


Unsupervised classification of non-linear dynamics in optical fiber propagation using intensity clustering

Anastasiia Sheveleva^a, Andrei V. Ermolaev^b, John M. Dudley^{b,c}, Christophe Finot^{a,*} 

^a Laboratoire Interdisciplinaire Carnot de Bourgogne, UMR 6303 CNRS-Université de Bourgogne, Dijon, France

^b Institut FEMTO-ST, UMR 6174 CNRS-Université de Franche-Comté, Besançon, France

^c Institut Universitaire de France (IUF), Paris, France

ARTICLE INFO

Communicated by: Dmitry Pelinovsky

Keywords:

Nonlinear fiber optics
nonlinear dynamics
pulse propagation
unsupervised classification

ABSTRACT

We demonstrate that centroid-based clustering of normalized intensity profiles is able to successfully isolate different classes of pulses associated with physically distinct regimes of nonlinear and dispersive propagation in optical fiber. Remarkable for its simplicity, this approach shows how analysis of only the temporal intensity profiles of propagating pulses, even at relatively limited sampling resolution, reveal sufficient similarities to allow physical classification of different classes of propagation behavior.

1. Introduction

Machine learning techniques have seen rapid and impressive development in the field of ultrafast photonics in the last 5 years [1–3]. A particular area that has especially benefited from this work is the study of nonlinear dynamics in optical waveguides, where different machine learning techniques have been applied to the study of output waveform properties [4–7], supercontinuum evolution maps [8,9], extreme event emergence [10] and the dynamics of four-wave mixing [11,12]. The methods used in this work have been varied, including simple feedforward neural network architectures [4,10–12], recurrent neural networks [8] convolutional networks [5] and more complex combinations of different deep learning algorithms [5,7]. Other work has used physics-informed methods to solve the underlying propagation equations [13], inverse-problem methods for prediction [4,6,12], sparse regression to discover dynamical models from data [14,15], and dominant-balance methods to automate detection of dominant physics during nonlinear evolution [16–18].

In this paper, we report a further application of machine learning methods in nonlinear fiber optics, using simple clustering techniques to analyze and classify intensity profiles resulting from a range of propagation scenarios. Specifically, we show that simple centroid-based clustering, such as the widely used K-means algorithm [19,20], can successfully reveal patterns during higher-order soliton and wave breaking evolution dynamics, and can also distinguish normal and

anomalous dispersion-regime dynamics in specific regimes of propagation. These results reveal that clustering provides a complementary tool to existing analytical and numerical methods to automatically highlight patterns and similarities between different regimes of evolution, potentially yielding new insights into the underlying physics. Moreover, although motivated by the study of dynamics in optical fiber propagation, our analysis is based on an ideal nonlinear Schrödinger equation model, so that our results can be readily extended to other fields such as atom optics, hydrodynamics and plasma physics.

2. Propagation model and methods

We consider propagation in optical fiber described by the dimensional nonlinear Schrödinger equation (NLSE) [21]:

$$i \frac{\partial \psi}{\partial z} - \frac{\beta_2}{2} \frac{\partial^2 \psi}{\partial t^2} + \gamma |\psi|^2 \psi = 0 \quad (1)$$

where $\psi(t, z)$ is the complex pulse envelope, γ and β_2 are respectively the nonlinear and dispersion coefficients, z and t are respectively the propagation distance and comoving time. We define nonlinear and dispersive lengths as $L_{NL} = 1/\gamma P_0$ and $L_D = T_0^2/|\beta_2|$ respectively, where P_0 and T_0 are the peak power and temporal width of a hyperbolic secant input pulse $\psi(t, 0) = \sqrt{P_0} \operatorname{sech}(t/T_0)$. Loss and noise are neglected. The NLSE in normalized form is then given by:

* Corresponding author.

E-mail address: christophe.finot@u-bourgogne.fr (C. Finot).

$$i \frac{\partial u}{\partial \xi} - \frac{\text{sgn}(\beta_2)}{2} \frac{\partial^2 u}{\partial \tau^2} + N^2 |u|^2 u = 0 \quad (2)$$

where $\xi = z/L_D$, $\tau = t/T_0$, and $N^2 = L_D/L_{NL}$ defines the soliton number N . The normalized input pulse is $u(\tau, \xi) = \text{sech}(\tau)$.

Our aim here is to use centroid-based clustering to differentiate different regimes of nonlinear and dispersive pulse propagation, based solely on analyzing temporal intensity profiles $I(\tau, \xi) = |u(\tau, \xi)|^2$. We stress that this is a severe constraint because we have no access to the temporal phase of the pulse or any spectral information. To generate large data sets of intensity profiles, we perform numerical simulations to solve the NLSE [21] for different initial conditions by randomly scanning N over the range 1–5 (continuously i.e. not only integer values), and at different propagation distances by extracting $I(\tau, \xi)$ at distances ξ over the range 0.0–2.5. We generated two ensembles of data consisting of 20,000 independent numerical simulations for each case of normal ($\beta_2 > 0$) and anomalous ($\beta_2 < 0$) dispersion regime propagation. Given the temporal symmetry of our initial conditions and the fact we are in an ideal NLSE model, the temporal intensity profiles are always symmetric, and so we can restrict ourselves to performing clustering for only positive-valued τ i.e. for $I(\tau > 0, \xi)$. We also introduce a normalized intensity profile $I_n(\tau, \xi) = I(\tau, \xi)/I_{\max}(\xi)$ with I_{\max} is the maximum intensity at a given ξ . Considering the intensity profile in this way allows us to characterize the structure of the intensity profiles in terms of their computed kurtosis as we describe below. Our NLSE simulations used 2^{13} temporal grid points over a temporal window of normalized τ -width 100. We first present some general results in Fig. 1 (a), showing three typical examples of normalized intensity profiles under three different propagation conditions: (green) $\xi = 0.8$, $N = 4.1$, $\text{sgn}(\beta_2) = -1$; (red) $\xi = 2$, $N = 4.5$, $\text{sgn}(\beta_2) = +1$; (blue) $\xi = 1.5$, $N = 2.9$, $\text{sgn}(\beta_2) = -1$. The profiles are plotted for positive times ($\tau > 0$) and illustrate the large qualitative differences between pulse profiles arising from propagation under different conditions, with shapes ranging from highly compressed pulses with subpulses associated with soliton evolution (green), to highly flattened and significantly broadened pulse envelopes associated with high power normal dispersion propagation dynamics (red) [21,22]. For improved clustering performance, we downsample the 4096 positive time points to 2^L points, with L ranging between 4 and 8, using a nonuniform sampling procedure in order to ensure that we capture both broad envelope and short compressed features in the intensity profiles. Specifically, we resampled our data on a logarithmically-spaced grid covering the normalized temporal interval [0 25] with 2^L samples. Examples of the resulting data obtained for $L = 8, 6$ and 4 are provided in panels (b) of Fig. 1 where we notice that even a reduced number of only 16 samples is still able to reproduce the general features of the pulse profile after propagation.

