Identification Rates for Block-correlated Gaussian Sources

Hanwei Wu, Qiwen Wang and Markus Flierl
School of Electrical Engineering and Computer Science
KTH Royal Insititute of Technology
Stockholm, Sweden
Email: {hanwei, qiwenw, mflierl}@kth.se

***Abstract*—Among many current data processing systems, the objectives are often not the reproduction of data, but to compute some answers based on the data responding to some queries. The similarity identification task is to identify the items in a database which are similar to a given query item regarding to a certain metric. The problem of compression for similarity identification has been studied in [1]. Unlike classic compression problems, the focus is not on reconstructing the original data. Instead, the compression rate is determined by the desired reliability of the answers. Specifically, the information measure *identification rate* of a compression scheme characterizes the minimum compression rate that can be achieved which guarantees reliable answers with respect to a given similarity threshold. In this paper, we study the component-based quadratic similarity identification for correlated sources. The blocks are first decorrelated by Karhunen-Loève transform. Then, the decorrelated data is processed by a distinct D-admissible system for each component. We derive the *identification rate* of component-based scheme for block-correlated Gaussian sources. In addition, we characterize the *identification rate* of a special setting where any information regarding to the component similarity thresholds is unknown while only the similarity threshold of the whole scheme is given. Furthermore, we prove that block-correlated Gaussian sources are the "most difficult" to code under the special setting.**

I. Introduction

The problem of efficient identification and data retrieval from large databases has become more relevant in recent years. Similarity identification requires that a database returns all the data items which are similar to a given query as defined by a similarity threshold. The notion of similarity is often defined by a specific metric measure, such as the Euclidean distance or the Hamming distance. It is required that false negative errors are not permitted in the retrieval process, since they cannot be detected by further processing. This is important for some applications, such as security cameras and criminal forensic databases. On the other hand, although false positive errors can be detected by further verification, they increase the computational cost on the server side, and hence, reduce efficiency. Therefore, the tradeoff between the compression rate and the reliability of the answers to a given query is of interest.

The problem of similarity identification of compressed data was first studied in [2] from an information-theoretic viewpoint. In this work, both false positive and false negative errors are allowed, as long as the error probability vanishes with the data block-length. Our setting though is closely related to the problem of compression for similarity queries as introduced in [3], [4]. In [3], [4] and this work, $false\ negative$ errors are not permitted. [3], [4] study the problem from an information-theoretic viewpoint and introduce the term $identification\ rate$. It characterizes the minimum compression rate that permits query answers with a vanishing false positive probability, while false negative errors are not allowed. [3], [4] provide the identification rate for Gaussian sources with quadratic distortion and for binary sources with Hamming distance. [3] also proves that, as with classical compression, the Gaussian source requires the largest compression rate among sources with a given variance.

Since most real-world data is correlated, it is of interest to investigate similarity identification schemes for correlated sources. [5] uses lossy compression as a building block to construct the TC-$\triangle$ (Type Covering signatures and triangle-inequality decision rule) and LC-$\triangle$ (Lossy Compression signatures and triangle-inequality decision rule) schemes. The results in [5] show that the compression rate of TC-$\triangle$ can achieve the identification rate for the binary-Hamming case. In [6], the authors present a shape-gain quantizer for i.i.d. Gaussian sequences: Scalar quantization is applied to the magnitude of the data vector. The shape (the projection on the unit sphere) is quantized using a warpped spherical code [7]. [8] proposes tree-structured vector quantizers that hierarchically cluster the data using k-center clustering. In [9], the authors compare two transform based similarity identification schemes to cope with exponentially growing codebooks for high-dimensional data. One of the proposed schemes, that is, the component-based approach, shows both good performance and low search complexity. However, for correlated sources, no analytical results for the identification rate of above schemes are provided.

The outline of this paper is as follows: In Section 2, we give a brief description of the problem's background and key concepts [1] for more detailed background and problem description. In Section 3, we discuss component-based approaches. We first derive the identification rate of component-based approaches for block-correlated Gaussian sources. Then, we characterize the identification rate for a special setting and show that the block-correlated Gaussian sources are the "most difficult" to

[1]We follow the problem setup and adopt most notations in [3] and [4]. Therefore, we refer to [3] and [4]

code. The conclusions are given in Section 4.

