Data-driven methods for analysing nonlinear propagation in optical fibres

Andrei Ermolaev

UBFC
UNIVERSITÉ
BOURGOGNE FRANCHE-COMTÉ

UNIVERSITÉ DE
FRANCHE-COMTÉ

THÈSE DE DOCTORAT DE L'ÉTABLISSEMENT UNIVERSITÉ BOURGOGNE FRANCHE-COMTÉ

PRÉPARÉE À L'UNIVERSITÉ DE FRANCHE-COMT É

École doctorale n°37

Sciences Pour l'Ingénieur et Microtechniques

Doctorat d'Optique et Photonique

par

Andrei V. Ermolaev

Data-driven methods for analysing nonlinear propagation in optical fibres

Méthodes basées sur les données pour l'analyse de la propagation non linéaire dans les fibres optiques

Thèse présentée et soutenue à Besançon, le 18 septembre 2024

Composition du Jury :

Alonso Miguel	Professeur, Ecole Centrale de Marseille, Institut Fresnel, Marseille, France	Rapporteur
Kutz J. Nathan	Professeur, University of Washington, Department of Applied Mathematics, Seattle, États-Unis	Rapporteur
Chernysheva Maria	Chargée de recherche (junior research group leader), Leibniz Institute of Photonic Technology, Jena, Allemagne	Examinateur
Finot Christophe	Professeur, Université de Bourgogne, Dijon, France	Examinateur
Jacquot Maxime	Professeur, Université Bourgogne - Franche-Comté, Besançon, France	Président du jury
Dudley John M.	Professeur, Université Bourgogne - Franche-Comté, Besançon, France	Directeur de thèse

N° | X | X | X |

"Scientific revolutions are more often driven by new tools than by new concepts."

- Thomas Kuhn, from *"The Structure of Scientific Revolutions"* [1]

Acknowledgments

I would first of all like to express my deep gratitude to my supervisor, John M. Dudley. His intuition and curiosity about new scientific approaches and tools, his constant willingness to engage in active discussions about results and future steps, his attentiveness as well as his sense of humor were an integral part of my PhD life and created an excellent working environment, for which I am very grateful to him. Working under his guidance has been an inspiring experience that has allowed me to learn a lot over these past three years.

I thank all the jury members — J. Nathan Kutz, Miguel Alonso, Maria Chernysheva, Maxime Jacquot, and Christophe Finot — for carefully reading my thesis and giving me valuable suggestions and ideas, and, of course, for the fruitful and encouraging discussion we had. I would like to further thank Christophe Finot and Goëry Genty for actively participating and contributing to all the scientific projects I worked on throughout my PhD. I would also like to acknowledge the ongoing commitment of J. Nathan Kutz and his colleagues to an open science and open data strategy. Their dedication to making research data and program codes accessible enabled us to focus directly on advancing and applying data-driven methods in nonlinear fiber optics.

I thank Bryan E. Kaiser for valuable discussions regarding the data-driven dominant balance approach and the implementation of the combinatorial model selection algorithm presented in Chapter 4 of this thesis. I am grateful to Anastasiia Sheveleva for numerous discussions on the experimental observation of ideal four-wave mixing dynamics in fibers and her contribution to the second chapter of this thesis. I want to also thank my office mates — Mehdi, Pierre-Jean and Jacob — for our engaging conversations, both scientific and otherwise. My special thanks are addressed to my friend Vahe, whose inquisitive mind often led to intriguing ideas through our lengthy debates.

I would also like to express my appreciation to all members of the administration and in particular the SPIM Doctoral School staff for their continued assistance and providing various opportunities. Additionally, I am thankful to John M. Dudley, Maxime Jacquot, François Courvoisier, Daniel Brunner, and Luc Froehly for their continuous help and support in pursuing my research career.

Of course, I can not forget to mention my close friends who have always stood by me. I would like to thank Vanya, Denis, Zlata and Fedja for their steadfast support and friendship since our school years. I would also like to mention the wonderful friends I made in Besançon — Amjad, Alexey, Paul, Arina, Valeria, Daria, Lera, Nastya, Vera, Alisa, Emmanuel — and many others. Lastly, I am grateful to my parents, Vadim and Elena, my sister Daria, my brother-in-law Yuri, their lovely daughter Nina and my grandparents, for their tireless support and encouragement.

Most of all, I am deeply grateful to my partner Veronika for her love and support, especially while I was writing my thesis in Achères, so I would like to acknowledge her contribution to all the four chapters of this thesis.

Contents

General introduction 1

1 Optical pulse propagation in fibres: Introduction 3

1.1 Optical pulse propagation in fibres . . . 3

1.1.1 Induced polarization . . . 4

1.1.2 Scalar model of the nonlinear pulse propagation . . . 5

1.1.3 Chromatic dispersion . . . 7

1.1.4 Nonlinear Schrödinger equation . . . 8

1.2 NLSE propagation regimes . . . 9

1.2.1 Dispersive propagation . . . 9

1.2.2 Self-phase modulation . . . 10

1.2.3 Temporal fibre solitons . . . 10

1.3 Modulation instability . . . 12

1.3.1 Linear stability analysis . . . 12

1.3.2 Modulation instability gain . . . 13

1.3.3 Noise-driven modulation instability . . . 14

1.3.4 Solitons on finite background . . . 15

1.4 Generalised nonlinear Schrödinger equation . . . 16

1.4.1 Raman scattering . . . 17

1.4.2 Soliton fission . . . 18

1.5 Machine learning methods in fibre optics: Overview . . . 19

1.5.1 State-of-the-art data-driven methods in fibre optics . . . 20

2 Data-driven discovery of FWM dynamics in fibres 23

2.1 Introduction . . . 23

2.2 Degenerate four wave mixing in optical fibres . . . 24

2.2.1 Hamiltonian formalism . . . 24

2.2.2 Phase space representation of dynamics . . . 26

2.3 Methodology of SINDy . . . 29

2.3.1 Extracting differential equation models from data . . . 29

2.3.2 Library of candidate functions 30
2.3.3 Sparse regression problem 30
2.3.4 Optimization methods 31
2.4 Results 32
2.4.1 Sparse identification of ideal FWM dynamics 32
2.4.2 Discovery of the ideal FWM model from noisy data 35
2.4.3 Application of SINDy to coupled amplitude equations 38
2.5 Summary, discussion & conclusion 41

3 Data-driven dominant balance analysis of MI 43
3.1 Introduction 43
3.2 Methodology of data-driven dominant balance 45
3.2.1 Equation space representation of dynamics 46
3.2.2 Gaussian mixture model framework 47
3.2.3 GMM clustering analysis of the Peregrine soliton dynamics 50
3.2.4 Model selection: Sparse principal component analysis 52
3.3 Results 57
3.3.1 Dominant balance analysis of SFB dynamics 57
3.3.2 Dominant balance analysis of noise-driven MI 61
3.4 Summary, discussion & conclusion 63

4 Automated unsupervised dominant balance search 65
4.1 Introduction 65
4.2 Combinatorial model selection algorithm 65
4.2.1 Dominant balance search criteria 67
4.2.2 Riemann wave propagation example 69
4.2.3 Uncertainties in the dominant balance models 72
4.2.4 Dominant balance search in spatio-spectral domain 74
4.3 Results 75
4.3.1 Riemann wave propagation 75
4.3.2 Optical wave breaking 77
4.3.3 Perturbed fundamental soliton propagation 79
4.3.4 Soliton fission induced by third-order dispersion 81
4.3.5 Soliton fission in the presence of TOD and Raman effects 82

Conclusion and perspectives 85

Bibliography **87**

A List of acronyms **99**

B Dominant balance analysis of optical wave breaking dynamics: Python code **101**

C List of publications **103**

General introduction

Over the past decades, the methods and tools of machine learning (ML) have led to remarkable technological advances that now impact many aspects of our lives. Automated image and speech recognition, the creation of large language models and the control of autonomous vehicles are just a few examples of real-world applications of ML. Moreover, the rapid development of this field is beginning to have a huge influence on the way we conduct research in various fields of science [2].