We used $L = 7$ except in the last section of this work where we discuss the influence of L . The two ensembles of 2×10^4 temporal intensity profiles after logarithmic resampling serve as unlabeled data of the clustering algorithm in order to have a dense and visually quasi-continuous cover of the (N^2, ξ) space. We have however checked that qualitatively similar clustering could be achieved with ensemble sizes down to ~ 200 .

We have also calculated, from the intensity profiles before resampling, the temporal moments of order two and four to characterize the intensity profile through its root mean square (rms) temporal width σ and its kurtosis κ [23] respectively. These are defined for a symmetric waveform as:

$$\sigma^2 = \frac{\int \tau^2 I(\tau) d\tau}{\int I(\tau) d\tau} \quad (3)$$

$$\kappa = \frac{\int \tau^4 I(\tau) d\tau / \int I(\tau) d\tau}{\sigma^4}, \quad (4)$$

where we note that the rms duration and the kurtosis of a hyperbolic

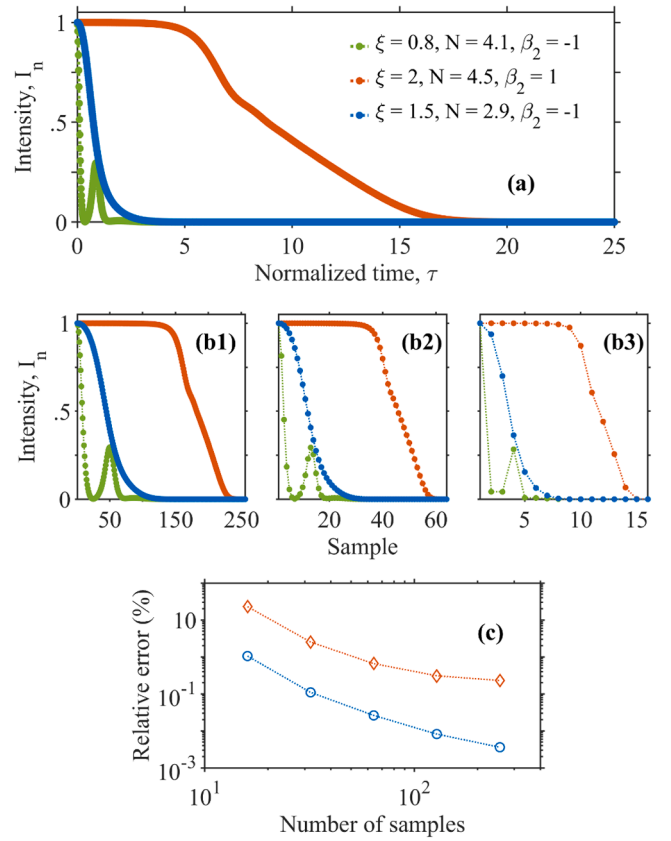


Fig. 1. (a) Examples of three normalized intensity profiles used as inputs for clustering. These profiles result from the propagation of a hyperbolic secant pulse under the following conditions: (green) $\xi = 0.8$, $N = 4.1$, $\text{sgn}(\beta_2) = -1$; (red) $\xi = 2$, $N = 4.5$, $\text{sgn}(\beta_2) = +1$; (blue) $\xi = 1.5$, $N = 2.9$, $\text{sgn}(\beta_2) = -1$. The profiles are plotted over a temporal window of normalized width 25 representing 2048 linearly spaced points. (b) The same temporal intensity profiles but after non-uniform resampling with 256, 64 and 16 samples ($L = 8, 6$ and 2 plotted on panels b1, b2 and b3, respectively). (c) Influence of the number of logarithmically spaced samples on the average relative error made on the numerical evaluation of the rms duration (blue circles) and kurtosis parameter (red diamonds).

secant pulse are $\sigma = 0.907$ and $\kappa = 4.2$ respectively. We evaluate in Fig. 1 (c) the impact of the number of resampled points on the average relative error in σ and κ , using as reference our initial intensity profiles which were densely sampled in τ . We see from the figure that for $L > 6$ (i.e. 64 temporal points), the moments are evaluated with a very low average error below 1% over the two data ensembles (normal and anomalous dispersion regime) of 4×10^4 profiles. However, when L decreases below 6, the error rapidly increases, which can be understood since the integrals in Eqs (3) and (4) are numerically approximated by the trapezoidal rule. To avoid any sampling-induced errors of this kind, in what follows, we evaluate σ and κ from the initial intensity profiles before resampling.

The clustering method we use here is the K-means (or Lloyd-Forgy) algorithm, which is one of the most widely adopted algorithms in data science and which has been extensively studied for many years [19,24]. Although there are certainly other clustering methods that could be considered, our motivation here was to consider how even the simplest approaches to clustering can be used to add physical insights based only on the analysis of pulse intensity profiles. In particular, we have not attempted here to include any comparison or benchmarking with other methods, but this could naturally form the basis of future work.

We also note here that, although K-means clustering has been previously implemented in the context of optical telecommunication to

mitigate non-linear distortions of phase-coherent transmissions [25,26], to the best of our knowledge it has never been exploited to get insights on the nonlinear dynamics. K-means is a centroid-based algorithm where the user must first define the desired number of clusters in which to partition the data. In our case, we choose between 2–5 clusters as we describe below. Centroids are then randomly created based on the number of clusters and the distance between data points, and each centroid is calculated so that each data point can be assigned to the nearest centroid. The metric to evaluate “distance” here is an important parameter, and different distance metrics can be defined in addition to the simple Euclidean distance. For this study, we used the cosine distance based on a cosine similarity calculation, as this is known to yield better clustering results in the presence of complex non-spherical cluster shapes [24]. The mean of the centroid is recalculated based on all the assigned data points, and this will then modify the position of the centroid. This process is iterated until it converges when each data is linked to a single cluster. In order to avoid local minima and obtain reproducible output, we run the algorithm multiple times (typically between 10–15) and keep the results leading to the lowest total sum of distances.

