SPECTRAL MEASUREMENT AND CLASSIFICATION IN THE ERA OF BIG DATA

Webler, F.S., Andersen, M
Laboratory of Integrated Performance In Design (LIPID), School of Architecture, Civil and Environmental Engineering (ENAC), École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, SWITZERLAND
forrest.webler@epfl.ch

Abstract

The measurement and classification of light is essential across many scientific disciplines. Devices used to measure light range from the highly precise scanning spectroradiometers to the more practical compact multichannel filter-array type imaging sensors and the ubiquitous RGB pixel. While there have been numerous successful efforts to reconstruct spectrum from RGB, RGB-to-spectrum reconstruction has historically been limited to natural scenes and other edge cases under strict constraints. However, information theory and recent advances in deep learning have shed new light on the vast amount of redundancy contained within data collected in the natural world, including light. In this paper, we will investigate how analytic methods can help map high dimensional spectra data to a low-dimensional feature space with minimal inductive bias. Through a better understanding of the intrinsic dimension of the data, we can use the features expressed in this representation to exploit regularities and make tasks like data compression, measurement and classification more efficient. The aim of this analysis is to help inform how and when low-dimensional representation of spectra is useful in practice for designing compact sensors as well as for lossy data compression and robust classification.

Keywords: Spectrometry, Classification, Big data, Sparse optimization, Metrology

1 Introduction

Light permeates our reality as electromagnetic radiation and is responsible for facilitating a multitude of interactions essential for life. While light exists as a continuous signal of emission frequencies, we often only describe light in terms of a lower dimensional representation. Consequently, the term *spectral resolution* can cause some confusion depending on the application and context when light is measured for scientific purposes. As most standards and norms relate to the visual appearance of light as observed by humans, a spectrum is reduced to three relative contributions of red, green, and blue (RGB). RGB was first formally introduced in the Young-Helmholtz trichromatic colour theory which elucidated the mythical properties of the eye by studying the additive properties of R, G, and B matching functions (Young, T., 1802). In 1861 history was made when James Clerk Maxwell produced the first colour photograph by experimenting with similar trichromatic filters (Maxwell, J.C., 1860; Longair, M.S., 2008). Nineteenth century physics, while primitive compared to modern standards, did uncover a fundamental property of (visible) light: despite the continuous *appearance* of light, the continuous spectral signals could be encoded in a low-rank embedding or *colour space.* In 1931 the concept of colour space was formalized by the International Commission on Illumination (Smith and Guild, 1931; CIE, 1932).

While trichromatic colour theory advanced our understanding of light and how it could be represented in ways that were easy to document and communicate, there was still unresolved work to be done regarding the subjectivity of colour. Standards and norms like those imposed at the 1931 CIE conference were adopted to ensure consistency across scientific fields and not to ensure absolute descriptivity of spectral information as observed by the existence of metamers in both the CIE RGB and XYZ colour spaces. Metamerism occurs when two spectral distributions map to the same coordinates in a pre-defined colour space (Luo, M.R., 2016). While this may be a moot point in the context of colour perception, it does mean that trichromatic matching functions developed over the last two centuries are ineffective in retaining spectral information. In other words, the embedding does not form a linearly independent spanning set (i.e., *basis*). In the following sections we will investigate how new methods in applied math can help contextualize our understanding of spectral representation and classification within

contemporary information theory. Using a data-driven approach we highlight exciting methods that are made possible by the existence of big data. Finally, we conclude with some outlooks into how the era of big data is redefining spectral measurement, classification and metrology in general.

2 An information theoretic approach to measurement, compression, and classification

Access to vast amounts of data challenges existing methods of measurement and classification so severely that some have claimed that the era of big data should be historically differentiated as the *fourth paradigm* of scientific discovery (Hey, 2009; Hey and Trefethen, 2020; Brunton et al., 2021). The more data we have to study; the more we know about a given class of observations. In the case of spectra, if we only had a very small number of blue-emitting illuminant spectra we might not know that it was possible to emit red-shifted photons. Over time as we observe more and more spectra, patterns emerge and these patterns (or *features*) enable us to characterise different natural phenomena. Towards this end, the challenge of big data is to find the methods that enable the discovery of the relevant features from a large set of observations. In applied machine learning, this is known as *feature mining.* Methods of feature mining range from relatively straightforward proper orthogonal decomposing (POD) to more complex autodidactic architectures (Brunton and Kutz, 2019). All of these methods are based on the principal of factor sparsity (i.e., Pareto principal) that the majority of information can be described by a minority of its features. In the context of visible light, this means that given access to many spectral observations, we only need to find those which span the *feature space*.