The notational conventions in this work are as follows. Uppercase nonboldface symbols such as X are used to denote random variables; and lowercase nonboldface symbols such as x are used to denote sample values of those random variables. Vectors and matrices of random variables or their sample values are denoted by boldface symbols. For example, $\mathbf{X}$ and $\mathbf{x}$ are vectors (or sometimes matrices from the context) of random variables X and its sample values x, respectively. The ith entry of a vector $\mathbf{X}$ is denoted by X_i.

II. Quadratic Similarity Queries

Let $\mathbf{y} = (y_1, y_2, .., y_n)^T$ denote the query sequence and $\mathbf{x} = (x_1, x_2, .., x_n)^T$ as the data sequence. A rate-R_{ID} identification system (T, g) consists of a *signature assignment* function: $T : \mathbb{R}^n \to \{1, 2, \cdots, 2^{nR_{\text{ID}}}\}$, and a *query function* g:$\{1, 2, \cdots, 2^{nR_{\text{ID}}}\} \times \mathbb{R}^n \to \{\text{no}, \text{maybe}\}$. The database keeps only a short signature $T(\mathbf{x})$ for each $\mathbf{x}$. And the output decision *no* or *maybe* of a query function indicates whether $\mathbf{x}$ and $\mathbf{y}$ are probably D_{ID}-similar or not. The sequences $\mathbf{x}$ and $\mathbf{y}$ are called D_{ID}-*similar* if $d(\mathbf{x}, \mathbf{y}) \leq D_{\text{ID}}$, where

$$d(\mathbf{x}, \mathbf{y}) \triangleq \frac{1}{n} \sum_{i=1}^{n} \rho(x_i, y_i), \tag{1}$$

ρ is an arbitrary per-letter distance measure, and D_{ID} is the *similarity threshold*. Specifically, the quadratic similarity is

$$d(\mathbf{x}, \mathbf{y}) \triangleq \frac{1}{n} \|\mathbf{x} - \mathbf{y}\|^2 = \frac{1}{n} \sum_{i=1}^{n} \|x_i - y_i\|^2, \tag{2}$$

where $\|\cdot\|$ is the standard Euclidean norm.

A similarity query retrieves all data items that are D_{ID}-similar. A scheme is called D_{ID}-*admissible* if we obtain $g(T(\mathbf{x}), \mathbf{y}) = \text{maybe}$ for any pair of data item and query $(\mathbf{x}, \mathbf{y})$ which is D_{ID}-*similar*.

Now, consider a probabilistic model for database and query. The objective is to design $D_{\text{ID}}-$admissible systems that minimize the probability of the output *maybe* for given distributions of database vectors $\mathbf{X}$ and query vectors $\mathbf{Y}$. According to [3], for a $D_{\text{ID}}-$admissible system, this probability is calculated as

$$\begin{aligned} &\Pr\{g(T(\mathbf{X}), \mathbf{Y}) = \text{maybe}\} \\ &= \Pr\{g(T(\mathbf{X}), \mathbf{Y}) = \text{maybe} | d(\mathbf{X}, \mathbf{Y}) \leq D_{\text{ID}}\} \\ &\quad \Pr\{d(\mathbf{X}, \mathbf{Y}) \leq D_{\text{ID}}\} \\ &+ \Pr\{g(T(\mathbf{X}), \mathbf{Y}) = \text{maybe}, d(\mathbf{X}, \mathbf{Y}) > D_{\text{ID}}\} \\ &= \Pr\{d(\mathbf{X}, \mathbf{Y}) \leq D_{\text{ID}}\} + \Pr(\varepsilon), \end{aligned} \tag{3}$$

where the second equality follows from $\Pr\{g(T(\mathbf{X}), \mathbf{Y}) = \text{maybe} | d(\mathbf{X}, \mathbf{Y}) \leq D_{\text{ID}}\} = 1$ by the requirement of D_{ID} admissibility. Hence, minimizing (3) is equivalent to minimizing the probability of false positives $\Pr(\varepsilon)$. That is, the probability $\Pr\{g(T(\mathbf{X}), \mathbf{Y}) = \text{maybe}\}$ can be used as a performance measure for the investigated schemes. In the following, we use the abbreviation $\Pr\{\text{maybe}\}$ for the probability that a system outputs maybe.