3. Clustering for anomalous dispersion regime dynamics

We begin by considering the clustering behavior for intensity profiles generated from soliton dynamics in the fiber anomalous dispersion regime. We consider 20,000 intensity profiles obtained by scanning N over the range 1–5 and ξ over the range 0.0–2.5, and these profiles are then used as input to the K-means algorithm. To illustrate how simple clustering can identify different dynamical characteristics, we first present results when partitioning into only 2 clusters and considering as input the non-normalized profiles $I(\tau)$. Fig. 2 (a1) displays these results by plotting the (N^2, ξ) pair associated with each profile, but assigning a different color depending on the cluster into which the profile is sorted.

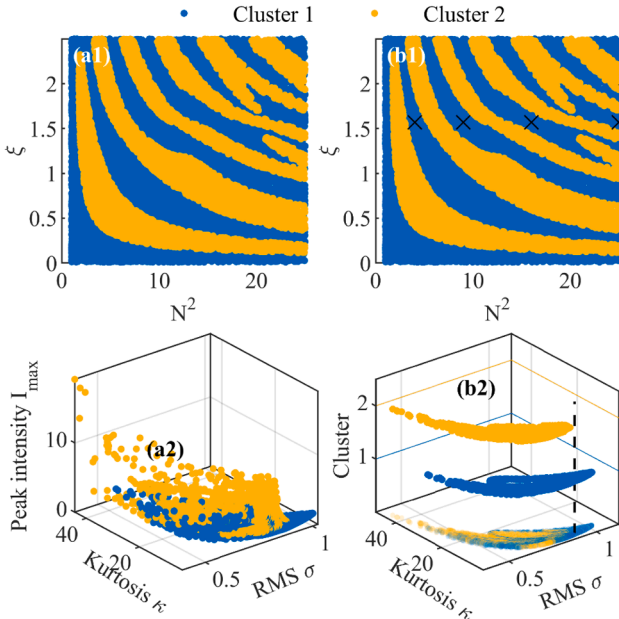


Fig. 2. Clustering into two sets of temporal intensity profiles generated for propagation in the anomalous dispersion regime. Panels (a) and (b) show results using non-normalized profiles $I(\tau)$ and normalized profiles $I_n(\tau)$ respectively. Panels a1 and b1 display clusters according to normalized propagation distance ξ and the soliton number N . The crosses indicate the parameters leading to ideal periodic recovery of the initial intensity profile. Panels a2 and b2 display clusters according to their rms duration σ and kurtosis κ as well as their peak power I_{\max} for the non-normalized dataset. The black dashed line in (b2) is a visual guideline for the properties of the input pulse.

This clearly shows a periodic band structure in the (N^2, ξ) plane, but to gain more insight, we plot in Fig. 2 (a2) a three-dimensional plot where the clustered profiles are plotted in terms of their root-mean-squared (rms) duration σ , their kurtosis κ and their peak intensity I_{\max} . It is clear from this plot that the main separation between the clusters arises from the difference in profile intensities, allowing us to conclude that the bands in Fig. 2 (a1) are identifying different stages of periodic temporal expansion and compression (yellow and blue bands respectively).

To consider how clustering can identify more general pulse shape characteristics in addition to peak intensity, Fig. 2 (b1) and Fig. 2 (b2) show results using normalized intensity profiles $I_n(\tau)$ as input. We see that Fig. 2 (b1) shows an identical band structure in the (N^2, ξ) plane, stressing that the knowledge of the peak-power does not provide major new insights. The clusters can be readily attributed to different stages of pulse evolution, with the yellow cluster linked to stages of noticeable compression [27,28] and significant deviation from the input profile, and the blue color to stages where there is only moderate change with respect to the input profile. This is particularly the case for short propagation distances where the dominance of the nonlinear regime (self-phase modulation) does not affect the temporal intensity profile [18]. Similarly, the blue regime dominates for areas with low nonlinearity and for values of $N \sim 1$ where the pulse preserves its shape or can even be stationary in the case of the fundamental soliton [21,28]. We also recognize the special conditions corresponding to integers values of N and propagation distances of $\xi = \pi/2$ (the black crosses in the (N^2, ξ) plane) associated with perfect soliton recurrence to its initial state [29]. Fig. 2 (b2) separates the previously obtained clusters in terms of computed temporal duration and kurtosis, with the projection in the rms/kurtosis (σ, κ) plane showing that the different regions strongly overlap. This highlights how an approach using the normalized profile is not equivalent for the anomalous dispersion regime of propagation to an approach considering only the temporal moments of the pulses, a point that we will further discuss in Figs. 3 and 4.

We now discuss how this clustering behavior changes when the number of clusters for partitioning is increased to 3. Results of this new clustering for the normalized profiles $I_n(\tau)$ are shown in Fig. 3 (a1). We again see a band structure in the (N^2, ξ) plane, with the blue regions again corresponding to intensity profiles that have experienced little variation with propagation. This cluster region is similar to that in the 2 cluster results in Fig. 2 (b1). However, when selecting 3 clusters, the single cluster in Fig. 2 (b1) (corresponding to pulses that underwent significant evolution in their intensity profiles) now splits into two subsets. And we see in Fig. 3 (a2) that in the rms/kurtosis (σ, κ) plane, these clusters are superimposed to a large degree so that clustering in terms of these features is insufficient to separate the different behaviors.

To better understand the physical characteristics that are being selected and associated with these different clusters, Fig. 3 (b1) and (b2) plots the longitudinal evolution of the intensity profile of pulses of soliton order $N = 4$ and $N = 5$, respectively, accompanied by the variations of the rms duration and kurtosis in panels (c1) and (c2) respectively. For solitons with integer values of N , using the inverse scattering transform [30], it is possible to interpret the longitudinal evolution as the result of the interaction of N distinct solitons with different powers that interfere with each other leading to increasingly complex temporal patterns [31]. Depending on the phase difference between these elementary soliton components, the output pulse exhibits temporal compression [27], splitting into two or more subpulses [28], or revival back towards its initial state [29]. We now see that the yellow bands in the (N^2, ξ) plane in Fig. 3 (b1) correspond to the zones undergoing maximum temporal compression, while the green bands correspond to the zone where the pulse has temporally split into several subpulses with an intensity at $\tau = 0$ approaching zero. This is confirmed by also showing on panel (a1) with the dashed curve, the empirical equation giving the distance ξ_c leading to the maximum compression as a function of the soliton number [32,33]:

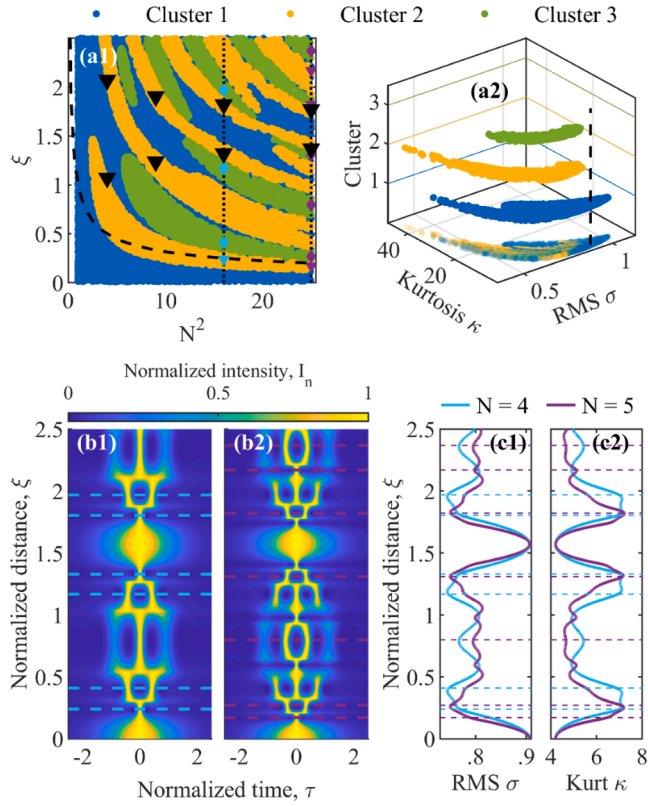


Fig. 3. . Clustering into three sets of the normalized temporal intensity profiles generated upon propagation in the anomalous dispersion regime. (a1) Clusters according to the normalized propagation distance ξ and soliton number N . The dashed black line marks the positions of the compression stages approximated by Eq. (5) and the black triangles highlight the parameters leading to recurrence of the compression stage for an integer value of N . Dashed cyan and purple lines label $N = 4$ and 5 , respectively. The cyan and purple triangles mark positions of the compressed pulses, while the cyan and purple diamonds – the split pulses' positions. (a2) Clusters according to their rms duration σ and their kurtosis κ . The black dashed line is a visual guideline for the properties of the input pulse. (b) Longitudinal evolution of the normalized intensity profile for $N = 4$ and 5 (panels b1 and b2 respectively). (c) Longitudinal evolution of c1: the corresponding rms duration σ and c2:kurtosis κ .

$$\xi_c \sim 1/N \quad (5)$$

Using the recurrence conditions and the longitudinal symmetry of the evolution achieved for integer values of N , we see that the corresponding points (black triangles) are well contained within the second cluster. The variations with propagation distance in the rms duration and kurtosis (panels (c)) highlight how these two parameters are not suitable to unambiguously characterize the output properties when the pulse is affected by splitting and sidelobes. Indeed, as an example, when considering the evolution of the 4th order soliton between distance 0.5 and 1, we observe only slight variations of σ and κ even though the output pulse goes through very different behavior such as pronounced compression or pulse splitting. Therefore, clustering in the anomalous regime based on σ and κ as shown in Fig. 4 (a) leads to a result that significantly differs from Fig. 3 (a). Even though the clusters plotted in the (σ, κ) plane are now well-separated, clustering is unable to catch the stage of pulse splitting.

Finally, we investigated how the addition of a fourth cluster to the K-means search algorithm impacts the results. With three clusters, we saw that the algorithm was able to differentiate between the stage where the output pulse is similar to the input, as well as the stages where there is compression or splitting into two pulses. One interesting stage of evolution that is not identified, however, is where the output pulse exhibits

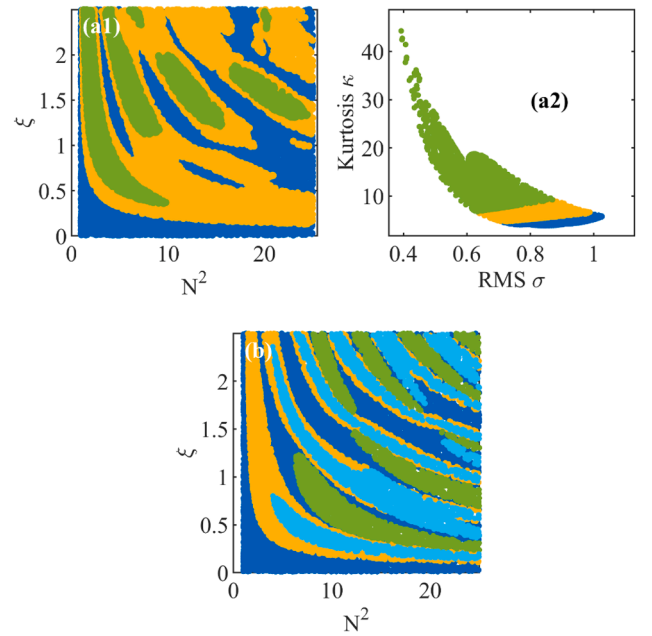


Fig. 4. . (a) Clustering into three sets based on σ and κ properties. (a1) Clusters according to the normalized propagation distance ξ and the soliton number N . (a2) Clusters according to their rms duration σ and their kurtosis κ . (b) Clustering into four sets of the normalized temporal intensity profiles. Clusters according to the normalized propagation distance ξ and the soliton number N .

strong central compression but with the emergence of pronounced sidelobes, leading to a three-pulse profile as in Fig. 3 (b) around $\xi \sim 0.4$ for $N = 5$. It is natural to consider whether such a stage is clearly identified when increasing the number of clusters to 4 and we show the results in panel (b) of Fig. 4. Specifically, when partitioning into 4 clusters, the main difference with the 3-cluster configuration is the splitting of the initial 'yellow cluster' attributed to compression into two new clusters now colored yellow and cyan. Although for some positions, the cyan cluster includes the 3 pulse profiles, this is not always the case, and it appears that the cyan cluster is mainly associated with the most compressed pulses, i.e. it has isolated the shortest profiles. So in this case we conclude that partitioning with 4 clusters does not add significant insight.

