# Information loss from dimensionality reduction in Gaussian distributed spectral data

Alexej Schelle
*IU Internationale Hochschule and*
*hema.to GmbH, Ainmiller Str. 22, 80801 Munich*

Hannes Lueling
*hema.to GmbH, Ainmiller Str. 22, 80801 Munich*
(Dated: December 7, 2022)

Understanding the loss of information in spectral analytics is a crucial first step towards finding root causes for failures and uncertainties using spectral data in artificial intelligence models built from modern complex data science applications. Here, we show from a very simple entropy model analysis with quantum statistics of spectral data, that the relative loss of information from dimensionality reduction due to projection of an initial multi-dimensional state onto two-dimensional diagrams is less than one percent in the parameter range of small data sets with sample sizes on the order of few hundreds data samples. From our analysis, we also conclude that the structure, density and expectation value of the entropy probability distribution increases with the sample number and sample size using artificial data models derived from random sampling Monte Carlo simulation methods.

**Purpose:** Whitepaper on Github.

## I. INTRODUCTION

Artificial intelligence models are considered as a very powerful tool for quantitative predictions and environmental modelling in a very large range of physical and economic applications. Typically based on data training with high-dimensional representations of state vectors which describe the system under study and typically cannot be directly related to the actual physical processes or the equations describing the latter, AI models are most favourably implemented in modern IT technologies such as Python Machine Leaning. Interpreting and applying predictions of artificial intelligence models requires a solid quantitative understanding of the underlying accuracy and numerical limits of the applied algorithms and mathematical assumptions and methods. Despite that artificial intelligence models have been proven to be a successful tool for predictive analytics in a very large range of applications, such as medicine, economics or weather forecasting [1], a current problem of artificial intelligence models based on technologies such as Python Machine Learning and Big Data is that most artificial intelligence models are not generalisable in the sense that once trained to good accuracy on the basis of a specific data set, the AI model cannot be applied to similar datasets with slightly different parameters without firstly training the model to the new dataset.

In this context, a specific unresolved question is whether predictions from the training of AI models with multi-dimensional data can be compared to the manual analysis of the same data with reduced dimensionality. In this context, it is especially important to relate the accuracy of model predictions in any trial of the model realisation, especially when applied to disease diagnostics, which is a realistic example for the situation, where numerical model predictions are made from high-dimensional data, whereas the manual analysis is performed on the basis of two- or three-dimensional datasets. As a simple example, one may manually compare predictions of an artificial intelligence software such as hema.to describing probabilities for leukemia diagnosis to the standard diagnosis made from combinations of two-dimensional representations of cytograms [2].

Generally, there exists a large range of relevant software programs for generating AI model predictions based on artificial intelligence models built from spectral data, which intrinsically make use of a dimensionality reduction in the application and analysis of the model results. For multi-dimensional datasets, where standard accuracy and sensitivity measures defined from correlations of true-positive and true-negative outcomes do only measure the accuracy of pairwise correlations between two predicted classes, a more accurate measure is needed in order to proof that the standard accuracy and sensitivity measure (defined only pairwise on two classes) do correctly quantify the correlations between the multi-dimensional input vector which the model was trained on and the manual outcome analytics of an observer which compares his observations with the model predictions. In our case, the convergence of a specific mathematical quantity, the entropy calculated from a multi-dimensional probability measure for the system to satisfy certain model properties is used as a generalised measure for the accuracy of the AI model with parameters defined from an artificial multi-dimensional dataset with a given initial shape of the probability distribution.

Understanding the predictive outcome of an artificial intelligence model based on spectral data is often performed by analysing the event density and intensity patterns and structures manually, in order to relate them to specific classifications such as different types of disease classes [3]. In this type of analysis, it is crucial to understand the accuracy both of the predictive model as well

as the standard analysis, in order to provide consistent predictions also in the case, where forecasts of artificial intelligence models and standard human analysis tend to contradict. In the sequel of the present analysis, it is therefore shown within a very simple entropy model that the deviations and losses of information present to artificial intelligence models based on spectral data is less than one percent from the loss of information obtained by projections of the initial multi-dimensional state vector onto pairwise, two-dimensional representations of the multi-dimensional dataset.