For given distributions P_X and P_Y and a similarity threshold D_{ID}, a rate R is said to be D_{ID}-*achievable* if there exits a sequence of D_{ID}-admissible schemes $(T^{(n)}, g^{(n)})$ that can achieve a vanishing $\Pr\{\text{maybe}\}$ as n approaches infinity:

$$\lim_{n \to \infty} \Pr\{g^{(n)}(T^{(n)}(\mathbf{X}), \mathbf{Y}) = \text{maybe}\} = 0. \tag{4}$$

The *identification rate* R^*_{ID} of the source is defined as the infimum of all D_{ID}-*achievable* rates.

III. Identification Rates of Dependent Normally-Distributed Variables

A. Component-based Approach

Consider a concatenation of N independent blocks of correlated zero-mean Gaussian random variables with blocklength M for D_{ID}-similarity identification, where $n = MN$ is the length of the source sequence. The blocks can be decorrelated by the KLT. For each $m \in \{1, 2, \ldots, M\}$, we first collect the m-th elements of all N blocks and arrange them into M separate subsequences with subsequence length N. The subsequences can be represented by an $M \times N$ matrix $\tilde{\mathbf{X}}_D$, where each row represents one subsequence. Let $\Sigma_{\tilde{\mathbf{X}}_D} = \frac{1}{N}\mathbb{E}[\tilde{\mathbf{X}}_D \tilde{\mathbf{X}}_D^T]$ be the covariance matrix of $\tilde{\mathbf{X}}_D$. Let Φ be the eigenmatrix of $\Sigma_{\tilde{\mathbf{X}}_D}$, $\Sigma_{\tilde{\mathbf{X}}_D}\Phi = \Phi\Lambda$, where Λ is a diagonal matrix with eigenvalues λ_i as diagonal entries. Since the covariance matrix is real symmetric, then its eigenvectors are orthonormal, $\Phi^T\Phi = \mathbf{I}$. Hence, the KLT uses the transpose of the eigenmatrix Φ of $\Sigma_{\tilde{\mathbf{X}}_D}$ to decorrelate the data:

$$\frac{1}{N}\mathbb{E}[\Phi^T \tilde{\mathbf{X}}_D (\Phi^T \tilde{\mathbf{X}}_D)^T] = \Phi^T \Sigma_{\tilde{\mathbf{X}}_D} \Phi = \Lambda \tag{5}$$

Since uncorrelated jointly distributed Gaussian random variables imply independence, the KLT outputs M independent subsequences consisting of i.i.d. Gaussian variables with corresponding variances. We use a $D_{\text{ID}}^{(m)}$-admissible system for the m-th row of the data matrix after the KLT $\mathbf{X}_D = \Phi^T \tilde{\mathbf{X}}_D$. We call it the m-th component, and $D_{\text{ID}}^{(m)}$ is the similarity threshold for the m-th component.

The query $\mathbf{Y}$ has the same distribution as the database vectors $\mathbf{X}$. Hence we can use the same KLT to decorrelate the query. The m-th component scheme answers *maybe* if the transformed m-th query-database pair $(\mathbf{x}^{(m)}, \mathbf{y}^{(m)})$ satisfies $\frac{1}{N}\sum_{i=1}^{N} \|x_i^{(m)} - y_i^{(m)}\|^2 \leq D_{\text{ID}}^{(m)}$.

We define the achievable rate of the component-based scheme as the average of the component rates $R_{\text{ID}}^{(m)}$, *i.e.,*

$$R_{\text{ID}} = \frac{1}{M} \sum_{m=1}^{M} R_{\text{ID}}^{(m)}. \tag{6}$$

The similarity threshold for a component-based scheme is defined as the upper bound

$$\begin{aligned} d(x, y) &= \frac{1}{M} \sum_{m=1}^{M} \frac{1}{N} \sum_{i=1}^{N} ||x_i^{(m)} - y_i^{(m)}||^2 \\ &\leq \frac{1}{M} \sum_{m=1}^{M} D_{\text{ID}}^{(m)} \\ &:= D_{\text{ID}}. \end{aligned} \tag{7}$$

Note, the eigenbasis of the KLT is orthonormal and the distances in the component space are preserved.

B. Identification Rate R_{ID}^{C*}

We define the *identification rate of a component-based scheme* R_{ID}^{C*} as the infimum of all D_{ID}-achievable rates which can be achieved by a component-based approach as defined above. The identification rate R_{ID}^{C*} for correlated Gaussian sources is given in Theorem 1.