4. Clustering for normal dispersion regime dynamics

Pulse evolution in the normal dispersion regime differs drastically from the anomalous case. Instead of periodic evolution dynamics and temporal localization associated with soliton effects, the interaction between the Kerr nonlinearity and normal dispersion is marked by effects such as optical wave breaking, self-similar propagation, and temporal broadening [34,35]. We study here the clustering of intensity profiles associated with normal dispersion propagation using the same approach as above, with 2×10^4 intensity profiles scanning N over the range 1–5 and ξ over the range 0.0–2.5, except that now we take $\text{sgn}(\beta_2) > 0$, and the input parameter N is no longer interpreted as a soliton number.

Clustering into three clusters as shown in Fig. 5 (a1) leads to the isolation of three successive stages of dynamics associated with well-known characteristics of nonlinear and dispersive propagation [21]. The structure of the clusters is very different from the previous case, with each cluster here consisting of a single continuous block. The blue cluster, where the rms duration of the pulse does not vary significantly, corresponds to the initial propagation regime when the spectral broadening and chirp induced by the nonlinearity has not yet translated into significant temporal broadening. However, the shape of the temporal

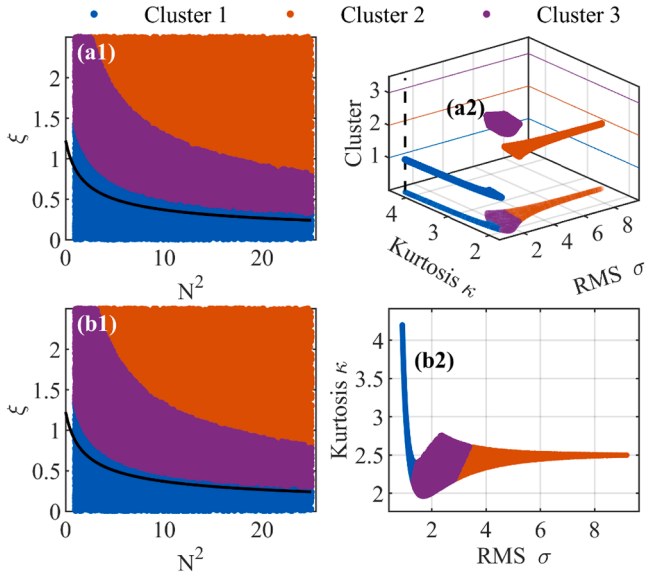


Fig. 5. Clustering into three sets for pulses generated during propagation in the normal dispersion regime. Fig. 5 (a) plots the clustering results in terms of the normalized intensity profile whilst Fig. 5 (b) plots the results in terms of σ and κ properties. Panels a1 and b1 plot the clusters according to the normalized propagation distance ξ and the number N . The black line marks the wave breaking condition (Eq. (6)). Panels a2 and b2 plot the clusters according to their rms duration σ and their kurtosis κ .

profile does change in this regime, as can be observed in the significant drop of the kurtosis [36] apparent from Fig. 5 (a2). The boundary of the blue cluster can be associated with a known analytical result that estimates the distance ξ_{WB} after which wave breaking occurs for a hyperbolic secant pulse [34,35]

$$\xi_{WB} = \sqrt{\frac{3}{2}} \frac{1}{\sqrt{N^2 + 1}} \quad (6)$$

which is plotted as the solid line in Fig. 4 (a1). Although strictly speaking accurate only for large N , the trend is nonetheless remarkably consistent with the results of the clustering. In the purple cluster that is associated with intermediate stage of evolution, the pulse undergoes both temporal broadening and change in shape, representing a transition stage before the pulse shape tends to evolve much less, with intensity profiles in the blue cluster showing little variation in kurtosis while still undergoing temporal broadening. This stage of evolution is well-understood, associated with the temporal profile taking on the same shape as the pulse spectrum (the dispersive Fourier transform or “spectron” regime also known as the gain-free similariton [22,37–39]). If the number of clusters is increased up to 4, it will mainly result in the splitting of the intermediate cluster into two subclusters of similar shape.

Finally in this section, we note that in contrast to the case of anomalous dispersion regime propagation where the clusters are mixed in the (σ, κ) plane, the clusters are much better isolated for normal dispersion regime propagation, with no overlap between them. In this context, it is interesting to consider whether direct clustering according to the rms duration and kurtosis could reveal a similar picture. These results are plotted in panels (b) of Fig. 5, and we can see that they are close to identical to the clusters based on the normalized intensity alone. The kurtosis seems therefore a relevant parameter to describe the pulse properties for this case where the pulse does not experience breaking into substructures. We noted however that the kurtosis-based clustering was less robust than the one based on the intensity profile, and that other configurations could emerge from the K-means algorithm, requiring the algorithm to be run multiple times to ensure consistent results. Cases such as this can be compared more systematically using silhouette

analysis [40], and whilst we have not performed this in detail for every result in this paper, for this particular case we find that σ/κ clustering is associated with a lower score (i.e. overlapping poorly separated clusters) than clustering based on normalized intensity.

5. Clustering profiles for both anomalous and normal dispersion regime dynamics

We now evaluate the performance of the clustering algorithm when input data combines intensity profiles arising from both anomalous and normal dispersion regime propagation. That is, we now input 40×10^3 profiles into the algorithm combining intensity profiles arising from soliton dynamics as well as normal dispersion broadening and wave breaking. In Fig. 6 panels (a), (b), (c) and (d), we show results of this clustering into two, three, four and five clusters respectively.

We first consider the results in Fig. 6 (a1) where we plot how the two clusters (red and blue, plotted in the (N^2, ξ) plane) sort the intensity

