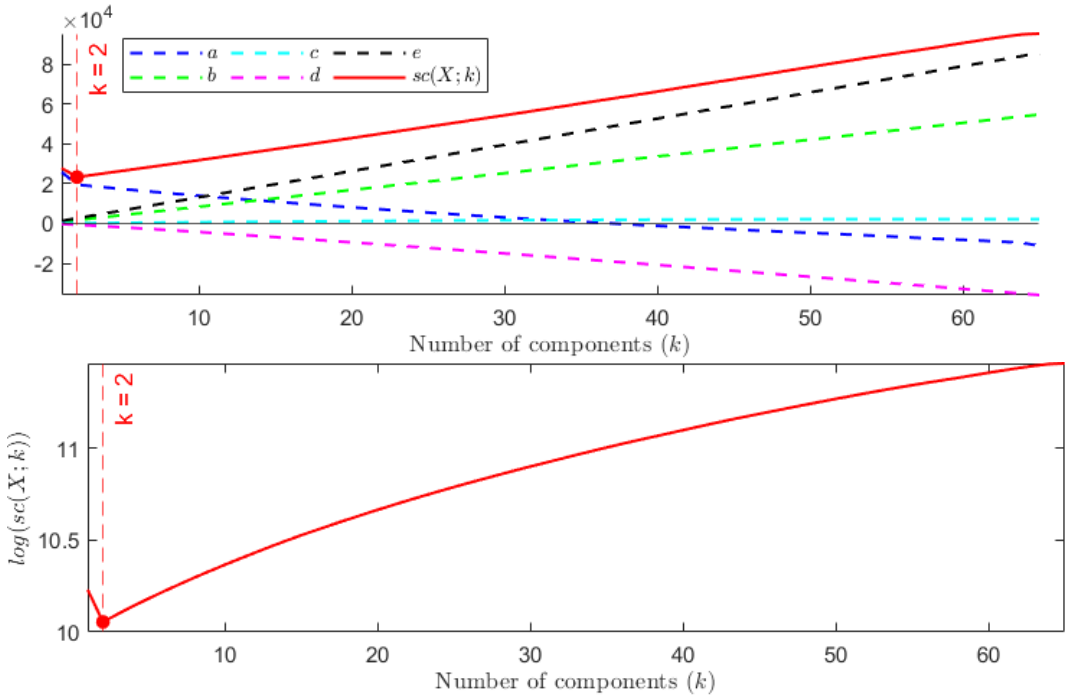

Figure 3. TEST 2. (**Top**) Plot of $sc(\mathbf{X};k)$ and its components versus the number of components k. Signals have been uniformly sampled so that the dimension of $\mathbf{X}$ is $n \times m = 66 \times 90$. (**Bottom**) Plot of $log(sc(\mathbf{X};k))$.

TEST 3 The third test aims at using the proposed NML-based feature reduction method in a more interesting (real) case concerning hyperspectral image classification.

For classification purposes, the conventional approach consists of first reducing the dimensionality of the data by applying PCA, and then feeding the transformed data to an SVM (support vector machine), which classifies them. It is straightforward that the selection of the right number of new components is a core problem, and often, several trials are needed to find the best classification score, resulting in a time-consuming and computationally expensive process.

Our intent is to determine whether minimizing $sc(\mathbf{X};k)$ allows us to simplify the process, i.e., if it could be a good choice to simply select the first $\hat{k}$ components, where $\hat{k}$ minimizes $sc(\mathbf{X};k)$. For the numerical experiment, the procedure adopted in [78] is taken as a reference, and the results concerning the Indian Pines dataset are compared with the ones presented there. Accordingly, the training set for the SVM is composed of 10% of the samples in each class, randomly selected and normalized; these samples are the columns of the matrix $\mathbf{X}$ that is analyzed. As depicted in Figure 4, the value $\hat{k}$ which minimizes $sc(\mathbf{X};k)$ is 22, which is consistent with the best classification result for PCA+SVM obtained in [78], as shown in Figure 5.

In this case, the ϵ-dependent term e plays a key role in determining the trend of $sc(\mathbf{X};k)$ for two reasons: first, the arguments of the logarithms in the terms b and e have the same magnitude; second, the dimensions n and m of the matrix $\mathbf{X}$ are such that $n << m$, so that the term e overwhelms the term b as k grows. It turns out that, in this case, the selection of the quantization step ϵ is crucial. As in the first test, the presented results refer to $\epsilon = 10^{-8}$ and the required computing time has been about 1.10 s.