Theorem 1. *Consider correlated Gaussian sources that are decorrelated by a KLT with component variances $\sigma_1^2 \geq \cdots \geq \sigma_M^2$. An M-component scheme with $D_{ID}^{(m)}$-admissible systems for each component m and logic AND decision for the final output, can achieve the identification rate*

$$R_{ID}^{C*} = \frac{1}{M}\sum_{m=1}^{M} \log\left(\frac{2\sigma_m^2}{2\sigma_m^2 - D_{ID}^{(m)}}\right), \tag{8}$$

where

$$D_{ID}^{(m)} = \max\left(0, 2\sigma_m^2 - \frac{1}{v\ln(2)}\right) \tag{9}$$

with $v \in \left\{\min_m\left(\frac{1}{2\ln(2)\sigma_m^2}\right), \infty\right\}$.

Proof. Each component of a M-component scheme uses a $D_{\text{ID}}^{(m)}$-admissible system. An overall output of maybe can only be achieved if all component systems output maybe. For a given D_{ID}, the identification rate, i.e., the infimum of the achievable rate, is obtained if the component similarities $D_{\text{ID}}^{(1)}, \ldots, D_{\text{ID}}^{(M)}$ satisfy

$$\begin{aligned} \min_{D_{\text{ID}}^{(1)},\ldots,D_{\text{ID}}^{(M)}} R_{\text{ID}} &= \frac{1}{M}\sum_{m=1}^{M} R_{\text{ID}}^{(m)}\left(D_{\text{ID}}^{(m)}\right) \\ \text{s.t.} \quad & \frac{1}{M}\sum_{m=1}^{M} D_{\text{ID}}^{(m)} \geq D_{\text{ID}}, \\ \text{s.t.} \quad & D_{\text{ID}}^{(m)} \geq 0. \end{aligned} \tag{10}$$

Note, all rate similarity functions $R_{\text{ID}}^{(m)}\left(D_{\text{ID}}^{(m)}\right)$ of the components are convex and strictly increasing. Hence, we consider the equivalent problem

$$\begin{aligned} \min \quad & J = R_{\text{ID}} - vD_{\text{ID}} \\ \text{s.t.} \quad & D_{\text{ID}}^{(m)} \geq 0, \end{aligned} \tag{11}$$

where v is a positive Lagrangian multiplier. By definition, similarities are non-negative.

Recall that the identification rate for i.i.d. Gaussian sources is [4]

$$R_{\text{ID}}^{*}(D_{\text{ID}}) = \begin{cases} \log(\frac{2\sigma^2}{2\sigma^2 - D_{\text{ID}}}) & \text{for } 0 \leq D_{\text{ID}} < 2\sigma^2 \\ \infty & \text{for } D_{\text{ID}} \geq 2\sigma^2. \end{cases} \tag{12}$$

The derivative of the cost function J with respect to $D_{\text{ID}}^{(m)}$ is:

$$\frac{\partial J}{\partial D_{\text{ID}}^{(m)}} = \frac{1}{\ln(2)(2\sigma_m^2 - D_{\text{ID}}^{(m)})} - v \tag{13}$$

By setting (13) to zero, we obtain that $D_{\text{ID}}^{(m)}$ is determined by the component variance σ_m^2 and the value of v

$$D_{\text{ID}}^{(m)} = 2\sigma_m^2 - \frac{1}{v\ln(2)}. \tag{14}$$

In order to satisfy the non-negative constraint of $D_{\text{ID}}^{(m)}$, each component is only activated when the multiplier v is larger than its $v_{\min}^{(m)}$

$$v \geq v_{\min}^{(m)} = \frac{1}{2\sigma_m^2\ln(2)}. \tag{15}$$

Note, the smallest v for the scheme is $\frac{1}{2\sigma_1^2\ln(2)}$.

Then, we can sweep over permitted values of the Lagrangian parameter $v \in \left\{\frac{1}{2\sigma_1^2\ln(2)}, \infty\right\}$ to obtain the $(R_{\text{ID}}^{C*}, D_{\text{ID}})$ curve, and each component similarity threshold is determined by (9). □

The theorem shows that the optimal identification rate can be achieved by activating the components according to their variances after the KLT. At the lowest rate, only the component with the largest variance is activated. In this case, the M-component scheme uses only one component. Then, as the rate increases, the remaining components are activated in the order of their component variances. The activated components operate according to the Pareto condition.