2.1 Data-driven encoding

While Maxwell identified filters based on available solutions like ferrous thiocyanide (green) out of convenience and availability (Dougal et al., 2006), there was no evidence that his filters were optimised for spectral reconstruction. The advent of big data means that we can now amass observations into a matrix and solve for the best low-rank approximation via some decomposition operation. In this section we investigate a specific method of decomposition known as *symmetric non-negative factorization* (SymNMF) developed by Kuang et al., (Kuang et al., 2012) coupled with QR-factorization for sparse sensor placement (Manohar *et al.,* 2019) and show how combing these methods leads to an interpretable data-driven approach to signal encoding.

2.1.1 Symmetric non-negative matrix factorization

SymNMF works by finding a low-rank approximation **H** of the affinity matrix **A** of a data matrix **Y**. The affinity matrix is derived by computing the differences between observations in **Y** relative to an appropriate dissimilarity metric. For spectral distributions we use the spectral angle mapper (SAM) algorithm (Yuhas *et al.,* 1992) although other metrics based on alternative geometries could be used. The columns of the low-rank matrix **H** form a linearly independent basis set and can be used to identify clusters by assigning indices to the columns of **H** with the largest values. Mathematically the factorization is written as

$$\min_{H \geq 0} \|A - HH\|_F^2$$

where

A is the affinity matrix

H is the low-rank approximation.

The advantage of SymNMF over classical non-negative matrix factorization is that it more naturally captures the structure of the original data matrix even when the data are non-uniformly distributed (as is often the case in large datasets). The rank of **H** is specified a priori but can also be determined by computing the Bayesian information criterion (BIC) of the objective function. In Figure 1 we show what the basis modes look like for $r = 7$ together with the objective function and a density plot illustrating the distribution of the spectra used to derive **A**.

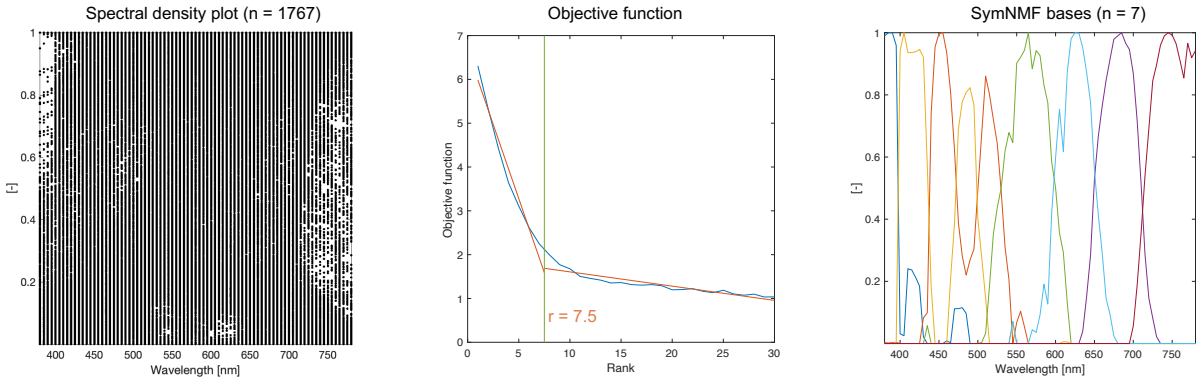

Figure 1 – (Left) Density plot showing all spectra used in the analysis with noticeable deficiencies in coverage for high-intensity blue, low-intensity green, and mid-intensity red frequencies. (Centre) Objective function showing the results for r = 1 to r = 30 with a Pareto-optimal solution for r = 7.5. (Right) SymNMF-derived basis modes (n = 7).

The information comprised in Figure 1 illustrates how regularities across spectral distributions can inform a representative set of basis modes that emulate sensor response functions like those used in multi-spectral imaging devices. The difference here is that these 'response functions' were *derived* by finding the $r = n$ rank approximation to our data matrix **Y**. Since **Y** is general (most of the density plot is covered), it is likely that if we could design a sensor with the SymNMF basis, we could recover the majority of observed spectra to a high degree of accuracy. Unlike basis derived via POD, the SymNMF basis shown in Figure 2 are feasible to construct because they are non-negative and nearly Gaussian. We can also see how SymNMF basis compare in the case of $r = 3$ compared against classic non-negative matrix factorization (NMF) and the CIE XYZ colour matching functions.