ML is a broad term describing the use of statistical methods and numerical algorithms to solve problems without explicit instructions. The ability of this diverse range of algorithms and approaches to recognise patterns in data, suggest optimal solutions, as well as to predict and control the evolution of complex dynamical systems, makes ML particularly suitable for research purposes [3, 4]. The development of ML-based methodologies for scientific applications is very timely given the increasing amount of data being collected in fields such as biology, neuroscience, astronomy and physics [5–7]. Recently, this has resulted in the emergence of data-driven methods specifically designed to control, predict, and interpret the dynamics of complex systems in these diverse areas of science [8, 9].

Current applications of ML in optics and photonics are numerous, ranging from reverse engineering of novel photonic structures, optical sensing [10], to active control and optimisation of mode-locked lasers [11–13]. Significant progress in these areas has been made through the use of genetic search algorithms and deep learning strategies enabling the development of model-free approaches to control and predict the dynamics of optical systems with many degrees of freedom (control parameters) [14][1].

At the same time, there is growing interest in developing data-driven methods that allow us to extract interpretable and generalisable models that provide a deeper understanding of the underlying physics [15], as opposed to implementing model-free approaches. The ability of these novel data-driven methods to complement conventional theoretical approaches in interpretation of the physics and provide model-based control strategies has the potential to open up many opportunities and applications in the field of nonlinear fiber optics, especially in highly nonlinear and chaotic propagation regimes.

In this thesis we present the application of data-driven approaches (based on both supervised and unsupervised learning) to different scenarios of nonlinear dynamics in optical fibres. The aim of this thesis is to demonstrate how these emerging data-driven techniques can be applied to the field of nonlinear fibre optics to improve our understanding of the physics underlying various propagation regimes and even make discoveries based on dynamical data.

[1] An overview of the state-of-the-art ML techniques in fiber optics is presented in Section 1.5 of Chapter 1.

Thesis outline

This thesis is organised as follows. In Chapter 1 we first present the background theory on optical pulse propagation in fibers. We derive the differential equation models governing the propagation and describe various fundamental phenomena arising under variety of conditions. In the end of Chapter 1 we also present an overview of the state-of-the-art data-driven methods in nonlinear fiber optics.

In Chapter 2 we focus on one of the recently proposed data-driven techniques, called the sparse identification of nonlinear dynamics (SINDy) [16], that enables to extract the governing differential equation models directly from dynamical data. We then present the first application of this technique to the fundamental process of four-wave mixing in optical fibers. In particular, we focus on the ability of SINDy to extract underlying differential equation models in nonlinear fiber optics from noisy dynamical data.

Chapter 3 presents another data-driven technique designed to identify the dominant physical interactions in local spatio-temporal regions [17] that we apply to the analysis of noise-driven modulation instability (MI) dynamics in fibers.

Finally, in Chapter 4 we describe the fully unsupervised approach [18] to the data-driven dominant balance technique presented in Chapter 3. We then apply this automated data-driven dominant balance search algorithm to study the dominant physical regimes in various scenarios of optical pulse propagation in fibers.

1

Introduction to optical pulse propagation in fibres

In this chapter, we describe the underlying physical principles governing the propagation of optical pulses in a fibre. We derive the basic equations describing the nonlinear and dispersive processes that play a major role in the dynamics. We then describe a number of fundamental effects and phenomena of nonlinear fibre optics such as the Kerr effect, the modulation instability, optical solitons and the Raman scattering.

1.1/ Optical pulse propagation in fibres

The dynamics of optical waves in non-magnetic media without charges and with zero current density can be described by the following set of Maxwell's equations, which are written here using the Gaussian unit system [19]

$$\nabla \cdot \mathbf{D} = 0, \tag{1.1a}$$

$$\nabla \cdot \mathbf{B} = 0, \tag{1.1b}$$

$$\nabla \times \mathbf{B} = \frac{1}{c}\frac{\partial \mathbf{D}}{\partial t}, \tag{1.1c}$$

$$\nabla \times \mathbf{E} = -\frac{1}{c}\frac{\partial \mathbf{B}}{\partial t}, \tag{1.1d}$$

where ∇ stands for the nabla operator $\hat{x}(\partial/\partial x) + \hat{y}(\partial/\partial y) + \hat{z}(\partial/\partial z)$ with $\hat{x}, \hat{y}, \hat{z}$ being the unit vectors along x, y, z directions, respectively, $(\cdot)$ and $(\times)$ denote the dot and cross products, and c is the speed of light in vacuum. Vectors $\mathbf{E} = \mathbf{E}(\mathbf{r}, t)$[1] and $\mathbf{B} = \mathbf{B}(\mathbf{r}, t)$ are the electric and magnetic field, respectively, and $\mathbf{D} = \mathbf{D}(\mathbf{r}, t)$ is the displacement field vector that satisfies the constitutive relation

$$\mathbf{D} = \mathbf{E} + 4\pi\mathbf{P}, \tag{1.2}$$

where $\mathbf{P} = \mathbf{P}(\mathbf{r}, t)$ is the induced polarisation. Now, taking the curl of equation 1.1d and then substituting Eq. 1.1c into it, yields

$$\nabla^2\mathbf{E} - \nabla(\nabla \cdot \mathbf{E}) - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\mathbf{E} - \frac{4\pi}{c^2}\frac{\partial^2}{\partial t^2}\mathbf{P} = 0. \tag{1.3}$$

Equation 1.3 is a rather general wave equation that is usually presented in a simplified form after making approximations specific to the particular optical problem [20].

[1] $\mathbf{r}$ is the radius vector $\hat{x}x + \hat{y}y + \hat{z}z$.

1.1.1/ Induced polarization

The induced polarisation is linked to the electric field vector by the following law

$$\mathbf{P} = \Upsilon \mathbf{E}, \tag{1.4}$$

where Υ is, in general, some nonlinear and nonlocal operator [21]. The calculation of $\mathbf{P}$ may require a quantum-mechanical approach to accurately account for the optical response of a medium in the vicinity of resonances [22]. For example, the polarisation can be expressed in terms of a dipole moment operator using the density matrix formalism [23].

Far from atomic resonances, the following general relation can be used to link $\mathbf{P}$ and $\mathbf{E}$

$$\begin{aligned}\mathbf{P}(\mathbf{r},t) = &\int_{-\infty}^{t} dt' \chi^{(1)}(t-t') \cdot \mathbf{E}(\mathbf{r},t') + \int_{-\infty}^{t} dt' \int_{-\infty}^{t} dt'' \chi^{(2)}(t-t', t-t'') : \mathbf{E}(\mathbf{r},t')\mathbf{E}(\mathbf{r},t'') + \\ &\int_{-\infty}^{t} dt' \int_{-\infty}^{t} dt'' \int_{-\infty}^{t} dt''' \chi^{(3)}(t-t', t-t'', t-t''') \vdots \mathbf{E}(\mathbf{r},t')\mathbf{E}(\mathbf{r},t'')\mathbf{E}(\mathbf{r},t''') + \ldots,\end{aligned} \tag{1.5}$$

where $\chi^{(i)}$ are tensor terms obtained from the expansion of the electric susceptibility. Note that Eq. 1.5 was obtained when neglecting the nonlocal response of the polarization to the external field, but taking into account its nonlinear nature. Spatial dispersion effects, however, may play an essential role near the interface between different media and in the confined volumes, where the electric susceptibility may also depend on spatial variables [24–26].

The first integral term in Eq. 1.5 describes the linear component of the polarisation $\mathbf{P}_L$, while the higher-order terms represent the nonlinear part of the polarisation $\mathbf{P}_{NL}$. The second order nonlinearity describes such effects as second harmonic generation, sum-frequency generation, and optical rectification. The third order susceptibility $\chi^{(3)}$ is responsible for various phenomena, including self-phase modulation, third harmonic generation and four wave mixing (FWM).