C. Special Setting

In this section, we discuss the special setting that only the similarity threshold D_{ID} of an M-component scheme is given, and any information regarding the component similarity thresholds $D_{\text{ID}}^{(m)}$ is unknown. For this special setting, we first need to guarantee that the component-based schemes are still D_{ID}-admissible after applying the transform. This means that the similarity queries in the transform domain should retrieve all the true positive data items that should be retrieved in the original space. Proposition 1 shows the constraints imposed on $D_{\text{ID}}^{(m)}$ of each component to maintain the D_{ID}-admissibility of the scheme.

Proposition 1. *The D_{ID}-admissibility of a M-component scheme is maintained if each individual component uses an MD_{ID}-admissible scheme.*

Proof. In the original space, a vector $\tilde{\mathbf{x}}$ with vector length $n = NM$ in the database should be labeled as *maybe* if

$$\frac{1}{M}\sum_{m=1}^{M}\frac{1}{N}\|\tilde{\mathbf{x}}^{(m)} - \tilde{\mathbf{y}}^{(m)}\|^2 \leq D_{\text{ID}} \tag{16}$$

$$\sum_{m=1}^{M}\frac{1}{N}\|\tilde{\mathbf{x}}^{(m)} - \tilde{\mathbf{y}}^{(m)}\|^2 \leq MD_{\text{ID}}. \tag{17}$$

Since the orthogonal transform preserves the Euclidean norm, the D_{ID} constraint is still valid in the transform domain. Then, it is easy to see that if one component is with $\frac{1}{N}\|\tilde{\mathbf{x}}^{(m)} - \tilde{\mathbf{y}}^{(m)}\|^2 > MD_{\text{ID}}$, then the whole vector $\mathbf{x}$ can not be D_{ID}-similar to the query. Hence, the D_{ID}-similarity of the M-component scheme requires that every component should satisfy

$$\frac{1}{N}\|\tilde{\mathbf{x}}^{(m)} - \tilde{\mathbf{y}}^{(m)}\|^2 \leq MD_{\text{ID}} = D_{\text{ID}}^{(m)}. \tag{18}$$

□

We define the identification rate $R_{\text{ID}}^{\text{CS}}$ as the infimum of D_{ID}-achievable rates that can be achieved for this special setting. Equipped with Proposition 1, we give $R_{\text{ID}}^{\text{CS}}$ in Proposition 2.

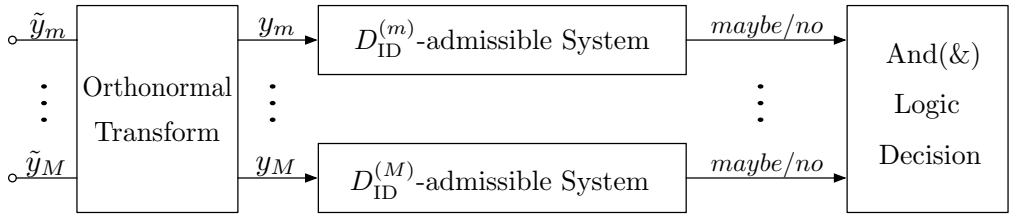

Fig. 1: Component-based approach for similarity identification.

Proposition 2. *The R_{ID}^{CS} for correlated Gaussian sources is*

$$R_{\mathrm{ID}}^{\mathrm{CS}} = \begin{cases} \frac{1}{M}\log\frac{1}{1-\frac{MD_{\mathrm{ID}}}{2\max(\sigma_m^2)}} & \text{for } 0 \le D_{\mathrm{ID}} < \frac{2}{M}\max(\sigma_m^2) \\ \infty & \text{for } D_{\mathrm{ID}} \ge \frac{2}{M}\max(\sigma_m^2). \end{cases} \tag{19}$$

Proof. The database vectors $\mathbf{x}$ are labeled as *maybe* if and only if all its transform components are determined as *maybe* with similarity threshold MD_{ID}. Define the event $A_m = \left\{d\left(T_m^{-1}\left(T_m\left(\mathbf{X}^{(m)}\right)\right), \mathbf{Y}^{(m)}\right) \le MD_{\mathrm{ID}}\right\}$, where $T_m(\cdot)$ is the signature function for the m-th component, and where $T_m^{-1}(k) = \{\mathbf{x} : T_m(\mathbf{x}) = k\}$ represents the set of vectors that have the same signature. Since the transform components of Gaussian sources are independent, we can write $\Pr\{\text{maybe}\}$ for the overall scheme as