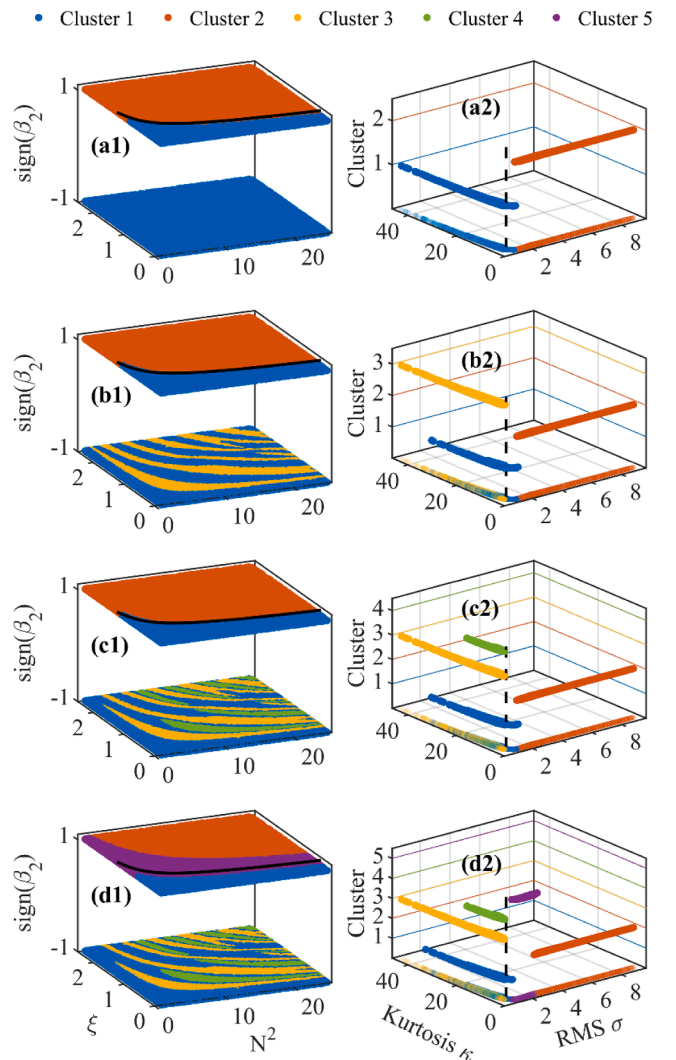


Fig. 6. Clustering of the temporal intensity profiles generated upon propagation in the normal and anomalous regimes of dispersion. Figs. 5(a), (b), (c) and (d) show results according to the different number of clusters: 2, 3, 4 and 5 respectively. Panels a1,b1,c1, d1 are clusters displayed according to the normalized propagation distance ξ , the soliton number N and the dispersion regime ($\text{sign}(\beta_2)$). The black line marks the wave breaking condition (Eq. (6)). Panels a2, b2, c2, d2 are the clusters plotted according to their rms duration σ and their kurtosis κ . The black dashed line is a visual guideline for the properties of the input pulse.

profiles relative to the known values of dispersion $\text{sgn}(\beta_2) = \pm 1$ in which they were generated. The classification shows that based only on the output normalized intensity profile, it is possible to infer to a large extent the dispersion regime in which pulse propagation took place. However, the classification is not perfect and for short propagation distances, some profiles known to arise from normal dispersion regime propagation are sorted into the same blue cluster associated primarily with soliton-like profiles in the anomalous dispersion regime. However, this can be readily understood because these are profiles associated with the initial stages of normal dispersion propagation (below the wave breaking distance) where the combination of dispersion and nonlinearity is not yet sufficient to lead to major variation in the temporal pulse duration. In fact, we can see from Fig. 6 (a2) where the clusters are plotted in the (σ, κ) plane that the algorithm selects different characteristics of the pulse profiles that follow distinct branches: one in which the duration changes little but the pulse kurtosis is significantly modified (blue); and another where the pulse shape changes little but its duration significantly changes (red).

When classifying into three and four clusters, it is essentially within the anomalous dispersion regime that we see additional cluster structure appear. Indeed, as can be seen in Fig. 6 (b1) and (b2), moving to three clusters leads to the identification of intensity profiles linked to periodic soliton compression dynamics, as discussed in section 3. For four clusters as shown in Fig. 6 (c1) and (c2), it is once again in the anomalous dispersion regime that new structure emerges, which in this case corresponds to the typical splitting of the higher soliton pulse into multiple subpulses. Finally, when the number of clusters is increased to 5, the results for the normal dispersion regime exhibits an intermediate cluster as discussed in Section 4. Therefore, mixing profiles resulting from focusing (anomalous) and defocusing (normal) nonlinear propagation leads to results fully consistent with the clusters previously observed when the two propagation regimes of dispersion are investigated separately.

In this regard, it is also of interest to explicitly plot the centroids of the five clusters in Fig. 7 (a), and because the centroid does not actually result from any physical propagation, we also plot in Fig. 7 (b) the normalized intensity profiles of the pulse in the dataset that is the closest to each centroid. This provides additional insights into the “typical profiles” in each cluster, and confirms our physical discussion above. Cluster 1 is associated with profiles close to the initial pulse, typical of low propagation distances or parameters leading to a recurrence of the initial higher-order soliton condition. Cluster 2 corresponds to pulse evolution beyond the wave breaking distance where the pulse has experienced strong temporal broadening and reshaping towards a flattened profile. Cluster 3 corresponds to significantly compressed profiles whereas the centroid of Cluster 4 is typical of the splitting of the higher order soliton into two subpulses with reduced intensity at $\tau = 0$. The centroid of Cluster 5 is typical of the intensity profile observed slightly

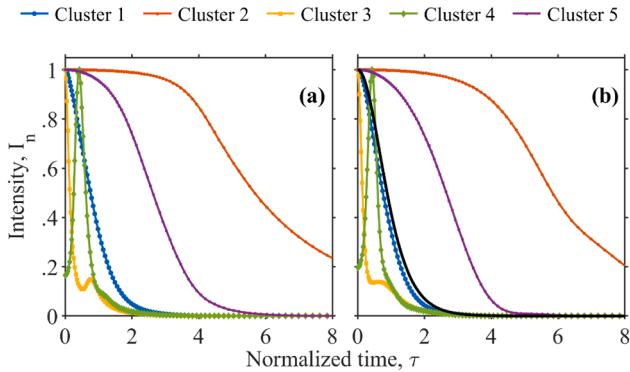


Fig. 7. (a) Centroids of the five clusters. (b) Normalized temporal intensity profiles in our dataset that are the closest to each centroid. The black curve is the input condition.

after the wave breaking point where the pulse exhibits a central parabolic like structure.

Finally, we have also studied the impact of the number of points considered after logarithmic resampling of the time grid. The clusters obtained for 2^L points with L between 8 and 4 are thus compared in Fig. 8. Although we saw in Section 1 that reducing the number of temporal samples can lead to a significant error in the evaluation of kurtosis for $L < 6$, this does not appear to be the case for the clustering results. Thus, even with only 16 points considered (panel d), the differences appearing are remarkably minor and only concern the anomalous regime which presents the most varied structures. Note that this number of temporal samples could even be further reduced if one only considers a single regime of propagation.