In optical fibres made of silica, the second order effects $\chi^{(2)}$ can be neglected due to the centro-symmetry of its structure. Thus, the second integral in Eq. 1.5 can be omitted. The main contribution to the nonlinear part of the polarisation comes from the third-order nonlinearity, since higher-order effects associated with higher-order tensor terms $\chi^{(4)}, \chi^{(5)}, \ldots$ are usually much weaker. The linear part of the polarisation satisfies the following constitutive relation for the linear part of the field displacement vector

$$\mathbf{D}_L = \mathbf{E} + 4\pi \mathbf{P}_L = \varepsilon \cdot \mathbf{E}, \tag{1.6}$$

where ε is the dielectric permittivity of a medium.

Equation 1.3 can now be rewritten as follows

$$\nabla^2 \mathbf{E} - \frac{1}{c^2}\frac{\partial^2}{\partial t^2}\mathbf{E} - \frac{4\pi}{c^2}\frac{\partial^2}{\partial t^2}\mathbf{P_L} = \frac{4\pi}{c^2}\frac{\partial^2}{\partial t^2}\mathbf{P_{NL}}, \tag{1.7}$$

where $\mathbf{P}_{NL}$ only accounts for third-order nonlinear effects represented by the tensor $\chi^{(3)}$. Note that the second term in Eq. 1.3 was neglected in Eq. 1.7 as ε is assumed to be independent of $\mathbf{r}$ [27], so with the use of Eqs. 1.1a and 1.6 one gets

$$\nabla \cdot \mathbf{D} = \varepsilon \nabla \cdot \mathbf{E} = 0. \tag{1.8}$$

1.1.2/ Scalar model of the nonlinear pulse propagation

As in many problems in optics, a complete description of the propagation of a vector field in a medium is a rather cumbersome task, so that simplified scalar models are usually used to simulate the dynamics. To derive such a model in the case of optical pulse propagation in a fibre, it is necessary to introduce additional assumptions about the structure of the optical field. First, we can assume that the optical field maintains its direction of polarisation throughout the propagation. Second, the propagating optical field is considered to be quasi-monochromatic that implies that its spectral width $\Delta\omega \ll \omega_0$ with ω_0 being the carrier frequency of a pulse. In isotropic media (such as silica), using the slowly varying pulse approximation, the field of a monochromatic beam propagating along the z-axis as well as the associated polarisation components can be written in the following way

$$\mathbf{E} = \frac{1}{2}\hat{x}\left[E(\mathbf{r},t)\exp(-i\omega_0 t) + c.c.\right], \tag{1.9a}$$

$$\mathbf{P}_L = \frac{1}{2}\hat{x}\left[P_L(\mathbf{r},t)\exp(-i\omega_0 t) + c.c.\right], \tag{1.9b}$$

$$\mathbf{P}_{NL} = \frac{1}{2}\hat{x}\left[P_{NL}(\mathbf{r},t)\exp(-i\omega_0 t) + c.c.\right], \tag{1.9c}$$

where we separated the rapidly varying parts of electric field[2] and polarisation components by introducing the slowly varying functions $E(\mathbf{r},t), P_L(\mathbf{r},t), P_{NL}(\mathbf{r},t)$ with respect to the optical period. The above approximations allow us to significantly simplify the structure of the polarisation components. First, due to the isotropic properties of silica both $\chi^{(1)}$ and $\chi^{(3)}$ tensors can be represented by single components $\chi^{(1)}_{xx}$ and $\chi^{(3)}_{xxxx}$, repectively, describing the corresponding linear and nonlinear response of the medium to the incident field linearly polarised along $\hat{x}$. Next, substituting Eq. 1.9 into Eq. 1.5, assuming the instantaneous nonlinear response[3] and neglecting third harmonic generation process, yields the following result for the polarisation components

$$\begin{aligned} P_L(\mathbf{r},t) = \int_{-\infty}^{\infty} dt' \chi^{(1)}_{xx}(t-t')E(\mathbf{r},t')\exp\left[i\omega_0(t-t')\right] = \\ \frac{1}{2\pi}\int_{-\infty}^{\infty} d\omega \tilde{\chi}^{(1)}_{xx}(\omega)\tilde{E}(\mathbf{r},\omega-\omega_0)\exp\left[-i(\omega-\omega_0)t\right], \end{aligned} \tag{1.10}$$

where the tilde over the variable E denotes the Fourier transform

$$\tilde{E}(\mathbf{r},\omega) = \int_{-\infty}^{\infty} dt E(\mathbf{r},t)\exp\left(i\omega t\right) = \mathcal{F}[E(\mathbf{r},t)], \tag{1.11}$$

and

$$P_{NL} \approx \frac{3}{4}\chi^{(3)}_{xxxx}|E(\mathbf{r},t)|^2 E(\mathbf{r},t) = \varepsilon_{NL}E(\mathbf{r},t). \tag{1.12}$$

Solution of Eq. 1.7 can be found by considering its right-hand-side as a small perturbation. In other words, the nonlinear contribution to the total polarisation is assumed to be much smaller than the contribution from the linear part of polarisation. This assumption is justified due to relatively small changes of refractive index induced by nonlinearity [22]. On this basis, to find a first-order perturbation theory solution of Eq. 1.7 for the moment

[2] In the above equation, *c.c.* stands for the complex conjugate

[3] Mathematically this corresponds to expressing the time varying part of $\chi^{(3)}$ as delta functions $\delta(t-t')$.

we can neglect the dependence of P_{NL} on the field intensity $|E(\mathbf{r},t)|^2$ and consider ε_{NL} as a constant term. Finally, by substituting Eqs. 1.9, 1.10, 1.12 into Eq. 1.7 and taking the Fourier transform of both sides of it, we find that the Fourier transform of the field $\tilde{E}(\mathbf{r},\omega-\omega_0)$ satisfies the Helmholtz equation [22]

$$\nabla^2\tilde{E}+\varepsilon(\omega)\frac{\omega^2}{c^2}\tilde{E}=0, \tag{1.13}$$

where $\varepsilon(\omega)=1+4\pi\chi^{(1)}_{xx}+\varepsilon_{NL}$. The solution of Eq. 1.13 can be found via method of separation of variables assuming a solution of the form

$$\tilde{E}(\mathbf{r},t)=F(x,y)\tilde{A}(z,\omega-\omega_0)\exp(i\beta_0 z), \tag{1.14}$$

where $\tilde{A}(z,\omega-\omega_0)$ is the slowly varying function of z, $F(x,y)$ is the transverse modal distribution, and β_0 is the wave number at ω_0. Substitution of Eq. 1.14 into the Helmholtz equation (see Eq. 1.13) gives

$$\frac{1}{F}\left(\frac{\partial^2F}{\partial x^2}+\frac{\partial^2F}{\partial y^2}\right)+\varepsilon(\omega)\frac{\omega^2}{c^2}=\beta_0^2-2i\beta_0\frac{1}{\tilde{A}}\frac{\partial\tilde{A}}{\partial z}-\frac{1}{\tilde{A}}\frac{\partial^2\tilde{A}}{\partial z^2}=\tilde{\beta}^2. \tag{1.15}$$

As the left-hand side part only depends on the spatial variables x and y and the right-hand side part only depends on the spatial variable z, we can put both sides to be equal to some constant value $\tilde{\beta}^2$ ($\tilde{\beta}$ is the wavenumber to be defiend later). As $\tilde{A}$ is the slowly varying function of z, we can neglect the second-order derivative term $\partial^2\tilde{A}/\partial z^2$, since $\partial^2\tilde{A}/\partial z^2\ll\tilde{A}$. Therefore, $F(x,y)$ and $\tilde{A}(z,\omega-\omega_0)$ satisfy the following set of equations

$$\frac{\partial^2F}{\partial x^2}+\frac{\partial^2F}{\partial y^2}+\left[\varepsilon(\omega)\frac{\omega^2}{c^2}-\tilde{\beta}^2\right]F=0, \tag{1.16a}$$

$$2i\beta_0\frac{\partial\tilde{A}}{\partial z}+\left(\tilde{\beta}^2-\beta_0^2\right)\tilde{A}=0. \tag{1.16b}$$