$$\Pr\{A_1 \cap A_2 \cap, \ldots, \cap A_M\} = \prod_{m=1}^{M}\Pr\{A_m\} = \prod_{m=1}^{M} P_m, \tag{20}$$

where P_m is $\Pr\{\text{maybe}\}$ of the m-th component. Note, $\Pr\{\text{maybe}\}$ converges to 1 if $R < R_{\mathrm{ID}}$. Hence, we have $P_m = 1$ if the component is assigned a rate that is smaller than its corresponding $R_{\mathrm{ID}}^{*(m)}(MD_{\mathrm{ID}})$. The scheme's $\Pr\{\text{maybe}\}$ (20) is a product of component $\Pr\{\text{maybe}\}$. Therefore, to make sure $\Pr\{\text{maybe}\}$ vanishes, it only requires one of the M schemes to be assigned a rate slightly over its corresponding identification rate, $R^* = R_{\mathrm{ID}}^{*(m)}(MD_{\mathrm{ID}}) + \triangle R$, where $\triangle R$ is arbitrarily small.

Since R_{ID}^* of i.i.d. Gaussian sources increases along with the ratio of $\frac{D_{\mathrm{ID}}}{\sigma^2}$, we only need the component with the largest variance to operate with a rate above its corresponding identification rate. Since the component variances are equal to their corresponding eigenvalues, then we can write the $R_{\mathrm{ID}}^{\mathrm{CS}}$ as (19). □

When compared with R_{ID}^{C*}, we can see that $R_{\mathrm{ID}}^{\mathrm{CS}}$ is optimal in the lowest rate region when only the component with the largest eigenvalue is activated.

Finally, we compare $R_{\mathrm{ID}}^{\mathrm{CS}}$ with the rate of a scheme without transform, i.e., ignoring the correlation among the Gaussian sources with variance σ^2:

$$R_{\mathrm{ID}}^{WT} - R_{\mathrm{ID}}^{\mathrm{CS}} = \tag{21}$$

$$= \frac{1}{M}\left(\log\left(\frac{1}{1-\frac{MD_{\mathrm{ID}}}{2\sigma^2}}\right) - \log\left(\frac{1}{1-\frac{MD_{\mathrm{ID}}}{2\max(\sigma_m^2)}}\right)\right) \tag{22}$$

$$= \frac{1}{M}\log\left(\frac{\sigma^2}{\max(\sigma_i^2)}\frac{2\max(\sigma_m^2) - MD_{\mathrm{ID}}}{2\sigma^2 - MD_{\mathrm{ID}}}\right) \tag{23}$$

In the following, we consider non-Gaussian correlated sources and derive an upper bound on the identification rate $R_{\mathrm{ID}}^{\mathrm{CS}}$, achieved by the component-based approach under the special setting. Moreover, it can be concluded that for all correlated sources with the same largest component variance, Gaussian sources require the highest identification rate $R_{\mathrm{ID}}^{\mathrm{CS}}$. Proposition 3 below summarizes the result.

Proposition 3. *Given a D_{ID}-admissible M-component scheme where a correlated source is characterized by the largest finite KLT component variance σ^2 ($\sigma^2 > \frac{MD_{ID}}{2}$), then the rate R_{ID}^{CS} achieved by the component-based approach under the special setting for this source is bounded from above by $\frac{1}{M}\log\frac{1}{1-\frac{MD_{ID}}{2\sigma^2}}$.*

Proof. According to the Fréchet inequalities, we can write $\Pr\{\text{maybe}\}$ for the overall scheme as

$$\Pr\{A_1 \cap, \ldots, \cap A_M\} \le \min(\Pr(A_1), \cdots, \Pr(A_M)) \tag{24}$$

Similar to the proof of Proposition 2, it suffices if we only require that one component achieves a vanishing $\Pr\{\text{maybe}\}$. The other components can simply follow a trivial scheme with rate 0, and always output maybe. Therefore, the identification rate $R_{\mathrm{ID}}^{*(m)}$ of any component is a D_{ID}-achievable rate for the M-component scheme; and $R_{\mathrm{ID}}^{\mathrm{CS}}$ for a non-Gaussian source is the $R_{\mathrm{ID}}^{*(m)}$ of the component with the largest variance σ^2 divided by M. Theorem 6 of [10] shows that Gaussian sources require the largest R_{ID}^* among all sources with the same variance. Hence, if a correlated non-Gaussian source obtains the largest component variance σ^2 which is the same as the largest component variance of a correlated Gaussian sources after KLT, it follows that the achieved rate $R_{\mathrm{ID}}^{\mathrm{CS}}$ for the non-Gaussian source is bounded from above by the $R_{\mathrm{ID}}^{\mathrm{CS}}$ (19) for correlated Gaussian sources, *i.e.,* $\frac{1}{M}\log\frac{1}{1-\frac{MD_{\mathrm{ID}}}{2\sigma^2}}$. □