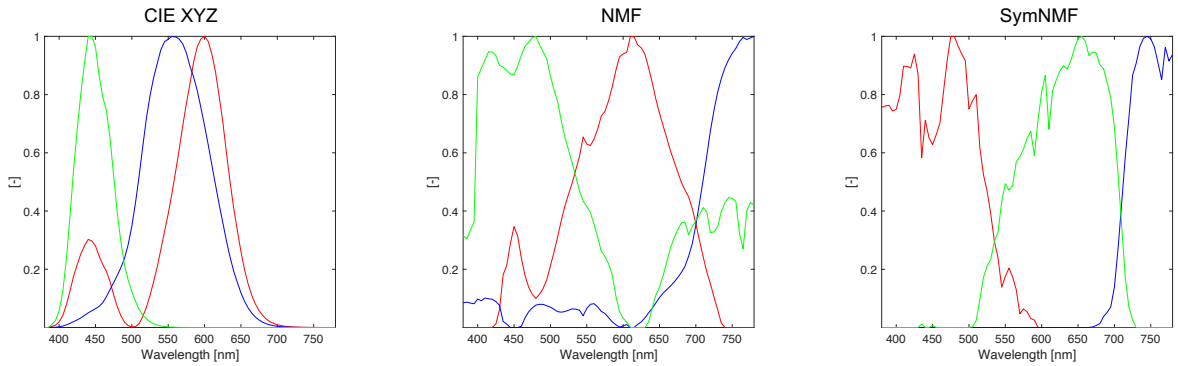

Figure 2 – Comparison of the CIE XYZ colour matching functions, the NMF-derived basis modes and the SymNMF-derived basis modes.

2.1.2 Optimized sensor placement with QR-factorization

In many applications including the design of imaging sensors, the basis modes themselves are infeasible to use as the exact sensor response functions. This is typically the case when the basis itself is decoupled from physical interpretation as would be the case if we did not constrain our bases to be non-negative. Regardless, the basis modes themselves are not meant to necessarily also be the best response functions since they merely guarantee that they form an orthogonal spanning set of the training data. Recently, work has focused on deriving optimal sensor locations from a learned basis in order to facilitate a method of sensing called *tailored sensing* (Manohar *et al.,* 2018). Tailored sensing is a data-driven method of learning the optimal sensor locations from prior data in the hopes that few strategically placed sensors can capture the same (if not more) information than many randomly placed sensors. The framework for this method was presented by Manohar et al. in 2018 and has demonstrated success in sensing

applications ranging from ocean currents to dynamic flow fields (Clark *et al.,* 2020)**.** Tailored sensing only requires that the number of sample points is greater than or equal to the intrinsic rank of the data, which can be determined by computing the first *r* basis modes such that $\mathbf{\Psi_r}$ is set equal to the SymNMF basis **H** which represents the majority of the information comprised in **Y.** Via the principle of factor sparsity, we can use $\mathbf{\Psi}_r$ to infer optimal sensor locations given that we will use $\mathbf{\Psi}_r$ as the basis for reconstruction. Manohar et al., demonstrated this could be done using a well-known decomposition method called QR-factorization (Businger and Golub, 1965). The set up for QR-factorization is straightforward where a matrix of point sensor locations P is determined by factoring the basis Ψ_r such that

$$\mathbf{\Psi}_r\mathbf{P} = \mathbf{QR}$$

where

$\mathbf{\Psi_r}$	is the tailored basis
P	is the pivot matrix of sensor locations
Q	is an orthogonal matrix
R	is an upper triangular matrix

The pivot matrix **P** can be computed directly from $\mathbf{\Psi_r}$ and comprises *r* columns corresponding to the optimal location along sampling axis. In the context of an imaging device, we can assume that each column of **P** corresponds to a photodiode with delta-function responsivity at a particular wavelength. For applications in compression, this means storing only the value of the signal at the locations specified by the pivots. While feasibility constraints exist for point sensors in real-world applications such as capturing the *spectral diet* of humans (Webler et al., 2019), the theory is clear: information from historical data (i.e., *domain knowledge*) can be used to optimise sensor placement. While exciting, these methods are predicated on access to a continuously updated master repository of real-world data. Whether or not the data we currently have is sufficient to infer the missing structures and properties in unseen spectral distributions is difficult to prove but based on the outcomes of dimension reduction algorithms like POD there may be a high probability that we have at least captured the majority of the total information needed to infer the remaining unknowns. To demonstrate how this can translate to the measurement and compressibility of spectral observations for low-complexity (e.g., daylight) and high-complexity (e.g., CIE F11) signals, we plot the recovery rate for each class of signals using uniform sampling and non-uniform data-driven sampling via SymNMF and QR-factorization (see Figure 3).