6. Conclusions

The major result of this work has been to show that centroid-based clustering of normalized intensity profiles can successfully isolate different classes of pulses associated with physically distinct regimes of nonlinear and dispersive fiber propagation. Remarkable for its simplicity, this approach shows how only temporal intensity profiles (i. e. without spectral intensity or phase information) with a very limited

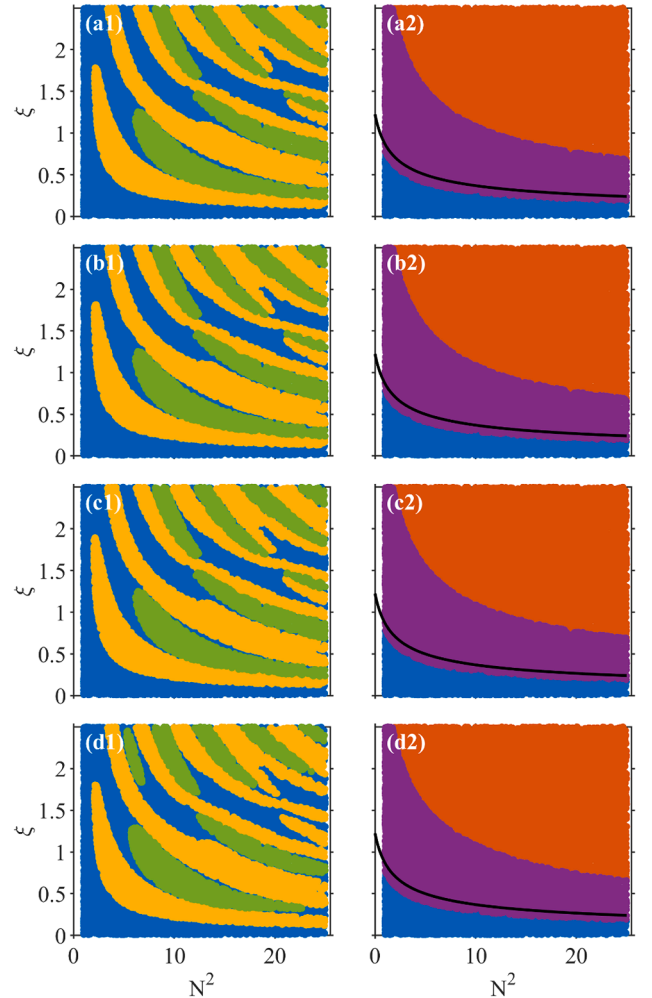


Fig. 8. Clustering of the temporal intensity profiles generated upon propagation in the anomalous and normal regimes of dispersion (panels 1 and 2, respectively). Fig. 8(a)–(d) show results for logarithmically spaced temporal grid with 2^L samples for $L = 8, 6, 5$ and 4, respectively. The color code is the same as panel Fig. 6(d). The black line indicates the wave breaking condition (Eq. (6)).

number of points nonetheless reveal sufficient similarities to allow physical classification of different propagation behavior. Extensions of this work could be in numerous directions, including for example the use of simultaneous spectral intensity profiles to generalize the technique to more difficult cases with initial chirp leading to effects such as spectral compression [41,42] which would be hard if not impossible to distinguish using temporal profile clustering only. Also, whilst we have considered background-free pulses in this analysis, it could be readily extended to nonlinear and dispersive pulse structures upon a continuous background, opening the possibility to obtain insights and empirical intuition into the properties and emergence of extreme rogue wave events. It is also possible that going beyond the simple NLSE and including additional higher order linear or non-linear terms will allow a useful cluster analysis of the highly complex process of supercontinuum generation. And in addition to NLSE-related problems, the process could readily be adapted to handle propagation in cavities typical of lasers or resonators supporting dissipative or cavity solitons. And of course, extensions of the clustering algorithm beyond the use of K-means may also prove useful [20], although this is beyond the scope of this present paper.

Funding

The work was funded by the Agence Nationale de la Recherche (Optimal project-ANR-20-CE30-0004; ANR-17-EURE-0002) and the Région Bourgogne-Franche-Comté.

CRediT authorship contribution statement

Anastasiia Sheveleva: Writing – review & editing, Visualization, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Andrei V. Ermolaev:** Writing – review & editing, Validation. **John M. Dudley:** Writing – original draft, Validation, Funding acquisition. **Christophe Finot:** Writing – original draft, Validation, Supervision, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