Example 1. *Consider the case of two correlated Gaussian variables with $\mathbb{E}[\tilde{X}_1] = \mathbb{E}[\tilde{X}_2] = 0$, $\mathbb{E}[\tilde{X}_1^2] = \mathbb{E}[\tilde{X}_2^2] = \sigma^2$ and $\mathbb{E}[\tilde{X}_1\tilde{X}_2] = \rho$, where ρ is the correlation coefficient. After applying the Karhunen-Loève transform, the resulting variances are $\mathbb{E}[X_1^2] = \sigma_1^2 = \sigma^2(1+\rho)$, $\mathbb{E}[X_2^2] = \sigma_2^2 = \sigma^2(1-\rho)$ and $\mathbb{E}[X_1X_2] = 0$. Therefore, the R_{ID}^{CS} for two correlated Gaussian random variables is*

$$R_{\mathrm{ID}}^{\mathrm{CS}} = \begin{cases} \frac{1}{2}\log\frac{1}{1-\frac{D_{\mathrm{ID}}}{\sigma_1^2}} & \text{for } 0 \le D_{\mathrm{ID}} < \sigma_1^2 \\ \infty & \text{for } D_{\mathrm{ID}} \ge \sigma_1^2. \end{cases} \tag{25}$$

In Fig. (2), we compare $R_{\mathrm{ID}}^{\mathrm{CS}}$ for Gaussian sources with different correlation coefficients. It can be observed that the scheme requires higher rates to achieve a vanishing $\Pr\{\text{maybe}\}$ as the data correlation becomes weaker.

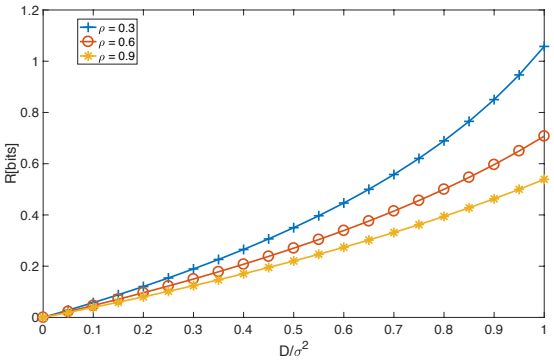

Fig. 2: Comparison of the component-based $R_{\mathrm{ID}}^{\mathrm{CS}}$ for Gaussian sources with different correlation coefficients.

In addition, we consider the special case $\rho = 1$, where the two-dimensional Gaussian vectors are completely determined by one Gaussian random variable. Hence, R_{ID}^{*} of the complete correlated sources should be half of the identification rate for i.i.d. random variables. Fig. (3) shows that the $R_{\mathrm{ID}}^{\mathrm{CS}}$ achieved by the scheme is half of the R_{ID}^{*} of i.i.d. Gaussian source. Therefore, the efficiency of the component-based scheme depends on the correlation of the data.

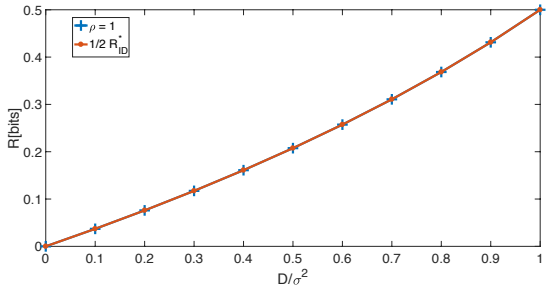

Fig. 3: Comparison of the component-based $R_{\mathrm{ID}}^{\mathrm{CS}}$ for $\rho = 1$ with R_{ID}^{*} of i.i.d. Gaussian sources.