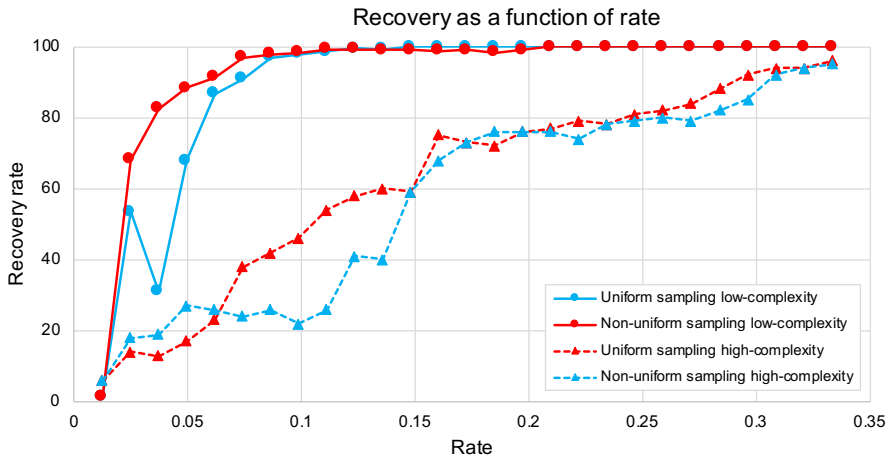

Figure 3 – Effect of tailored sensing with QR-derived bases on reconstruction of spectral information. The rate is defined as the number of sensors divided by the length of the signal.

It is clear from the results in Figure 3 that there are real advantages to picking sensor locations or pivot points (in the case of compression) based on a tailored basis. Due to the increasing availability of data, we should only benefit more from greater access to online datasets. What is interesting from both Figure 2 and 3 is that for low-complexity signals (i.e., those occurring from natural sources and their interaction with matter), the number of samples at which the signal is well defined is roughly between $n = 7$ and $n = 8$. When Isaac Newton did his first studies on light circa 1670, he claimed daylight was split into seven basic colours (Newton, I., 2014). Whether or not this is simply a coincidence, it is interesting to note that both the drop in the objective function and saturation for natural spectra occur when comprised of roughly seven basis modes.

3 Impact on existing standards and practices

Metrology is perhaps the most important field in science. Without standards and norms, scientists cannot compare observations share data or communicate their findings universally. Big data poses an existential threat to metrology because the more we know about our world, the less we have to discover and the more we have only to validate. From a mathematical perspective, this boils down to the fact that capturing truly novel information (i.e., *discovery*) requires imparting the fewest number of assumptions. On the other hand, if prior information is known *a priori*, we only need to capture the intrinsic information 'signature' of light to encode spectral information making it possible to design more compact spectral dosimeters. Whether or not we currently have collected enough data to derive inferences about all remaining unseen observations is functionally unprovable but we can say that they are *known unknowns* in that there are unlikely to be any true surprises. Towards this end, we propose a new framework for spectrometry that establishes definitions based on the intrinsic dimension according to our current knowledge of spectra. We summarize the main goals in two action points:

1. Curate an ever-expanding central repository of known measured spectra

2. Derive an official basis Ψ from the central repository using SymNMF factorization in order to update data-driven metrics for measurement and classification.

Adhering to and servicing these points would allow for a transparent and modern understanding of spectral distributions whilst providing researchers and manufactures with information theoretic bounds needed to make informative decisions in experimental and product design. We believe that such an approach would not only be more defensible from a theoretical perspective, but it would also be functionally useful in the same way that a library of congress is useful.

4 Conclusion

Big data offers exciting opportunities for scientific discovery but also challenges traditional metrological standards and practices. Re-examining these methods through the lens of information theory and the principle of factor sparsity can help contextualize metrology in the era of big data specifically in domains like spectral sensing where regularities and patterns dominate observational data. In other words, the idea is to investigate how methods like compressed sensing and sparse representation can help guide new standards and practices for measurement in a data-rich world.