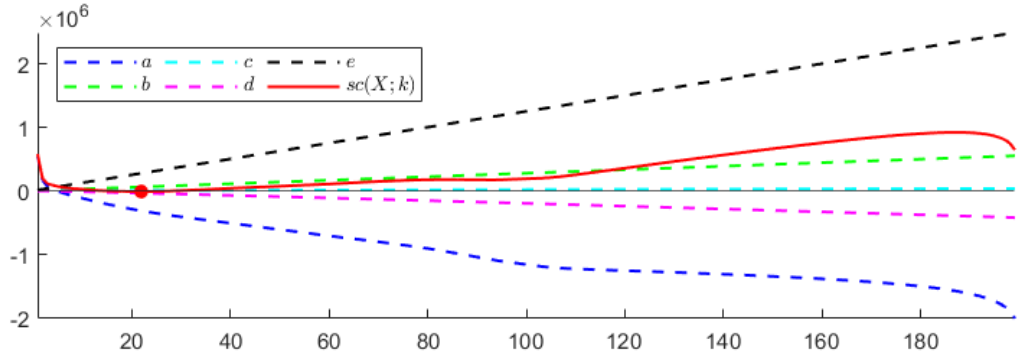

Figure 4. TEST 3. Plot of $sc(\mathbf{X};k)$ and its components versus the number of components k; the minimum is attained at $k = 22$.

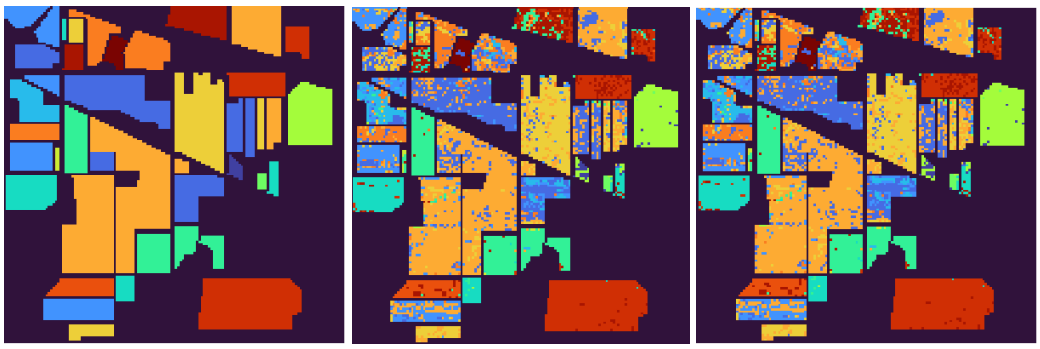

Figure 5. TEST 3. (**Left**) Ground-truth Indian Pines image; (**Middle**) classification image using the best result of PCA-SVM in [78]; (**Right**) classification image using the PCA-SVM method and the number of components estimated using the stochastic complexity, as in Equation (5).

To conclude this section, Table 1 contains the number of components selected using standard criteria for the selection of components in PCA, i.e., the percentage of the total

variance and Bartlett's test. The table refers to the three tests described above. As it can be observed, the MDL criterion is able to select the number of components closer or equal to the expected one in almost all tests, showing some robustness to the task. This is due to the fact that the MDL criterion tends to maximize the accuracy with the least cost. This confirms the potential of the MDL criterion in feature reduction procedures and offers a new and different approach to the solution of the selection of the best principal components' number.

Table 1. Number of principal components selected for the three tests by using different criteria: percentage of variance to be retained (90%, 95%, 99%), Bartlett's test with significance level α equal to 0.05 and 0.01, and the MDL criterion. The last column contains the expected number.