Finally, we compare $R_{\mathrm{ID}}^{\mathrm{CS}}$ for the special setting with the optimal R_{ID}^{C*} for $\rho = 0$ and $\rho = 0.7$ in Fig. (4). We also add R_{ID}^{*} of i.i.d. Gaussian sources as a baseline for the comparison. When $\rho = 0$, the input sources are i.i.d. Gaussian. We can observe that R_{ID}^{C*} is the same as R_{ID}^{*}. For the component-based approach, the optimal rate R_{ID}^{C*} is close to $R_{\mathrm{ID}}^{\mathrm{CS}}$ in the low rate region, and $R_{\mathrm{ID}}^{\mathrm{CS}}$ starts to become suboptimal after the second component is activated for R_{ID}^{C*}. The reason of the suboptimality of $R_{\mathrm{ID}}^{\mathrm{CS}}$ is that the special setting $D_{\mathrm{ID}}^{(1)} = 2D_{\mathrm{ID}}$ is enforced and cannot be optimized according to (10).

IV. Conclusions

In this work, we derive the identification rate of a component-based approach for block-correlated Gaussian sources. We characterize the identification rate for a special setting of the component-based approach. From the example,

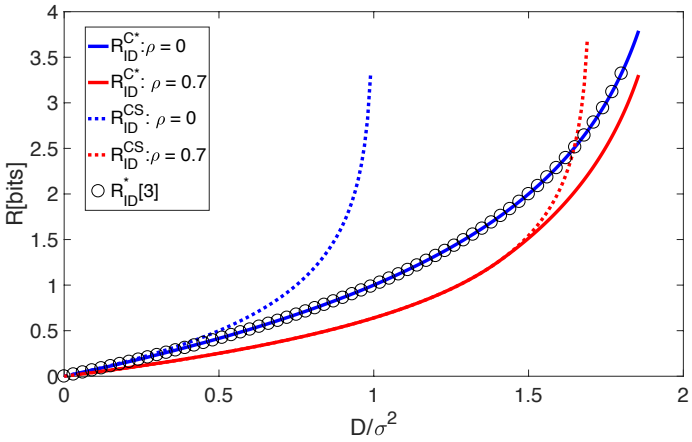

Fig. 4: Comparison of R_{ID}^{C*} and $R_{\mathrm{ID}}^{\mathrm{CS}}$ for $\rho = 0$ and $\rho = 0.7$. R_{ID}^{*} for i.i.d. Gaussian sources is given for reference.

we show that the identification rate for this special setting becomes suboptimal after the second component is activated. Furthermore, we show that block-correlated Gaussian sources are the "most difficult" to code under the special setting.

References

[1] A. Ingber and T. Weissman, "Compression for similarity identification: Fundamental limits," in *IEEE International Symposium on Information Theory*, Jun. 2014, pp. 1–5.

[2] R. Ahlswede, E. H. Yang, and Z. Zhang, "Identification via compressed data," *IEEE Trans. Inf. Theory*, vol. 43, no. 1, pp. 48–70, 1997.

[3] A. Ingber, T. Courtade, and T. Weissman, "Compression for quadratic similarity queries," *IEEE Trans. Inf. Theory*, vol. 61, no. 5, pp. 2729 –2747, May 2015.

[4] A. Ingber and T. Weissman, "The minimal compression rate for similarity identification," [Online]. Available: http://arxiv.org/abs/1312.2063.

[5] I. Ochoa, A. Ingber, and T.Weissman, "Compression schemes for similarity queries," in *Proc. of the IEEE Data Compression Conference*, Mar. 2014.

[6] F. Steiner, S. Dempfle, A. Ingber, and T. Weissman, "Compression for quadratic similarity queries: finite blocklength and practical schemes," *IEEE Trans. Inf. Theory*, vol. 62, no. 5, pp. 2737 –2747, May 2016.

[7] J. Hamkins and K. Zeger, "Asymptotically dense spherical codes. i. wrapped spherical codes," *IEEE Trans. Inf. Theory*, vol. 43, no. 6, pp. 1774–1785, 1997.

[8] H. Wu, Q. Wang, and M. Flierl, "Tree-structured vector quantization for similarity queries," in *2017 Data Compression Conference (DCC)*, April 2017.

[9] H. Wu and M. Flierl, "Transform-based compression for quadratic similarity queries," in *Conference on Signals, Systems, and Computers*, Oct. 2017, pp. 377–381.

[10] A. Ingber, T. Courtade, and T.Weissman, "Quadratic similarity queries on compressed data," in *2013 Data Compression Conference (DCC)*, 2013.