A Compression Distance for Bayesian Parametric Models

Lionel Gueguen*

Joint Research Centre, European Commission
Via E. Fermi, 2749, I-21027 Ispra (Va), Italy

The 8th Conference on Image Information Mining

*Lionel Gueguen is, since Sept. 2012, with DigitalGlobe Inc., Research & Development Department
Outline

1. Introduction
2. Proposed Compression Distance
   - Minimum Description Length
   - Normalized Compression Distance
   - Conditional Coding Length Approximation
3. Transform Invariance
   - Transform Invariance Definition
   - Toward An Invariant Compression Distance
4. Illustrative Case
   - Realizations of Gaussian Variables
   - Compression Distance Visualization
5. Conclusion
Context

- Compression raised interests in information mining since decreasing the coding length of a numerical object exploits its properties/characteristics/features;
- Data compression based image pattern representation was used for the first time in (Watanabe et al. 2002) for the classification of satellite images;
- It inspired later studies dealing with compression based information extraction from satellite patches (Cerra et al. 2010) or satellite image time series (Gueguen et al. 2008);
- These studies leded to the use of the Information Distance and its normalized version (Cilibrasi et al. 2005).
The Normalized Compression Distance (NCD) between two objects is obtained as the ratio of coding lengths.

\[ d(x, y) = \frac{K(xy) - \min(K(x), K(y))}{\max(K(x), K(y))} \]

- The coding lengths are obtained by real compressors such as the Lempel-Ziv coder;
- For example, image patches are transformed into strings before compression (Watanabe et al. 2002, Cerra et al. 2008) or a JPEG-Lossless like compressor was used in (Gueguen et al. 2008).
Motivation

- Data mining techniques making use of the Compression Distance are generally opposed to feature based information mining;
- Nevertheless, this point of view is arguable since the NCD relies on the compression technique which captures some properties of any object;
- Therefore, the NCD is dependent on the used compressor and the properties that it captures;
- We present the NCD applied to Bayesian parametric models which are intrinsically linked to compression and which have been extensively used in literature to characterize textures or spectral distributions;
Outline

1. Introduction
2. Proposed Compression Distance
   - Minimum Description Length
   - Normalized Compression Distance
   - Conditional Coding Length Approximation
3. Transform Invariance
   - Transform Invariance Definition
   - Toward An Invariant Compression Distance
4. Illustrative Case
   - Realizations of Gaussian Variables
   - Compression Distance Visualization
5. Conclusion
Outline

1. Introduction
2. Proposed Compression Distance
   - Minimum Description Length
   - Normalized Compression Distance
   - Conditional Coding Length Approximation
3. Transform Invariance
   - Transform Invariance Definition
   - Toward An Invariant Compression Distance
4. Illustrative Case
   - Realizations of Gaussian Variables
   - Compression Distance Visualization
5. Conclusion
Bayesian Parametric Models

- Models are used in information processing in order to capture the properties of interest of a signal;
- Models provide a two part representation of a signal: one embeds the signal properties and the other encodes the remaining uncertainty;

Notations of Generic Bayesian Models

Let \( \{M_1, \ldots M_m\} \) be a set of parametric models with their associated multidimensional parameter vectors \( \{\theta_1, \ldots \theta_m\} \). Let \( x^n = [x_1, \ldots x_n] \) be a signal of i.i.d. random variables. Given a couple of parameters and model \((\theta_i, M_i)\), the uncertainty which remains about the object is given by the conditional probability distribution \( p_{M_i}(x^n | \theta_i) \).
The Minimum Description Length criterion privileges the shortest description of an object (Rissanen 1995).

The optimal couple of properties and model \((\hat{\theta}_x, M_x)\) which minimizes the representation length of a signal \(x^n\) is given by:

\[
(\hat{\theta}_x, M_x) = \arg \min_{\theta_i, M_i} - \log(p_{M_i}(x^n | \theta_i)) + \frac{\text{dim}(\theta_i)}{2} \log(n),
\]

where \(\text{dim}(\theta_i)\) is the parameter vector dimension.

- \(\frac{\text{dim}(\theta_i)}{2} \log(n)\) is the code length of the parameters;
- \(- \log(p_{M_i}(x^n | \theta_i))\) is the code length of the remaining uncertainty.
The MDL criterion is originally a model selection technique; Additionally, it provides an approximation of the shortest coding length of a signal $x^n$ (Barron et al. 1998), which can be used for computing the Normalized Compression Distance.