Further solution requires the perturbation theory approach. First, the dielectric constant can be represented as

$$\varepsilon=(n+\Delta n)^2\approx n^2+2n\Delta n, \tag{1.17}$$

with Δn being a small perturbation

$$\Delta n=n_2|E|^2+i\frac{c\alpha}{2\omega}, \tag{1.18}$$

where α is the parameter describing the linear losses[4]. The dependence of the refractive index on the field intensity is known as the optical Kerr effect, which plays a central role in nonlinear optics. The zeroth order perturbation theory solution of Eq. 1.16a can be found by setting $\Delta n=0$ which allows to obtain the approximation of the modal distribution $F(x,y)$ and the wavenumber $\beta(\omega)$. Then, the first order perturbation theory solution yields

$$\tilde{\beta}(\omega)=\beta(\omega)+\Delta\beta(\omega), \tag{1.19}$$

where

$$\Delta\beta(\omega)=\frac{\omega^2n(\omega)}{c^2\beta(\omega)}\frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}dxdy\Delta n(\omega)|F(x,y)|^2}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}dxdy|F(x,y)|^2}. \tag{1.20}$$

[4]Note that to keep our notation similar to that used in the literature, we sometimes define the same symbols to mean different things in different chapters. Whenever this is the case, it will be associated with a footnote.

1.1.3/ Chromatic dispersion

The dependence of the refractive index on the frequency $n(\omega)$ [or equivalently of the dielectric permittivity $\varepsilon(\omega)$] is known as the chromatic dispersion. At the fundamental level, the chromatic dispersion arises from the oscillations of bound electrons under the effect of the incident field [19]. In optical fibres, chromatic dispersion plays a major role in optical pulse propagation, as it causes different spectral components to propagate at different phase velocities. Mathematically, chromatic dispersion in optical fibres is accounted for by $\beta(\omega)$ (see Eqs. 1.16, 1.19). With the use of approximation $\tilde{\beta}(\omega)^2 - \beta_0^2 \approx 2\beta_0[\tilde{\beta}(\omega) - \beta_0]$, Eq. 1.16b can be written in the following form

$$\frac{\partial \tilde{A}}{\partial z} = i[\beta(\omega) + \Delta\beta(\omega) - \beta_0]\,\tilde{A}, \tag{1.21}$$

clearly illustrating how the propagating spectral components acquire a phase shift that depends on the material dispersion [through $\beta(\omega)$] and the intensity [through $\Delta\beta(\omega)$]. To solve the Eq. 1.21, we must obtain the exact functional form of $\beta(\omega)$, which is quite a difficult task. In this case, it is common to expand $\beta(\omega)$ in a Taylor series with respect to the frequency ω_0

$$\beta(\omega) = n(\omega)\frac{\omega}{c} = \beta_0 + (\omega - \omega_0)\beta_1 + \frac{1}{2}(\omega - \omega_0)^2\beta_2 + \frac{1}{6}(\omega - \omega_0)^3\beta_3 + \ldots, \tag{1.22}$$

with $\beta_0 = \beta(\omega_0)$ and

$$\beta_m = \left.\frac{d^m\beta}{d\omega^m}\right|_{\omega=\omega_0}, \quad m = 1, 2, 3, \ldots . \tag{1.23}$$

The parameters β_1 and β_2 from the Taylor expansion are linked to the refractive index in the following way

$$\beta_1 = \frac{1}{v_g} = \frac{n_g}{c} = \frac{1}{c}\left(n + \omega\frac{dn}{d\omega}\right), \tag{1.24a}$$

$$\beta_2 = \frac{1}{c}\left(2\frac{dn}{d\omega} + \omega\frac{d^2n}{d\omega^2}\right), \tag{1.24b}$$

and have the direct physical connection with the group velocity v_g and the group velocity dispersion (GVD) governed by the β_2 parameter and responsible for pulse broadening. The effect of higher-order dispersion is usually comparably smaller than that of the GVD. However, higher-order terms have to be included in the vicinity of zero-dispersion wavelength (the spectral region where the GVD contribution vanishes) and for ultrashort pulses (see Section 1.4). For example, the term proportional to the β_3 coefficient in Eq. 1.22 is called the third order dispersion (TOD).

At the moment we neglect the contributions related to the higher-order dispersion. This assumption is justified for optical pulses with a spectral width $\Delta\omega \ll \omega_0$. It is important to note that although chromatic dispersion can be introduced into the propagation equation independently of nonlinearity, nonlinear effects in optical fibres manifest themselves quite differently depending on the sign of the GVD parameter. It is common to distinguish two different regimes of the optical pulse propagation: the normal dispersion regime when $\beta_2 > 0$ and the anomalous dispersion regime when $\beta_2 < 0$. Of particular interest is the anomalous dispersion regime, where the red-shifted spectral components travel faster than the blue-shifted. This effect is known to lead to the propagation scenario in which a balance between chromatic dispersion and nonlinearity can support soliton propagation [22]. This point will be discussed in detail in Section 1.2.3

1.1.4/ Nonlinear Schrödinger equation

The propagation equation can be written in the time domain by taking the inverse Fourier transform of Eq. 1.21 in the form

$$A(z,t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} d\omega \tilde{A}(z,\omega-\omega_0)\exp -i(\omega-\omega_0)t = \mathcal{F}^{-1}[\tilde{A}(z,\omega-\omega_0)]. \tag{1.25}$$

For the $\Delta\beta(\omega)$ term in the right-hand side of Eq. 1.16b we can perform a similar expansion as in Eq. 1.19

$$\Delta\beta(\omega) = \Delta\beta_0 + (\omega-\omega_0)\Delta\beta_1 + \frac{1}{2}(\omega-\omega_0)^2\Delta\beta_2 + \frac{1}{6}(\omega-\omega_0)^3\Delta\beta_3 + \ldots \tag{1.26}$$

and keep only the zeroth-order term that accounts for the fibre loss and nonlinearity, such that $\Delta\beta \approx \Delta\beta_0$. This yields[5]

$$\frac{\partial A}{\partial z} + \beta_1\frac{\partial A}{\partial t} + i\frac{\beta_2}{2}\frac{\partial^2 A}{\partial t^2} = i\Delta\beta_0 A. \tag{1.27}$$

Finally, assuming that the transverse field distribution $F(x,y)$ does not vary much over the pulse bandwidth, using $\beta(\omega) \approx n(\omega)\omega/c$ and neglecting the fibre losses, we get

$$\frac{\partial A}{\partial z} + \beta_1\frac{\partial A}{\partial t} + i\frac{\beta_2}{2}\frac{\partial^2 A}{\partial t^2} = i\gamma(\omega_0)|A|^2 A, \tag{1.28}$$

where $\gamma(\omega_0)$ is the nonlinearity parameter, defined as

$$\gamma(\omega_0) = \frac{n_2(\omega_0)\omega_0}{cA_{\text{eff}}}, \tag{1.29}$$

with

$$A_{\text{eff}} = \frac{\left(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dxdy|F(x,y)|^2\right)^2}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} dxdy|F(x,y)|^4}. \tag{1.30}$$

Equation 1.28 can be written in the reference frame associated with the group velocity of the optical pulse. Using the following transformation

$$T = t - z/v_g = t - z\beta_1, \tag{1.31}$$

one gets

$$i\frac{\partial A}{\partial z} - \frac{\beta_2}{2}\frac{\partial^2 A}{\partial T^2} + \gamma|A|^2 A = 0. \tag{1.32}$$

Equation 1.32 is referred to as the nonlinear Schrödinger equation (NLSE) because of its similarity to the quantum-mechanical Schrödinger equation with a nonlinear potential. This particular mathematical form of NLSE appears in various fields of nonlinear science and describes the nonlinear wave dynamics in hydrodynamical systems, Bose-Einstein condensates, and plasma.