Test	90%	95%	99%	Bartlett's Test ($\alpha = 0.05$)	Bartlett's Test ($\alpha = 0.01$)	MDL	True Value
Test 1	1	1	1	30	30	5	5
Test 2	37	52	75	90	90	90	3
Test 2 (decimated data)	2	6	24	63	63	2	3
Test 3	2	6	27	161	159	22	22

5. Conclusions

This short review has shown some of the main features of MDL by referring to a few specific applications. MDL is appealing in data approximation-based problems, as it simply uses available data and models to make the best choice. In fact, the rationale of discarding the hypothesis that a 'true' distribution produced the current data is a conceptual step forward in data analysis. In addition, apart from the model selection problem, MDL has shown, in its different declinations, to be an effective tool for many other applications. The selection of the suitable number of features to adopt in the classification process is only the latest of the several applications where it plays a fundamental role. In addition, this specific use opens new possible ways in machine/deep learning that implicitly or explicitly depend on both the type and the number of features. On the other hand, as the presented simulations have shown, MDL often suffers from an explicit or implicit dependence on one or more parameters that have to be set. Usually, this is not a critical step, as setting them often is easier than competing approaches. However, this is one of the main points to be investigated in the future research.

Author Contributions: Formal analysis, D.V.; Methodology, V.B.; Software, M.L.C.; Supervision, D.V.; Writing—original draft, V.B. and M.L.C.; Writing—review & editing, V.B. and D.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by the Italian national research group GNCS (INdAM). This research has been accomplished within RITA (Research ITalian network on Approximation).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank the anonymous reviewers for their valuable comments and suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chandrashekar, G.; Sahin, F. A survey on feature selection methods. *Comput. Electr. Eng.* **2014**, *40*, 16–28. [CrossRef]
2. Ferreira, A.J.; Figueiredo, M.A.T. Efficient feature selection filters for high-dimensional data. *Pattern Recognit. Lett.* **2012**, *33*, 1794–1804. [CrossRef]