The couple $(\hat{\theta}_x, M_x)$ represents the object compression based properties and the shortest coding length of $x^n$ is expressed by:

$$ l(x^n) = - \log(p_{M_x}(x^n | \theta_x)) + \frac{\dim(\theta_x)}{2} \log(n). \quad (2) $$
The Normalized Information Distance was first introduced in (Cilibrasi et al. 2005), which reports that:

Every admissible distance expressing similarity according to some feature, ..., is minorized by the Normalized Information Distance. Note that every feature of the data gives rise to a similarity, and, conversely, every similarity can be thought of as expressing some feature. We stress once more that different pairs of objects may have different dominating features. Yet every such dominant similarity is detected by the Normalized Information Distance.
The Normalized Information Distance is approximated by the Normalized Compression Distance (NCD) and it is expressed between two objects by:

\[
e(x, y) = \frac{l(x, y) - \min \{l(x), l(y)\}}{\max \{l(x), l(y)\}},
\]

(3)

\[
= \frac{\max \{l(x | y), l(y | x)\}}{\max \{l(x), l(y)\}},
\]

(4)

where \(l(x)\), \(l(x, y)\) and \(l(x | y)\) are the minimal code lengths necessary to represent \(x\), the couple \((x, y)\) and \(x\) knowing \(y\), respectively.
First Order Approximation

For two signals $x^n$ and $y^m$, their shortest coding length are given by the MDL criterion. It remains to approximate the conditional coding lengths $l(x^n|y^m)$ and $l(y^m|x^n)$. A straightforward approximation (Gueguen et al. 2007) for Bayesian parametric models can be expressed as follows:

The "Model Based Similarity Metric"

When describing one object $x^n$ knowing the two-part representation of another object $y^m$, only the properties are of interest since the uncertainty part $y^m | \theta_y$ does not help in representing $x^n$. Therefore, the conditional coding length $l(x^n | y^m)$ is upper bounded and approximated by $l(x^n | y^m) \approx -\log(p_{M_y}(x^n | \theta_y))$. 
Second Order Approximation

- When representing an object $x^n$ from the knowledge of $y^m$, only its properties $(\theta_y, M_y)$ should be exploited (First Order);
- Additionally, the most suited properties of $y^m$ for representing $x^n$ should be identified (Second Order).

Some Notations

- Let $k = \text{dim}(\theta_y)$ be the dimension of the model $M_y$;
- Let $r$ be a subset of the index set $\{1, \ldots, k\}$;
- Let $\theta^r_y$ be the parameter values whose index are in the subset $r$ and $\theta^{\bar{r}}_y$ the remaining values, where $\bar{r} = \{1, \ldots, k\} - r$. 
Given the properties \((\theta_y, M_y)\) of \(y^m\), one should identify the properties (the set of index) which fit best the other signal \(x^n\). One subpart of the parameter vector is used directly and the remaining part is inferred from the signal itself.

### Conditional Coding Length Approximation

\[
I(x^n | \theta_y, M_y) = \min_{r \subseteq \{1, \ldots, k\}, \theta_r} - \log(p_{M_y}(x^n | \theta^r, \theta^r_y)) + \frac{\dim(\theta^r)}{2} \log(n),
\]

(5)

\[
I(x^n | y^m) = \min \{ I(x^n | \theta_y, M_y), I(x^n) \}.
\]

(6)
Outline

1. Introduction
2. Proposed Compression Distance
   - Minimum Description Length
   - Normalized Compression Distance
   - Conditional Coding Length Approximation
3. Transform Invariance
   - Transform Invariance Definition
   - Toward An Invariant Compression Distance
4. Illustrative Case
   - Realizations of Gaussian Variables
   - Compression Distance Visualization
5. Conclusion
A feature function $f$, which extracts the properties of an object $x$, is said to be invariant to a transform $T$, if $\forall x$, $f(x) = f(T(x))$;

Thus a distance defined on the features $f(x)$ is invariant to the transform $T$;

In many cases, a given transform $T$ acts solely on a subset $t \subseteq \{1, \ldots, k\}$ of the $k$ parameters of a given model $M$;

**Some Notations**

- Let a transform $T$ act only on the features $\theta^t_x$ of $x^n$;
- Let $\tilde{\theta}^t_x$, $\tilde{t} = \{1, \ldots, k\} - t$ be the invariant parameters;
- Let $\tilde{t}^T_M$ be the set of parameter indices which are invariant to the transform $T$ given a model $M$. 
Toward An Invariant Compression Distance

- The impact of a transform can be decreased on the calculus of the Conditional Coding Length;
- Only the invariant parameter values should be considered as helping in reducing the coding length of $x^n$;
- The remaining parameters which vary with $T$ should be inferred from $x^n$;