[5] Note that these last steps are equivalent to inserting the ansatz $\mathbf{E} = \frac{1}{2}\hat{x}\Big[F(x,y)A(z,t)\exp(i\beta_0 z - i\omega_0 t) + c.c.\Big]$ into the Eq. 1.7 and introducing the same assumptions as described.

1.2/ NLSE PROPAGATION REGIMES

The NLSE introduced in Section 1.1 supports different propagation regimes depending on the properties of input optical pulse (in particular, initial pulse width T_0 and the peak power P_0)[6], as well as the fibre parameters β_2 and γ. In order to characterise these different propagation regimes, it is common to introduce two length scales indicating typical propagation distances at which nonlinearity and GVD dominate the dynamics

$$L_D = \frac{T_0^2}{|\beta_2|}, \tag{1.33a}$$

$$L_{NL} = \frac{1}{\gamma P_0}, \tag{1.33b}$$

where L_D and L_{NL} are called the dispersion and nonlinear lengths, respectively. An important physical quantity is the ratio of the introduced characteristic lengths

$$\frac{L_D}{L_{NL}} = \frac{\gamma P_0 T_0^2}{|\beta_2|} \tag{1.34}$$

connecting all four parameters T_0, P_0, γ, and β_2 together. The specific limiting cases when $L_D/L_{NL} \ll 1$ and $L_D/L_{NL} \gg 1$ will be discussed in the following sections.

1.2.1/ DISPERSIVE PROPAGATION

The dispersion-dominated regime occurs when $L_D/L_{NL} \ll 1$, which leads to a negligibly small relative contribution of nonlinearity. Mathematically, this regime can be studied by setting $\gamma = 0$ in Eq. 1.32, which yields the following simplified equation

$$i\frac{\partial A}{\partial z} - \frac{\beta_2}{2}\frac{\partial^2 A}{\partial T^2} = 0, \tag{1.35}$$

with a similar mathematical structure to the paraxial wave equation governing the diffraction of continous wave light in one of its transverse directions [19]. The solution of Eq. 1.35 can be readily found using the Fourier-transform method

$$\tilde{A}(z,\omega) = \tilde{A}(0,\omega)\exp\left(i\frac{\beta_2}{2}\omega^2 z\right), \tag{1.36}$$

$$A(z,T) = \frac{1}{2\pi}\int_{-\infty}^{\infty} d\omega \tilde{A}(0,\omega)\exp\left(i\frac{\beta_2}{2}\omega^2 z - i\omega T\right). \tag{1.37}$$

It can be seen from Eq. 1.36 that in the absence of the nonlinearity, the second order dispersion does not affect the pulse spectrum. At the same time, from Eq. 1.37 it is evident how the GVD leads to the time domain pulse broadening and phase modulation, while allowing an initially symmetric optical pulse to preserve symmetry [21]. Specifically, the pulse broadening arises due to the time delay in the arrival of different spectral components. In the normal dispersion regime ($\beta_2 > 0$), the red components travel faster than blue components, and vice versa in the anomalous dispersion regime ($\beta_2 < 0$). These effects, however, can be compensated by nonlinearity, as will be shown in the following sections.

[6]Pulse width definition will be given in Section 1.2.3

1.2.2/ Self-phase modulation

Another limiting case $L_D/L_{NL} \gg 1$ corresponds to optical pulse propagation dominated by nonlinearity, with GVD contribution being comparably small. This regime is associated with the phenomenon called self-phase modulation (SPM) [22] resulting in the spectral broadening of the input pulse[7]. Following a similar approach as in Section 1.2.1, we can now neglect the β_2 term in Eq. 1.32

$$i\frac{\partial A}{\partial z} + \gamma|A|^2 A = 0. \tag{1.38}$$

Next, by substituting the following ansatz $A(z,T) = |A(z,T)|\exp(i\phi_{NL})$ in Eq. 1.38, one can find that the pulse temporal amplitude does not change during the propagation and the general solution reads as

$$A(z,T) = A(0,T)\exp\left(i\phi_{NL}\right), \tag{1.39}$$

with

$$\phi_{NL}(z,T) = \gamma|A(0,T)|^2 z, \tag{1.40}$$

where ϕ_{NL} denotes the nonlinear phase acquired during the propagation. Thus, the SPM leads to the intensity-dependent phase shift that increases with distance. At the same time, the time dependence of ϕ_{NL} leads to spectral changes of the input pulse throughout its evolution. Specifically, the instantaneous optical frequency varies across the pulse with

$$\delta\omega(T) = -\frac{\partial\phi_{NL}(z,T)}{\partial T} = -\gamma\frac{\partial|A(0,T)|^2}{\partial T}z, \tag{1.41}$$

defining the frequency difference with respect to the central frequency ω_0. Thus, from Eq. 1.41, we can see that the SPM leads to spectral broadening of the pulse as new spectral components are generated continuously during propagation[8].

1.2.3/ Temporal fibre solitons

In the limiting cases discussed above, the simplified structure of the evolution equation allows to derive analytical solutions for the propagation of specific input pulses. A particular example is the input unchirped Gaussian pulse

$$A(0,T) = \sqrt{P_0}\exp\left(-\frac{T^2}{2T_0^2}\right), \tag{1.42}$$

for which the solution of the evolution equation can be written in an explicit analytical form when considering the isolated action of GVD or SPM effects [22][9]. Contrary to that, in the case when the GVD and SPM effects become comparable (i.e. for $L_D/L_{NL} \sim 1$), both effects have to be considered simultaneously. The combined action of GVD and SPM leads to qualitatively different propagation regimes. In this case, it still possible to study the pulse propagation using approximate analytical and semi-analytical methods for specific initial conditions. Alternatively, the NLSE evolution of an arbitrary input pulse

[7] This effect is known to be a temporal analog of the spatial self-focusing effect in noninear optics.
[8] The dependence of $\delta\omega$ on time is called frequency chirp.
[9] In the above formula T_0 stands for the half-width at $1/e$ intensity level and linked to the full width at half maximum through the following relation $T_{\mathrm{FWHM}} = 2(\ln 2)^{1/2}T_0 \approx 1.665T_0$.

(see Eq. 1.32) for given fibre parameters can be studied numerically using the split-step Fourier method [28].

It is instructive to first consider some special analytical solutions of the NLSE and their properties. The NLSE has infinitely many exact solutions, while more complex solutions can be expressed as a nonlinear superposition of simpler solutions [21]. The inverse scattering transform method developed by Zakharov and Shabat in 1972 [29] allows to find and construct the exact analytical solutions of the NLSE. Another important theoretical result is that the NLSE admits soliton solutions [30]. In nonlinear fibre optics, temporal solitons are pulse-like solutions arising due to the interaction of dispersive and nonlinear effects, which allows these pulses to propagate over long distances without distortion under certain conditions [31]. Such solutions of Eq. 1.32 can be found via inverse scattering transform method. In the anomalous dispersion regime the first-order soliton solution reads as[10]

$$A(z,T) = \sqrt{P_0}\,\mathrm{sech}\,(T/T_0)\exp\,(i\gamma P_0 z/2), \tag{1.43}$$

where sech stands for the hyperbolic secant function. This solution is also called the fundamental soliton [22] because its shape and spectrum remain unchanged during propagation due to the full compensation of the two chirp contributions arising from the GVD (for $\beta_2 < 0$) and SPM effects.

The soliton-like solutions in optical fibres can be characterised using the parameter N called the soliton number

$$N^2 = \frac{L_D}{L_{NL}} = \frac{\gamma P_0 T_0^2}{|\beta_2|}. \tag{1.44}$$

In particular, the propagation of the fundamental soliton (see Eq. 1.43) can be supported for such a combination of fibre parameters β_2 and γ, peak power P_0 and temporal duration T_0 of the input sech pulse that $N = 1$.