- G. Genty, L. Salmela, J.M. Dudley, D. Brunner, A. Kokhanovskiy, S. Kobtsev, S. K. Turitsyn, Machine learning and applications in ultrafast photonics, *Nat. Photon.* 15 (2021) 91–101.
- P. Freire, E. Manuylovich, J.E. Prilepsky, S.K. Turitsyn, Artificial neural networks for photonic applications—from algorithms to implementation: tutorial, *Adv. Opt. Photon.* 15 (2023) 739–834.
- J.W. Nevin, S. Nallaperuma, N.A. Shevchenko, X. Li, M.S. Faruk, S.J. Savory, Machine learning for optical fiber communication systems: An introduction and overview, *APL Photon.* 6 (2021) 121101.
- S. Boscolo, C. Finot, Artificial neural networks for nonlinear pulse shaping in optical fibers, *Opt. Laser Technol.* 131 (2020) 106439.
- N. Gautam, A. Choudhary, B. Lall, Comparative study of neural network architectures for modelling nonlinear optical pulse propagation, *Opt. Fiber Technol.* 64 (2021) 102540.
- S. Boscolo, J.M. Dudley, C. Finot, Modelling self-similar parabolic pulses in optical fibres with a neural network, *Result. Opt.* 3 (2021) 100066.
- H. Sui, H. Zhu, L. Cheng, B. Luo, S. Taccheo, X. Zou, L. Yan, Deep learning based pulse prediction of nonlinear dynamics in fiber optics, *Opt. Express* 29 (2021) 44080–44092.
- L. Salmela, N. Tspinakakis, A. Foi, C. Billet, J.M. Dudley, G. Genty, Predicting ultrafast nonlinear dynamics in fibre optics with a recurrent neural network, *Nat. Mach. Intell.* 3 (2021) 344–354.
- L.C.B. Silva, M.E.V. Segatto, Nonlinear autoregressive with external input neural network for predicting the nonlinear dynamics of supercontinuum generation in optical fibers, *J. Opt. Soc. Am. B* 40 (2023) 1292–1298.
- M. Närhi, L. Salmela, J. Toivonen, C. Billet, J.M. Dudley, G. Genty, Machine learning analysis of extreme events in optical fibre modulation instability, *Nat. Commun.* 9 (2018) 4923.
- A. Sheveleva, P. Colman, J.M. Dudley, C. Finot, Phase space topology of four-wave mixing reconstructed by a neural network, *Opt. Lett.* 47 (2022) 6317–6320.
- S. Boscolo, J.M. Dudley, C. Finot, Predicting nonlinear reshaping of periodic signals in optical fibre with a neural network, *Opt. Commun.* 542 (2023) 129563.
- X. Jiang, D. Wang, Q. Fan, M. Zhang, C. Lu, A.P.T. Lau, Physics-informed neural network for nonlinear dynamics in fiber optics, *Laser Photon. Rev.* (2022) 2100483.
- A.V. Ermolaev, A. Sheveleva, G. Genty, C. Finot, J.M. Dudley, Data-driven model discovery of ideal four-wave mixing in nonlinear fibre optics, *Sci. Rep.* 12 (2022) 1–11.
- M. Sorokina, S. Sygletos, S.K. Turitsyn, Sparse identification for nonlinear optical communication systems: SINO method, *Opt. Express* 24 (2016) 30433–30443.
- J.L. Callahan, J.V. Koch, B.W. Brunton, J.N. Kutz, S.L. Brunton, Learning dominant physical processes with data-driven balance models, *Nat. Commun.* 12 (2021) 1016.
- A.V. Ermolaev, M. Mabel, C. Finot, G. Genty, J.M. Dudley, Analysis of interaction dynamics and rogue wave localization in modulation instability using data-driven dominant balance, *Sci. Rep.* 13 (2023) 10462.
- A.V. Ermolaev, C. Finot, G. Genty, J.M. Dudley, Automating physical intuition in nonlinear fiber optics with unsupervised dominant balance search, *Opt. Lett.* 49 (2024) 4202–4205.
- S. Lloyd, Least squares quantization in PCM, *IEEE Transact. Inform. Theory* 28 (1982) 129–137.
- G. Seber, *Multivariate Observations*, John Wiley & Sons, 2009.
- G.P. Agrawal, *Nonlinear Fiber Optics*, Sixth Edition, Academic Press, San Francisco, CA, 2019.
- S.A. Akhmanov, V.A. Vysloukh, A.S. Chirkin, Self-action of wave packets in a nonlinear medium and femtosecond laser pulse generation, *Sov. Phys. Uspekhi* 29 (1986) 642.
- L.T. DeCarlo, On the meaning and use of kurtosis, *Psychol. Method.* 2 (1997) 292–307.
- J.W. Foreman, *Data smart: using data science to transform information into insight*, John Wiley & Sons, 2013.
- N.G. Gonzalez, D. Zibar, A. Caballero, I.T. Monroy, Experimental 2.5-Gb/s QPSK WDM phase-modulated radio-over-fiber link with digital demodulation by a K-means algorithm, *IEEE Photon. Technol. Lett.* 22 (2010) 335–337.
- J. Zhang, W. Chen, M. Gao, G. Shen, K-means-clustering-based fiber nonlinearity equalization techniques for 64-QAM coherent optical communication system, *Opt. Express* 25 (2017) 27570–27580.
- L.F. Mollenauer, R.H. Stolen, J.P. Gordon, W.J. Tomlinson, Extreme picosecond pulse narrowing by means of soliton effect in single-mode optical fibers, *Opt. Lett.* 8 (1983) 289–291.
- L.F. Mollenauer, R.H. Stolen, J.P. Gordon, Experimental observation of picosecond pulse narrowing and solitons in optical fibers, *Phys. Rev. Lett.* 45 (1980) 1095–1098.
- R.H. Stolen, L.F. Mollenauer, W.J. Tomlinson, Observation of pulse restoration at the soliton period in optical fibers, *Opt. Lett.* 8 (1983) 187–189.
- A. Shabat, V.E. Zakharov, Exact theory of two-dimensional self-focusing and one-dimensional self-modulation of waves in nonlinear media, *Sov. Phys. JETP* 34 (1972) 62.
- Y. Wang, F. Chen, S. Fu, J. Kong, A. Komarov, M. Klimczak, R. Buczyński, X. Tang, M. Tang, L. Zhao, Nonlinear Fourier transform assisted high-order soliton characterization, *New J. Phys.* 24 (2022) 033039.
- C.-M. Chen, P.L. Kelley, Nonlinear pulse compression in optical fibers: scaling laws and numerical analysis, *J. Opt. Soc. Am. B* 19 (2002) 1961–1967.
- G. Genty, S. Coen, J.M. Dudley, Fiber supercontinuum sources, *J. Opt. Soc. Am. B* 24 (2007) 1771–1785.
- D. Anderson, M. Desaix, M. Lisak, M.L. Quiroga-Teixeiro, Wave-breaking in nonlinear optical fibers, *J. Opt. Soc. Am. B* 9 (1992) 1358–1361.
- C. Finot, B. Kibler, L. Provost, S. Wabnitz, Beneficial impact of wave-breaking on coherent continuum formation in normally dispersive nonlinear fibers, *J. Opt. Soc. Am. B* 25 (2008) 1938–1948.
- C.-K. Rosenberg, D. Anderson, M. Desaix, P. Johansson, M. Lisak, Evolution of optical pulses towards wave breaking in highly nonlinear fibres, *Opt. Commun.* 273 (2007) 272–277.
- A. Zeytunyan, G. Yesayan, L. Mouradian, P. Kockaert, P. Emplit, F. Louradour, A. Barthélémy, Nonlinear-dispersive similariton of passive fiber, *J. Europ. Opt. Soc. Rap. Public.* 4 (2009) 09009.
- S.O. Iakushev, O.V. Shulika, I.A. Sukhoivanov, Passive nonlinear reshaping towards parabolic pulses in the steady-state regime in optical fibers, *Opt. Commun.* 285 (2012) 4493–4499.
- T. Jansson, Real-time Fourier transformation in dispersive optical fibers, *Opt. Lett.* 8 (1983) 232–234.
- P.J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, *J. Comput. Appl. Math.* 20 (1987) 53–65.
- A.V. Zohrabian, L.K. Mouradian, Compression of the spectrum of picosecond ultrashort pulses, *Quant. Electron.* 25 (1995) 1076.
- C. Finot, S. Boscolo, Design rules for nonlinear spectral compression in optical fibers, *J. Opt. Soc. Am. B* 33 (2016) 760–767.