**Transform aware conditional coding length**

$$l_T(x^n | \theta_y, M_y) = \min_{t_{My} \subseteq r, \theta^r} - \log(p_{My}(x^n | \theta^r, \theta^r_y)) + \frac{\dim(\theta^r)}{2} \log(n). \quad (7)$$
The shortest conditional coding length is given by
\[ l_T(x^n | y^n) = \min \{ l_T(x^n | \theta_y, M_y), l(x^n) \}; \]

The NCD derived from this approximation is denoted by
\[ d_T(x^n, y^m), \text{ and reduces the impact of } T; \]

By construction, \[ l(x^n | y^m) \leq l_T(x^n | y^m), \text{ such that } d(x^n, y^m) \leq d_T(x^n, y^m). \]
Let $x^n$ and $y^m$ represent $n$ and $m$ i.i.d. realizations of the Gaussian random variables $\mathcal{N}(\mu_x, \sigma_x)$ and $\mathcal{N}(\mu_y, \sigma_y)$.

**Shortest Coding Length**

\[
l_G(x^n) = n\left(\frac{1}{2} + \log(\sqrt{2\pi\sigma_x})\right) + \log(n),
\]

where $\frac{1}{2} + \log(\sqrt{2\pi\sigma_x})$ represents the average coding length of each element of $x^n$, while $\log(n)$ is the coding length necessary to represent $\{\mu_x, \sigma_x\}$ given $n$ samples.
Realizations of Gaussian Variables

Conditional Shortest Coding Length

\[ I_G(x^n \mid y^m) = \min \{ \]
\[ \frac{n}{2} \left( 1 + \log(2\pi \sigma_x^2) \right) + \log(n), \quad \text{no parameter} \]
\[ \frac{n}{2} \left( \frac{\sigma_x^2}{\sigma_y^2} + \log(2\pi \sigma_y^2) \right) + \frac{1}{2} \log(n), \quad \sigma_y \text{only considered} \]
\[ \frac{n}{2} \left( 1 + \log(2\pi (\sigma_x^2 + (\mu_x - \mu_y)^2)) \right) + \frac{1}{2} \log(n), \quad \mu_y \text{only considered} \]
\[ \frac{n}{2} \left( \frac{\sigma_x^2}{\sigma_y^2} + \frac{(\mu_x - \mu_y)^2}{\sigma_y^2} + \log(2\pi \sigma_y^2) \right) \quad \text{all considered} \]

(9)
$d(x^n, y^m)$ of Gaussian Variables

- $x^n$ is a realization vector of $\mathcal{N}_x = \mathcal{N}(0, 1)$ with $n = 50$;
- $d_G(x^n, y^m)$ is represented for $y^m$ being realizations from various Gaussian models $\mathcal{N}_y = \mathcal{N}(\mu_y, \sigma_y)$. 
Comparison to Kullback-Leibler Divergence

\[ d_G(x^n, y^m) \]

\[ d_{KL}(\mathcal{N}_x || \mathcal{N}_y) \]

\[ \hat{d}_G(x^n, y^m) \]

\[ d_{KL}(\mathcal{N}_y || \mathcal{N}_x) \]
Toward Transform Invariance

Invariance on Transform of $\mu$

Invariance on Transform of $\sigma$
Conclusion and Perspectives

Conclusion:
- Presented a compression distance for Bayesian models;
- Approximation of shortest coding length through the MDL principle;
- Ideas to make the distance invariant to transforms;
- Illustrative case for Gaussian variables;

Perspectives:
- Impact of the number of samples;
- Use the distance with Markov Random Field model of textures;
- Compare to Lempel-Ziv based compressor approximation;
- Derive a kernel from the distance for kernel based methods;
References

T. Watanabe, K. Sugawara, and H. Sugihara,
“A new pattern representation scheme using data compression,”

R. Cilibrasi and P. M. B. Vitanyi,
“Clustering by compression,”

A. Barron, J. Rissanen, and Bin Yu,
“The minimum description length principle in coding and modeling,”

D. Cerra and M. Datcu,
“Image classification and indexing using data compression based techniques,”

D. Cerra, A. Mallet, L. Gueguen, and M. Datcu,
“Algorithmic information theory-based analysis of earth observation images: An assessment,”

L. Gueguen and M. Datcu,
“A similarity metric for retrieval of compressed objects: Application for mining satellite image time series,”
*IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 4, pp. 562–575, 2008.

L. Gueguen and M. Datcu,
“The model based similarity metric,”