For input sech pulses satisfying the condition $N = 2, 3, 4, \ldots$ in Eq. 1.44, one can observe the propagation of higher-order solitons that exhibit recurrent periodic temporal and spectral evolution. For some of the higher-order solitons it is still possible to find the closed-form analytical solutions, however their mathematical structure becomes more cumbersome [22]. An important characteristic of all higher-order solitons is the soliton period z_0, which gives the propagation distance at which the higher-order soliton recovers its original shape

$$z_0 = \frac{\pi}{2} L_D = \frac{\pi}{2}\frac{T_0^2}{|\beta_2|}. \tag{1.45}$$

[10] "Dark soliton" solutions can also be supported in the normal dispersion regime ($\beta_2 > 0$).

1.3/ Modulation instability

Another important phenomenon arising due to the interplay between the dispersive and nonlinear effects is the modulation instability (MI) [32, 33][11]. The MI has been actively studied since the 1960s in diverse fields, including hydrodynamics [34], Bose-Einstein condensates, plasmas [35], and nonlinear fibre optics. In the context of the optical fibre propagation, the MI of the NLSE describes the instability of the continuous wave injected into an optical fibre in the presence of small initial perturbations that experience exponential growth at the expense of a strong input wave. Specifically, the MI is observed in the anomalous dispersion regime ($\beta_2 < 0$) and leads to the breakup of a continuous wave into a sequence of ultrashort pulses in the presence of a small periodic perturbation [36] or noise [37].

1.3.1/ Linear stability analysis

Let us first consider the propagation of a continuous wave in an optical fibre in the absence of perturbations. For the input continuous wave, the steady-state solution of Eq. 1.32 reads as[12]

$$\bar{A} = \sqrt{P_0}\exp(i\phi_{NL}), \tag{1.46}$$

where P_0 is the power of the incident continuous wave and $\phi_{NL} = \gamma P_0 z$ (see Eq. 1.40). Linear stability analysis consists in study of the effect of a small perturbation on the steady-state solution. First, the perturbed solution can be written as

$$A = (\sqrt{P_0} + a)\exp(i\phi_{NL}), \tag{1.47}$$

where $a = a(z,T)$ is a small, complex-valued function. Inserting Eq. 1.47 into the NLSE (see Eq. 1.32), yields

$$i\frac{\partial a}{\partial z} - \frac{\beta_2}{2}\frac{\partial^2 a}{\partial T^2} + \gamma P_0(a + a^*) = 0, \tag{1.48}$$

where only the linear terms with respect to a were retained due to the smallness of the perturbation ($|a|^2 \ll P_0$). Next, searching for the general solution for a in the following form

$$a(z,T) = u\exp\left[i(Kz - \Omega T)\right] + v\exp\left[-i(Kz - \Omega T)\right], \tag{1.49}$$

where K and Ω represent the wavenumber and frequency of the perturbation, respectively, Eq. 1.48 can be converted to the set of two linear equations for variables u and v [22]. Using standard methods for solving the set of linear equations for u and v, it can be shown that nontrivial solutions exist only if the following dispersion relation is satisfied

$$K = \pm\frac{1}{2}|\beta_2\Omega|\left[\Omega^2 + \mathrm{sgn}(\beta_2)\Omega_c^2\right]^{1/2}, \tag{1.50}$$

where $\mathrm{sgn}(\beta_2) = \pm 1$ is the sign function and $\Omega_c = 4\gamma P_0/|\beta_2|$. Finally, by substituting Eq. 1.50 into the general solution for a (see Eq. 1.49), it becomes obvious that the stability of the injected continuous wave is determined by the sign of the GVD parameter β_2. In the normal dispersion regime ($\beta_2 > 0$), the wave number K is real for every Ω, which makes the solution stable with respect to small perturbations. The opposite is observed in the anomalous dispersion regime ($\beta_2 < 0$), where K becomes imaginary for $|\Omega| < \Omega_c$, leading to an exponential growth of the perturbation with z. This process results in spontaneous temporal modulation of the input continuous wave and is called the MI.

[11] Also called the Benjamin-Feir or Bespalov-Talanov instability.
[12] See Section 1.2.2.

1.3.2/ Modulation instability gain

In order to characterise the development of MI, we can introduce the MI gain for $|\Omega| < \Omega_c$, which has the physical meaning of the growth rate of the perturbation with frequency Ω

$$g(\Omega) = 2\,\mathrm{Im}(K) = |\beta_2 \Omega| \left(\Omega_c^2 - \Omega^2\right)^{1/2}, \tag{1.51}$$

where Im stands for the imaginary part. It can be seen from Eq. 1.51 that $g(\Omega)$ is symmetric with respect to $\Omega = 0$ with $g(\Omega = 0) = 0$. Moreover, it is straightforward to find the spectral positions corresponding to the maximum amplification

$$\Omega_{\max} = \pm \frac{\Omega_c}{\sqrt{2}} = \pm \left(\frac{2\gamma P_0}{|\beta_2|}\right)^{1/2}, \tag{1.52}$$

at which the MI gain reaches its peak value

$$g_{\max} = g(\Omega_{\max}) = 2\gamma P_0. \tag{1.53}$$

Figure 1.1 shows the shape of MI gain spectrum curve plotted against $\Omega/\Omega_{\max}$ over the range $|\Omega| < \Omega_c$.

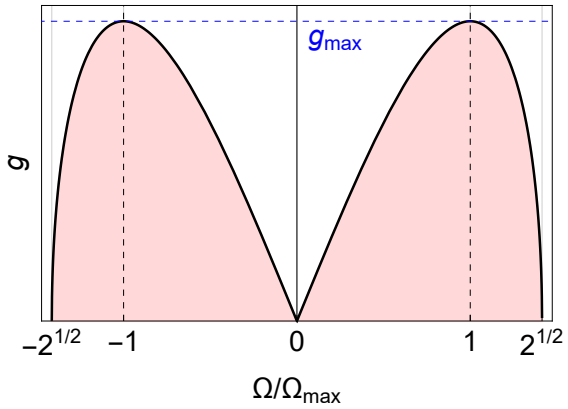


Figure 1.1: The MI gain spectrum plotted as a function of $\Omega/\Omega_{\max}$ (see Eq. 1.51). The black dashed vertical lines indicate the position corresponding to the maximum MI gain $g_{\max} = g(\Omega_{\max}) = 2\gamma P_0$ (highlighted by the blue dashed line).

The linear stability analysis results presented above describe only the initial stage of the MI propagation. Later MI dynamics leads to the emergence of localised ultrashort structures exhibiting temporal and/or spatial periodicity [21,36]. A complete theoretical description of the dynamics of MI under certain initial conditions requires the use of rather complex approaches of mathematical physics and statistics [38–42]. These aspects will be discussed in Chapter 3 of this thesis.

An interpretation of the MI dynamics in optical fibres can be found by considering the four-wave mixing (FWM) process with the phase-matching condition satisfied through the SPM [22]. The FWM is one of the third-order nonlinear parametric effects (characterised by $\chi^{(3)}$ susceptibility) describing the interaction of four optical field components with different frequencies (equivalent to the interaction of four photons in quantum-mechanics). In the most general case, the manifestation of the FWM process requires conservation of energy (matching of the frequencies) and fulfilment of the phase-matching condition

(matching of the wave vectors) for all four interacting components. However, the special case describing the interaction of a single intense pump optical field (at frequency ω_0) with two weak frequency sidebands (at frequencies ω_1 and ω_2) is of particular interest for optical fibres[13]. In this case, when the phase-matching condition is satisfied, two symmetric sidebands with a frequency shift Ω_s with respect to ω_0

$$\Omega_s = \omega_0 - \omega_1 = \omega_2 - \omega_0. \tag{1.54}$$

can be enhanced at the expense of a strong pump. It can be shown that in the anomalous dispersion regime the phase-matching condition for the degenerate FWM process occurs when $\Omega_s = (2\gamma P_0/|\beta_2|)$ with P_0 being here the input pump power. This frequency shift, in fact, coincides with the spectral position of the maximum MI gain obtained from the linear stability analysis (see Eq. 1.52). Indeed, a strong pump wave at frequency ω_0 transfers its energy most efficiently to the symmetric sidebands at $\omega_0 \pm \Omega_s$ as a result of the FWM process phase-matched by the SPM. This result highlights the idea of the frequency domain MI interpretation through the degenerate FWM process [22].