3. Jolliffe, I.; Cadima, J. Principal component analysis: A review and recent developments. *Philosphiocal Trans. A* **2016**, *374*, 20150202. [CrossRef]
4. Lee, D.; Seung, H. Learning the parts of objects by non-negative matrix factorization. *Nature* **1990**, *401*, 788–791. [CrossRef] [PubMed]
5. Tenenbaum, J.B.; De Silva, V.; Langford, J.C. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* **2000**, *290*, 2319–2323. [CrossRef] [PubMed]
6. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
7. McInnesand, L.; Healy, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* **2018**, arXiv:abs/1802.03426.
8. Vincent, P.; LaRochelle, H.; Bengio, Y.; Manzagol, P.A. Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th international conference on Machine learning-ICML'08, Helsinki, Finland, 5–9 July 2008; pp. 1096–1103.
9. Cox, M.; Cox, T. Multidimensional Scaling. In *Handbook of Data Visualization*; Springer Handbooks Comp. Statistics; Springer: Berlin/Heidelberg, Germany, 2008.
10. Rissanen, J. Modeling by the shortest data description. *Automatica* **1978**, *14*, 465–471. [CrossRef]
11. Rissanen, J. A universal prior for integers and estimation by minimum description length. *Ann. Stat.* **1983**, *11*, 416–431. [CrossRef]
12. Cover, T.; Thomas, J. *Elements of Information Theory*; Wiley Interscience: New York, NY, USA, 1991.
13. Myung, J.I.; Navarro, D.J.; Pitt, M.A. Model selection by normalized maximum likelihood. *J. Math. Psychol.* **2006**, *50*, 167–179. [CrossRef]
14. Grünwald, P.D.; Grunwald, A. *The Minimum Description Length Principle*; MIT Press: Cambridge, MA, USA, 2007.
15. Hu, B.; Rakthanmanon, T.; Hao, Y.; Evans, S.; Leonardi, S.; Keogh, E. Using the minimum description length to discover the intrinsic cardinality and dimansionality series. *Data Min. Knowl. Discov.* **2015**, *29*, 358–399. [CrossRef]
16. Cubero, R.J.; Marsili, M.; Roudi, Y. Minimum Description Length Codes Are Critical. *Entropy* **2018**, *20*, 755. [CrossRef] [PubMed]
17. Makalic, E.; Schmidt, D.F. Minimum Message Length Inference of the Exponential Distribution with Type I Censoring. *Entropy* **2021**, *23*, 1439. [CrossRef] [PubMed]
18. Adriaans, P.; Vitanyi, P.M.B. Approximation of the Two-Part MDL Code. *IEEE Trans. Inf. Theory* **2009**, *55*, 444–457. [CrossRef]
19. Murena, P.A.; Cornuéjols, A. Minimum Description Length Principle applied to structure adaptation for classification under concept drift. In Proceedings of the International Joint Conference on Neural Networks, Vancouver, BC, Canada, 24–29 July 2016; pp. 2842–2849.
20. Barron, A.; Rissanen, J.; Yu, B. The minimum description length principle in coding and modeling. *IEEE Trans. Inf. Theory* **1998**, *44*, 2743–2760. [CrossRef]
21. Grünwald, P. Minimum description length tutorial. In *Advances in Minimum Description Length: Theory and Applications*; Grünwald, P., Myung, I.J., Pitt, M.A., Eds.; MIT Press: Cambridge, MA, USA, 2005; pp. 23–80.
22. Hansen, M. H.; Yu, B. Minimum description length model selection criteria for generalized linear models. In *Lecture Notes–Monograph Series*; Institute of Mathematical Statistics: Beachwood, OH, USA, 2003; Volume 40, pp. 145–163.
23. Rissanen, J. Strong optimality of the normalized ml models as universal codes. *IEEE Trans. Inf. Theory* **2000**, *47*, 1712–1717. [CrossRef]
24. Bokde, D.; Girase, S.; Mukhopadhyay, D. Matrix factorization model in collaborative filtering algorithms: A survey. In Proceedings of the 4th International Conference on Advances in Computing, Communication and Control, Mumbai, India, 1–2 April 2015; pp. 136–146.
25. Eckart, C.; Young, G. The approximation of one matrix by another of lower rank. *Psychometrika* **1936**, *1*, 211–218. [CrossRef]
26. Udell, M.; Horn, C.; Zadeh, R.; Boyd, S. Generalized low rank models. *Found. Trends Mach. Learn.* **2016**, *9*, 1–118. [CrossRef]
27. Tavory, A. Determining Principal Component Cardinality Through the Principle of Minimum Description Length. In *Machine Learning, Optimization, and Data Science*; Nicosia, G, Ojha, V., Malfa, E.L., Jansen, G., Sciacca, V., Pardalos, P., Giuffrida, G., Umeton, R., Eds.: Springer: Berlin/Heidelberg, Germany, 2019; pp. 655–666; LOD 2019, LNCS 11943.
28. Grünwald, P.; Roos, T. Minimum description length revisited. *Int. J. Math. Ind.* **2019**, *11*, 1930001. [CrossRef]
29. Navarro, D.J.; Lee, M.D. Common and distinctive features in stimulus representation: A modified version of the contrast model. *Psychon. Bull. Rev.* **2004**, *11*, 961–974. [CrossRef]
30. Bruni, V.; Vitulano, D. An entropy based approach for SSIM speed up. *Signal Process.* **2017**, *135*, 198–209. [CrossRef]
31. Bruni, V.; Tartaglione, M.; Vitulano, D. A signal complexity-based approach for am–fm signal modes counting. *Mathematics* **2020**, *8*, 2170. [CrossRef]
32. Rissanen, J. *Stochastic Complexity in Statistical Inquiry*; World Scientific Publishing: Singapore, 1989.
33. Rissanen, J. Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Trans. Inf. Theory* **2001**, *47*, 1712–1717. [CrossRef]
34. Myung, I.J.; Pitt, M.A. Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychon. Bull. Rev.* **1997**, *4*, 79–95. [CrossRef]
35. Rissanen, J. MDL denoising. *IEEE Trans. Inf. Theory* **2000**, *46*, 2537–2542. [CrossRef]
36. Kontkanen, P.; Myllymaki, P.; Buntine, V.; Rissanen, J.; Tirri, H. *An MDL Framework for Data Clustering*; Helsinki Institute for Information Technology HIIT Technical Report; MIT Press: Cambridge, MA, USA, 2003.