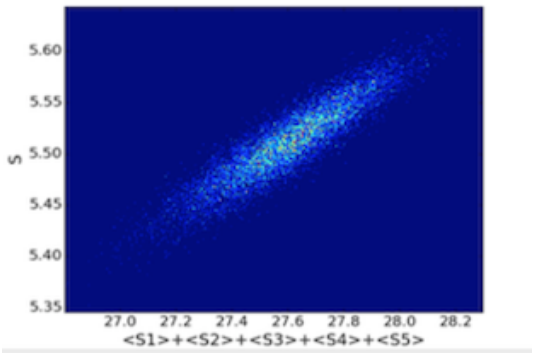


FIG. 1. (color online) Figure shows $10^4$ realisations of the total Shannon entropy versus the sum of conditional entropies in an artificial setup of $N = 400$ particles containing 5 frequency components with corresponding photon occupation numbers $N_k$. The mean value of the distribution is around 27.6 for the sum of conditional entropies, and 5.5 for the total entropy, and thus the relative information loss per frequency component measured by the Shannon entropy as defined in Eq. (1) is about one percent.

## II. THEORY

In the present work, modeling the entropy of N-dimensional spectral data within Gaussian quantum statistics is assumed to highlight the distribution of information in the context of photon emitting particles in a composite quantum system. As a starting point for our analysis of information loss from dimensionality reduction in such spectral representations of the multi-dimensional matrix representation in an spectral data file for data analytics, as a measure for information, we use the Shannon entropy defined as

$$S = -\sum_i p_i \log p_i \ , \tag{1}$$

where $p_i = p_i(E)$ is the probability

$$p_i(E) = \prod_{k=1}^{m} p_k^{(i)}(E) = \mathcal{Z}^{-1}(\sigma_k^2) \prod_{k=1}^{m} \mathrm{e}^{-\beta E - \frac{(E - < E_k^{(i)} >)^2}{2\sigma_k^2}} \tag{2}$$

for the system to occupy a quantum state with $N_k$ events of the frequency component $k$ and normalization $\mathcal{Z}(\sigma_k^2)$. In Eq. (2),

$$< E_k^{(i)} > = < N_k^{(i)} > \hbar\omega_k \ , \tag{3}$$

is the mean energy of a quantum state corresponding to the emission of $< N_k >$ photons from component $i$ with frequency $\omega_k$ on average. In the limit of large temperatures and small frequencies, Eq. (3) becomes a Gaussian sampling function for the number of events $N_k$ around their corresponding mean values $< N_k >$, weighted with the different frequency components $\omega_k$, which can in principle be either derived directly from spectral data analysis or from declaration, providing realistic quantification of the energy distribution, as defined in Eq. (3).

In our present approach, for a first estimate of information loss from dimensionality reduction, we reduce our analysis to an artificial data model by sampling the distribution of entropy as defined in Eq. (1) assuming artificial component frequencies in the optical range with non-trivial (average) occupations $< N_k >$ of the different frequency components. Please note that, from Eq. (2), it is straight forward to derive a mathematical expression for the intensity distribution of photon emission for the different components and to further implement a Monte-Carlo sampling method for the quantum energy and intensity distribution to very high accuracy. The total particle number is kept unchanged and constant during mathematical and numerical modelling [5].

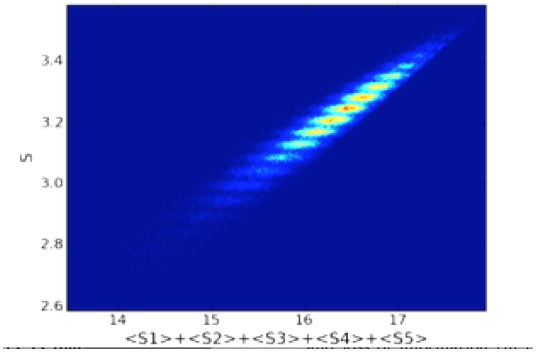


FIG. 2. (color online) Figure shows $10^5$ realisations of the total Shannon entropy versus sum of conditional entropies in an artificial setup of $N = 40$ particles containing 5 components with corresponding photon occupation numbers $N_k$. The structure of the entropy distribution can be clearly better recognised as the number of sampling steps increases.

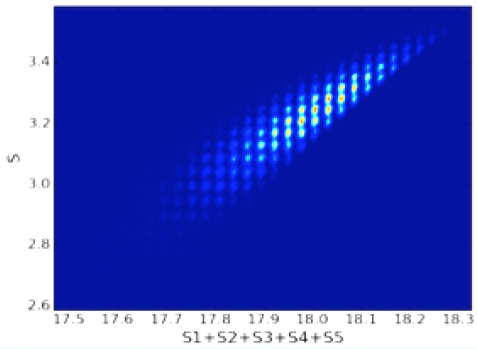


FIG. 3. (color online) Figure shows $10^5$ realisations of the total Shannon entropy versus sum of uncorrelated entropies in an artificial setup of $N = 40$ particles containing 5 components with corresponding photon occupation numbers $N_k$. It is observed that the relative loss of information is about five to ten times larger for (the sum of) uncorrelated entropies as compared to (the sum of) correlated, i.e. conditional entropies.