FWM is a central nonlinear process in optical fibre propagation, and in many areas of nonlinear science. The underlying theoretical background of FWM dynamics will be discussed later in Chapter 2.

1.3.3/ Noise-driven modulation instability

Isolated propagation of the pump wave in the anomalous dispersion regime of an optical fibre can also lead to MI in the presence of random noise perturbations. This phenomenon is known as spontaneous (or noise-driven) MI and was shown to lead to the random formation of localized and periodic spatio-temporal structures [43, 44].

In the most general case, the NLSE in the anomalous dispersion regime (also refered to as the one-dimensional focusing NLSE in the MI literature) and the initial conditions leading to the MI can be written in the following dimensionless form

$$i\frac{\partial \psi}{\partial \xi} + \frac{\partial^2 \psi}{\partial \tau^2} + |\psi|^2\psi = 0, \tag{1.55}$$

with

$$\psi(\xi = 0, \tau) = 1 + \epsilon(\tau), \tag{1.56}$$

where the dimensionless propagation distance, time and field envelope are defined as $\xi = z/L_{NL}$, $\tau = T/\sqrt{|\beta_2|L_{NL}/2}$, and $\psi(\xi, \tau) = A(z, T)/\sqrt{P_0}$, respectively, while $\epsilon(\tau)$ represents the imposed perturbation[14]. It is known that in the case of small periodic perturbations the solution of Eq. 1.56 can be written in an analytical form [36]. The situation turns out to be more complicated when $\epsilon(\tau)$ is considered to consist of localized perturbations [45] or random noise [46]. The appearance of the random localized nonlinear structures can be studied by introducing the concept of the solitonic gas [47] and by analyzing the statistics of extreme events [44, 48].

[13] This process is refered to as the degenerate FWM under the undepleted pump approximation.

[14] The normalisation in Eq. 1.55 is slightly different from that commonly appearing in the literature which uses a factor 1/2 before the second-order derivative term (see Eq. 2.1). For consistency with the literature and our own publications, we use the form in Eq. 1.55 here and in Chapters 3 and 4, and the normalisation in Eq 2.1 in Chapter 2 in the context of FWM dynamics.

In optical fibre systems MI occurs even in the presence of an uncorrelated random noise process. For the given initial conditions (see Eq. 1.56) the noise-driven MI governed by Eq. 1.55 can be simulated numerically via split-step Fourier method and using the one-photon-per-mode noise model to mimic the experimental conditions [49].

1.3.4/ Solitons on finite background

The NLSE possesses many analytical solution associated with MI dynamics [50]. Among them is the family of solutions known as the solitons on finite background (SFB). These solutions include the Akhmediev breather (AB), the Peregrine soliton (PS) [34] and Kuznetsov-Ma soliton (KM) [35] and can be written in the analytical form

$$\psi(\xi,\tau) = \left[1 + \frac{2(1-2a)\cosh(b\xi) + ib\sinh(b\xi)}{\sqrt{2a}\cos(\omega_m\tau) - \cosh(b\xi)}\right]\exp(i\xi), \tag{1.57}$$

where the physical behaviour of the solution is governed by a single parameter a through the arguments $b = [8a(1-2a)]^{1/2}$ and $\omega_m = [2(1-2a)]^{1/2}$. For example, with $a = 1/2$, $\omega_m = b = 0$ and we get a rational solution, known as the PS, doubly localised in ξ and τ [34]. For $a < 1/2$, ω_m and b are real-valued, and we obtain the τ-periodic AB solution, with ω_m and b taking the physical meaning of modulation frequency and exponential growth rate, respectively. Finally, when $a > 1/2$, ω_m and b become imaginary and we obtain the ξ-periodic KM solution. Figure 1.2 illustrates spatio-temporal properties of SFB solutions.

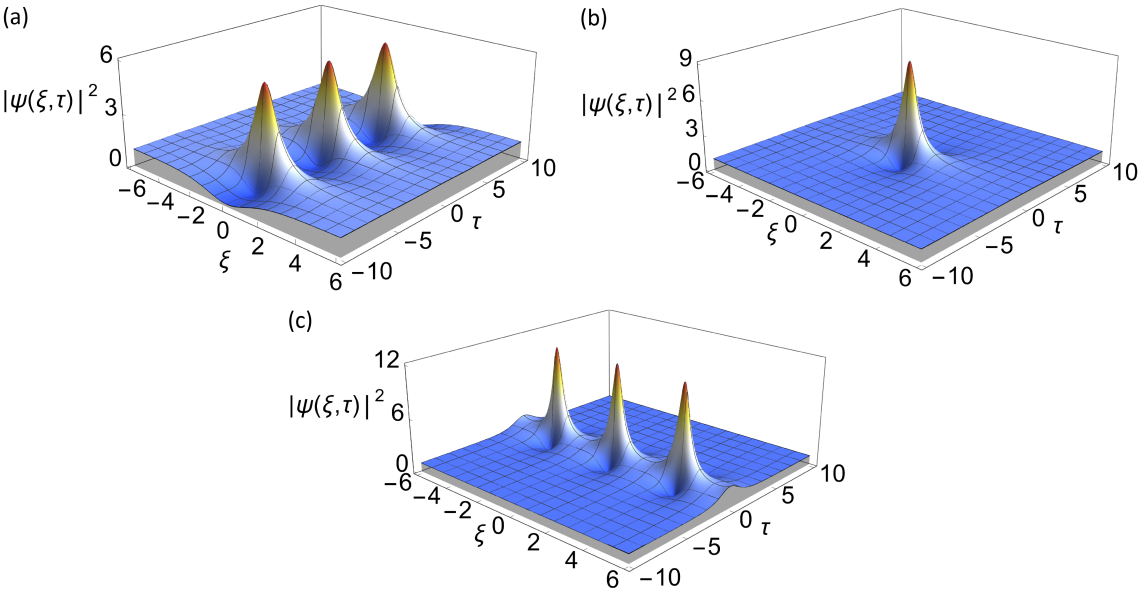


Figure 1.2: Spatio-temporal properties of solitons on finite background (see Eq. 1.57). **(a)** Akhmediev breather (AB). **(b)** Peregrine soliton (PS). **(c)** Kuznetsov-Ma soliton (KM).

These various SFB structures can be excited in a controlled way by coherent modulation of the input signal (e.g. a modulated continuous wave) and they can also emerge spontaneously during MI driven by random noise (technical or quantum noise). Many experiments since 2010 have observed a wide class of these SFB structures [51–53].

1.4/ Generalised nonlinear Schrödinger equation

As can be seen from the previous sections, the NLSE in its classical form (see Eq. 1.32) can be used to describe the fundamental nonlinear and dispersive effects arising in optical fibre propagation. The major nonlinear fibre optics phenomena discussed earlier in this chapter certainly do not exhaust the variety of NLSE propagation scenarios. Additional examples of normal dispersion propagation, including the optical wave breaking and the formation of Riemann wave shocks, will be provided later in Chapter 4.