37. Blier, L.; Ollivier, Y. The description length of deep learning models. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 3–8 December 2018; pp. 2220–2230.
38. Begum, N.; Hu, B.; Rakthanmanon, T.; Keogh, E. Towards a minimum description length based stopping criterion for semi-supervised time series classification. In Proceedings of the IEEE 14th International Conference on Information Reuse & Integration (2013), San Francisco, CA, USA, 14–16 August 2013; pp. 333–340.
39. Yamanishi, K.; Fukushima, S. Model Change Detection With the MDL Principle. *IEEE Trans. Inf. Theory* **2018**, *64*, 6115–6126. [CrossRef]
40. Yamanishi, K. Descriptive Dimensionality and Its Characterization of MDL-based Learning and Change Detection. *arXiv* **2019**, arXiv:1910.11540.
41. Hinton, G.E.; van Camp, D. Keeping Neural Networks Simple by Minimizing the Description Length of the Weights. In Proceedings of the 6th Annual Conference on Computational Learning Theory, Santa Cruz, CA, USA, 26–28 July 1993; pp. 5–13.
42. Lin, B. Regularity Normalization: Neuroscience-Inspired Unsupervised Attention across Neural Network Layers. *Entropy* **2022**, *24*, 59. [CrossRef]
43. Saxe, A.M.; Bansal, Y.; Dapello, J.; Advani, M.; Kolchinsky, A.; Tracey, B.D.; Cox, D.D. On the information bottleneck theory of deep learning. *J. Stat. Mech. Theory Exp.* **2019**, *2019*, 124020. [CrossRef]
44. Tishby, N.; Zaslavsky, N. Deep learning and the information bottleneck principle. In Proceedings of the IEEE Information Theory Workshop, Jerusalem, Israel, 11–15 October 2015; pp. 1–5.
45. Fang, J.; Ouyang, H.; Shen, V.; Dougherty, V.; Liu, W. Using the minimum description length principle to reduce the rate of false positives of best-fit algorithms. *EURASIP J. Bioinform. Syst. Biol.* **2014**, *13*, 13. [CrossRef]
46. Chaitankar, V.; Zhang, C.; Ghosh, P.; Gong, P.; Perkins, E.J.; Deng, Y. Predictive minimum description length principle approach to inferring gene regulatory networks. *Adv. Exp. Med. Biol.* **2011**, *696*, 37–43. [PubMed]
47. Fade, J.; Lefebvre, S.; Cézard, N. Minimum description length approach for unsupervised spectral unmixing of multiple interfering gas species. *Opt. Express* **2011**, *19*, 13862–13872. [CrossRef] [PubMed]
48. Wallace, R.S.; Kanade, T. Finding natural clusters having minimum description length. In Proceedings of the 10th International Conference on Pattern Recognition, Atlantic City, NJ, USA, 16–21 June 1990.
49. Hirai, S.; Yamanishi, K. Detecting Changes of Clustering Structures Using Normalized Maximum Likelihood Coding. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, Beijing, China, 12–16 August 2012; pp. 343–351.
50. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Autom. Control* **1974**, *19*, 716–723. [CrossRef]
51. Schwarz, G. Estimating the dimension of a model. *Ann. Stat.* **1978**, *6*, 461–464. [CrossRef]
52. Al-Qurabat, A.K.M.; Abou Jaoude, C.; Idrees, A.K. Two Tier Data Reduction Technique for Reducing Data Transmission in IoT Sensors. In Proceedings of the 15th International Wireless Communications & Mobile Computing Conference, Tangier, Morocco, 4–28 June 2019.
53. Squires, S.; Prügel-Bennett, A.; Niranjan, M. Minimum description length as an objective function for non-negative matrix factorization. *arXiv* **2019**, arXiv:1902.01632.
54. Pandey, G.; Dukkipati, A. Minimum description length principle for maximum entropy model selection. In Proceedings of the IEEE International Symposium on Information Theory, Istanbul, Turkey, 7–12 July 2013; pp. 1521–1525.
55. Shamir, G.I. Minimum description length (MDL) regularization for online learning. In Proceedings of the 1st International Workshop on Feature Extraction: Modern Questions and Challenges at NIPS 2015, PMLR 44:260-276, Montreal, QC, Canada, 11 December 2015.
56. Thodberg, H.H. Minimum Description Length Shape and Appearance Models. In Proceedings of the Biennial International Conference on Information Processing in Medical Imaging IPMI, Ambleside, UK, 20–25 July 2003; pp. 51–62.
57. Bariatti, F.; Cellier, P.; Ferré, S.; Berthold, M.R.; Feelders, A.; Krempl, G. GraphMDL: Graph Pattern Selection Based on Minimum Description Length. In *Advances in Intelligent Data Analysis XVIII*; Springer International Publishing: Berlin/Heidelberg, Germany, 2020; pp. 54–66.
58. Jang, K.E.; Tak, S.; Jung, J.; Jang, J.; Jeong, Y.; Ye, J.C. Wavelet minimum description length detrending for near-infrared spectroscopy. *J. Biomed. Opt.* **2009**, *14*, 034004. [CrossRef]
59. Hamid, E.Y.; Kawasaki, Z.I. Wavelet-based data compression of power system disturbances using the minimum description length criterion. *IEEE Trans. Power Deliv.* **2002**, *17*, 460–466. [CrossRef]
60. Ojanen, J.; Heikkonen, J. A soft thresholding approach for MDL denoising. In Proceedings of the 15th European Signal Processing Conference, Poznan, Poland, 3–7 September 2007; pp. 1083–1087.
61. Kumar, V.; Heikkonen, J.; Rissanen, J.; Kaski, K. Minimum description length denoising with histogram models. *IEEE Trans. Signal Process.* **2006**, *54*, 2922–2928. [CrossRef]
62. Wettig, H.; Kontkanen, P.; Myllymaki, P. Calculating the Normalized Maximum Likelihood Distribution for Bayesian Forests. In Proceedings of the IADIS International Conference Intelligent Systems and Agents, Lisbon, Portugal, 5–8 October 2007.
63. Jackson, D.A. Stopping rules in principal components analysis: A comparison of heuristical and statistical approaches. *Ecology* **1993**, *74*, 2204–2214. [CrossRef]
64. Jolliffe, I. *Principal Component Analysis*; Wiley Online Library: Hoboken, NJ, USA 2005.