In order to calculate the intensity distribution of the quantum distribution, i.e. photons which are emitted by the particles, we may hence define the average intensity of the light emitted from particles

$$< I_k^{(i)} >= \frac{< N_k^{(i)} > \hbar\omega_k}{A_i \tau_i} \ , \tag{4}$$

where $A_i$ is the effective surface of the $i^{\text{th}}$ particle (defined by the effective Rayleigh surface for the different quantum particles) and $\tau_i$ the average measurement time. The expression in Eq. (4) leads to a mathematical expression of the corresponding intensity probability distribution, as a function of component frequency and photon number occupation.

## III. RESULTS AND ANALYTICS

From our analysis of numerical results, we learn that relative loss of information is less than one percent starting from samples with particle numbers larger than few hundred particles at typical optical component frequencies, neglecting the temperature dependence for a first approximation. As shown in Fig. (1), for a total number of $N = 400$ particles, the relative loss of information as captured by the entropy measure as defined in Eq. (1) is around one percent. The relative loss decreases for larger sample sizes and entropy density. We observe a spreading of the entropy distribution and a better connection between different local areas of the entropy distribution for a larger initial state range estimation as defined by the parameter $\delta$, i.e. the ratio of the local number uncertainty in units of the local mean photon number (expectation value) - compare Fig. (2). Switching this parameter from 0.5 to 1.0 corresponds to assuming non-random failure modes, e. g. in the measurement process. The structure of the entropy sampling distribution is better pronounced as the number of samples increases. For increasing number of particles, both the absolute mean value and the density of the entropy distribution increases. Numerical results are obtained from sampling Eq. (1) with a Monte Carlo sampling method, with different ranges of occupation numbers and component frequencies corresponding to the optical frequency range between 400 and 800 nm. Total number of particles was assumed to be in the range of 10 to 1000 particles. The number of frequency components was assumed to be $N = 5$.

Finally, from Eq. (3), assuming Gaussian energy distribution of the photon energy emitted by the different particle component types, it is possible to show that the relative loss of information tends to zero when the total number of particles tends to infinity. However, even this result seems to indicate that there are almost no information losses from dimensionality reduction in spectral analytics in the limit of large particle sizes, one should keep in mind that for single cases, there may be still non-negligible deviations from projection onto two-dimensional subsets.

More sophisticated models and measures for information loss may in principle lead to slightly different quantitative scaling of the results.

## IV. PROPOSAL AND OUTLOOK

From the presented analysis with spectral data obtained from Markov sampling, one may conclude and propose that information loss from dimensionality reduction in the context of quantum particles emitting multi-component frequency photons is proposed to sample the spectral data with a Markov sampling algorithm. Within our analysis, the levels of deviation induced by dimensionality reduction indicate that misclassifications predicted by artificial intelligence models are not likely to be induced by projecting the initial state onto reduced subspaces of the initial data space in the limit of Gaussian correlated spectral data.

Financial support from hema.to GmbH for this analysis of information loss from dimensionality reduction in artificial entropy models for spectral analytics within the project grant hema.to analytics (0001) is acknowledged.

[1] Blood (2019) 134 (Supplement-1): 886, Wolfgang Kern, MD, Franz Elsner, PhD, Max Zhao, Nanditha Mallesh, Richard Schabath, PhD, Claudia Haferlach, MD, Peter Krawitz, MD, Hannes Lueling, PhD, Torsten Haferlach, MD.
[2] http://hema.to, F. Kunzweiler, H. Lueling, K. Miermans
[3] Blood (2021) 138 (20): 1917–1927, Christian Matek, Sebastian Krappe, Christian Münzenmayer, Torsten Haferlach, Carsten Marr.
[4] Cytometry: Part A, Max Zhao, Nanditha Mallesh, Richard Schabath, PhD, Alexander Hoellein, MD, Claudia Haferlach, MD, Torsten Haferlach, MD, Franz Elsner, PhD, Hannes Lueling, PhD, Peter Krawitz, MD, Wolfgang Kern, MD.
[5] Fluctuations and Noise Letters, Vol. 16, No. 1 (March) 2017, A. Schelle.