We propose that curating an open repository of spectral data is essential and that applying an online sparse dictionary learning algorithm on this set will give life to an evolving taxonomy of spectra that can be used to understand relationships between spectral features. Not only would such a taxonomy serve as a useful point of reference it could also be used to derive an orthogonal basis Ψ of *eigenspectra* to inform sensor placement for general applications. With this paper, we hope, even if these suggestions are not implemented directly, that our introduction of these concepts will inspire a new way of thinking amongst metrologists. The seismic changes we see in an increasingly data-driven world can be difficult to keep up with and it is normal that standards and practices change slowly. We advocate that in the absence of perfect knowledge, methods that make the fewest assumptions are ultimately the most advantageous.

Understanding the intrinsic dimension of spectral distributions is important in both theory and practice. In theory it is important to separate redundant and essential information for classification and formal definition while in practice, there are numerous advantages for identifying redundant information. Many applications require reducing size, weight, power, and cost (SaWP-C) which can only be done by collecting the lowest informative representation of the signal (i.e., *intrinsic dimension*). In this era of big data, we feel it is important that we continue to uncover how machine learning exploration, when it comes to data representation, can continue to inform our understanding of data and that our systems and methods used to describe the data are continuously updated to reflect these advances in technology.

References

Young, T., 1802. II. The Bakerian Lecture. On the theory of light and colours. *Philosophical transactions of the Royal Society of London*, (92), pp.12-48.

Maxwell, J.C., 1860. On the theory of compound colours, and the relations of the colours of the spectrum. *Proceedings of the Royal Society of London*, (10), pp.404-409.

Longair, M.S., 2008. Maxwell and the science of colour. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *366*(1871), pp.1685-1696.

Luo, M.R. ed., 2016. *Encyclopedia of color science and technology*. Springer Reference.

Hey, A.J. ed., 2009. *The fourth paradigm: data-intensive scientific discovery* (Vol. 1). Redmond, WA: Microsoft research.

Hey, T. and Trefethen, A., 2020. The fourth paradigm 10 years on. *Informatik Spektrum*, *42*(6), pp.441-447.

Brunton, S.L., Budišić, M., Kaiser, E. and Kutz, J.N., 2021. Modern Koopman theory for dynamical systems. *arXiv preprint arXiv:2102.12086*.
Brunton, S.L. and Kutz, J.N., 2019. Data-driven science and engineering: Machine learning, dynamical systems, and control. Cambridge University Press.

Dougal, R.C., Greated, C.A. and Marson, A.E., 2006. Then and now: James Clerk Maxwell and colour. *Optics & Laser Technology*, *38*(4-6), pp.210-218.

Kuang, D., Ding, C. and Park, H., 2012, April. Symmetric nonnegative matrix factorization for graph clustering. In *Proceedings of the 2012 SIAM international conference on data mining* (pp. 106-117). Society for Industrial and Applied Mathematics.

Manohar, K., Brunton, B.W., Kutz, J.N. and Brunton, S.L., 2018. *Data-driven sparse sensor placement for reconstruction: Demonstrating the benefits of exploiting known patterns*. IEEE Control Systems Magazine, 38(3), pp.63-86.

Yuhas, R.H., Goetz, A.F. and Boardman, J.W., 1992, June. Discrimination among semi-arid landscape endmembers using the spectral angle mapper (SAM) algorithm. In *Proc. Summaries 3rd Annu. JPL Airborne Geosci. Workshop* (Vol. 1, pp. 147-149).

Clark, E., Brunton, S.L. and Kutz, J.N., 2020. *Multi-fidelity sensor selection: Greedy algorithms to place cheap and expensive sensors with cost constraints*. IEEE Sensors Journal, 21(1), pp.600-611

Businger, P. and Golub, G.H., 1965. *Linear least squares solutions by Householder transformations*. Numerische Mathematik, 7(3), pp.269-276.

Webler, F.S., Spitschan, M., Foster, R.G., Andersen, M. and Peirson, S.N., 2019. What is the 'spectral diet'of humans?. Current opinion in behavioral sciences, 30, pp.80-86.

Newton, I., 2014. A letter of Mr. Isaac Newton, Professor of the Mathematicks in the University of Cambridge; containing his new theory about light and colors: sent by the author to the publisher from Cambridge, Febr. 6. 1671/72; in order to be communicated to the R. Society. *Philosophical Transactions of the Royal Society of London*, *6*(80), pp.3075-3087.