65. Okamoto, M. Optimality of principal components. In *Multivariate Analysis II*; Krishnaiah, P.R., Ed.; Academic Press: New York, NY, USA, 1969; pp. 673–685.
66. McCabe, G.P. Principal variables. *Technometrics* **1984**, *26*, 137–144. [CrossRef]
67. Cadima, J.; Cerdeira, J.O.; Minhoto, M. Computational aspects of algorithms for variable selection in the context of principal components. *Comp. Stat. Data Anal.* **2004**, *47*, 225–236. [CrossRef]
68. R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2015.
69. Saccenti, E.; Camacho, J. Determining the number of components in principal components analysis: A comparison of statistical, cross-validation and approximated methods. *Chemom. Intell. Lab. Syst.* **2015**, *149*, 99–116. [CrossRef]
70. Gabriel, K.R. The biplot graphical display of matrices with application to principal component analysis. *Biometrika* **1971**, *58*, 453–467. [CrossRef]
71. Cadima, J.; Jolliffe, I.T. On relationships between uncentred and column-centred principal component analysis. *Pak. J. Stat.* **2009**, *25*, 473–503.
72. Demmel, J.W. Applied Numerical Linear Algebra. In Proceedings of the SIAM, New Orleans, LA, USA, 13–15 July 1997.
73. Mirsky, L. Symmetric gauge functions and unitarily invariant norms. *Q. J. Math.* **1960**, *11*, 50–59. [CrossRef]
74. Baumgardner, M.F.; Biehl, L.L.; Landgrebe, D.A. *220 Band AVIRIS Hyperspectral Image Data Set: June 12, 1992 Indian Pine Test Site 3*; Purdue University Research Repository; Purdue University: West Lafayette, IN, USA, 2015.
75. Goldberger, A.L.; Amaral, L.; Glass, L.; Hausdorff, J.M.; Ivanov, P.; Mark, R.G.; Mietus, J.E.; Moody, G.B.; Peng, C.K.; Stanley, H.E. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation* **2000**, *101*, 215–220. [CrossRef]
76. Mallat, S. *A Wavelet Tour of Signal Processing*, 2nd ed.; Academic Press: Cambridge, MA, USA, 1999.
77. Gersho, A.; Gray, R.M. *Vector Quantization and Signal Compression*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 1991.
78. Shambulinga, M.; Sadashivappa, G. Hyperspectral Image Classification using Support Vector Machine with Guided Image Filter. *Int. J. Adv. Comput. Sci. Appl.* **2019**, *10*, 271–276.