The derivation of Eq. 1.32 presented in Section 1.1 was based on several simplifying assumptions allowing to neglect the higher-order dispersion and nonlinear contributions in the propagation equation model. These assumptions, however, lead to notable discrepancies between numerical simulations and experiments for optical pulses as short as $T_0 < 5$ ps [22, 54]. Therefore, the higher-order dispersive and nonlinear effects have to be taken into account when studying the ultrashort pulse propagation. In such regimes, the temporal evolution of $A(z,T)$ can be modeled using the generalised NLSE [55]

$$i\frac{\partial A}{\partial z}+i\frac{\alpha}{2}A+\sum_{m\geq 2}\frac{i^m}{m!}\beta_m\frac{\partial^m A}{\partial T^m}+\gamma\left(1+i\frac{\gamma_1}{\gamma}\frac{\partial}{\partial T}\right)\left[A(z,T)\int_{-\infty}^{\infty}dT'R(T')|A(z,T-T')|^2\right]=0, \tag{1.58}$$

that additionally takes into account the effects of fibre losses through the linear loss parameter α, higher-order dispersion terms (with m denoting the order of the Taylor series expansion of the chromatic dispersion), the self-steepening effect (proportional to the parameter γ_1/γ)[15] arising due to the dispersion of the nonlinear coefficient (see Eq. 1.26), and finally, the instantaneous and delayed Raman contributions. Under certain conditions, the combined action of dispersive and nonlinear effects inherent in the generalised NLSE can lead to the supercontinuum (SC) generation process, which consists in the extreme spectral broadening of the input narrow laser spectral profile [22, 54]. The SC has found many applications in science and engineering and is still an area of extensive research and development [54].

Let us now consider the structure of the generalised NLSE in more detail. The second term in Eq. 1.58 (responsible for the losses in a fibre) has to be included to mimic experimental conditions, although its physical effect on ultrashort pulse propagation is trivial in most cases[16]. In contrast, higher-order dispersion effects (third term in Eq. 1.58) now begin to play an important role in ultrafast dynamics even away from zero-dispersion wavelengths. The isolated action of the TOD on the soliton propagation can be studied via perturbation theory and is known to lead to a linear shift of the soliton peak $\delta T(z)$ with respect to z [22]

$$\delta T(z)=\left(\beta_3/6T_0^2\right)z=\delta_3\left(z/L_D\right). \tag{1.59}$$

Another important perturbation theory result is that higher-order dispersion effects lead to the emission of nonsolitonic radiation (analogous to Cherenkov radiation in particle physics) [56]. This phenomenon results in the energy transfer from solitons to narrowband resonances, the position of which can be obtained from phase-matching conditions. This effect becomes particularly noticeable in the presence of ejected Raman solitons (see Sections 1.4.1-1.4.2), whose perturbation leads to the efficient generation of dispersive waves with shorter wavelength (with respect to that of the solitons) [57].

[15] To leading order parameter γ_1/γ can be approximated by ω_0^{-1}, where $\gamma=\gamma(\omega_0)$ and $\gamma_1=(d\gamma/d\omega)_{\omega=\omega_0}$

[16] Note also that for short optical fibres this effect is usually neglected.

1.4.1/ Raman scattering

The integral term in Eq. 1.58 accounts for the Raman effect that is known to result in the energy transfer from high to low frequency components of the optical pulse (intrapulse Raman scattering) [22]. This particular higher-order nonlinear effect turns out to be the most crucial in the ultrafast soliton dynamics and SC generation in optical fibres.

The Raman effect in optical fibres originates from the scattering of photons on the vibrational modes of silica induced by the optical field [22]. Both the instantaneous (electronic) and delayed (nuclear) Raman contributions can be taken into account by introducing the following analytical model of the full Raman response function [58]

$$R(T) = (1 - f_R)\delta(T) + f_R h_R(T) = (1 - f_R)\delta(T) + f_R \frac{\tau_1^2 + \tau_2^2}{\tau_1 \tau_2^2} \exp(-T/\tau_2) \sin(T/\tau_1)\theta(T), \tag{1.60}$$

where $\theta(T)$ is the Heaviside step function, f_R is called the Raman fraction, while τ_1 and τ_2 are the fit parameters. The particular form of temporal Raman response function $h_R(T)$ was inspired by the problem of damped harmonic oscillator[17] and is shown in Fig. 1.3.

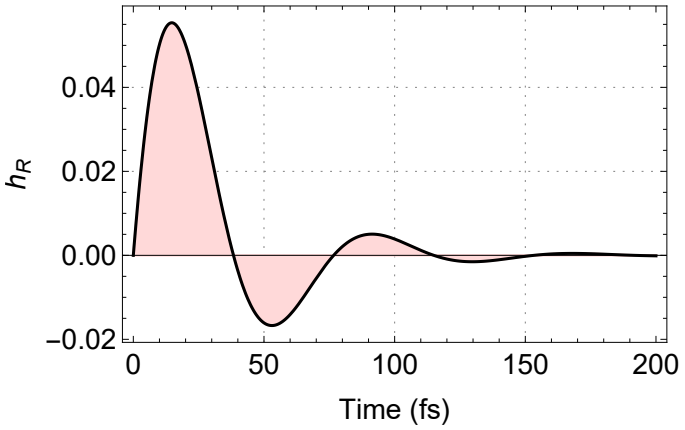


Figure 1.3: Analytical approximation of the temporal Raman response function $h_R(T)$ (see Eq. 1.60) plotted for $\tau_1 = 12.2$ fs and $\tau_1 = 16$ fs.

Typical values of the parameters τ_1, τ_2 and f_R for silica can be estimated from experiments and are 12.2 fs, 16 fs and 0.18, respectively. In practice, it is common to use the experimentally obtained Raman response function to more accurately account for the Raman contribution [59], which, however, does not qualitatively change the observed physics.

The isolated effect of Raman scattering was as well studied via perturbation theory approaches [22, 60]. It was found that it leads to a large temporal shift of the soliton peak, the appearance of small-amplitude temporal oscillatory tails, and a shift of the central frequency of the pulse towards longer wavelengths. This soliton self-frequency shift ω_R is the most noticeable effect of the delayed Raman contribution and can be expressed as $d\omega_R/dz \propto |\beta_2|/T_0^4$ [61].

Finally, it should be noted that the effect of self-steepening on the propagation of higher-order solitons is similar to that of intrapulse Raman scattering, although the latter tends to dominate the dynamics in most propagation scenarios.

[17] Note that the functional form of h_R is defined in such a way that $\int_{-\infty}^{\infty} dT h_R(T) = 1$.

1.4.2/ Soliton fission

Even when acting separately, the Raman effect, higher-order dispersion, as well as self-steepening can substantially perturb the higher-order soliton propagation and lead to its compression and to the phenomenon called the soliton fission [22]. Soliton fission consists in the breakup of the higher-order soliton into a series of fundamental solitons and plays the major role in the development of the SC from ultrashort input pulses [54].

The breakup of an input higher-order soliton (of order $N > 1$) into fundamental solitons begins after the fission distance L_{fiss} that corresponds to the maximum compression point of the injected soliton [62][18]. The explicit expressions for the peak powers and widths of the fundamental solitons of the form $A_j(z,T) = \sqrt{P_j}\,\text{sech}(T/T_j)$ ejected during the fission process can be found using the inverse scattering transform method [63]

$$P_j = P_0\frac{(2N-2j+1)^2}{N^2},\;\; T_j = \frac{T_0}{2N-2j+1}, \tag{1.61}$$

where $j = 1,\ldots,N$ and N represents the soliton number of the input soliton. The soliton fission process induced by the TOD and in the presence of TOD and Raman effect will be discussed in detail in Chapter 4.

The later propagation stage is associated with further spectral broadening emerging from the interaction of leading dispersive and nonlinear processes that result in rich and complex dynamics. Unlike the NLSE, Eq. 1.58 is generally non-integrable, thus numerical simulations have to be performed to study the SC generation process. The numerical integration scheme in this case can be as well implemented in the framework of the split-step Fourier method [22]. Specifically, the solution of the linear propagation step can be performed in the frequency domain, while the nonlinear step can be performed using the Runge-Kutta integration rule [54, 55, 57].

To accurately model the SC generation process, it is necessary to account for all terms in Eq. 1.58 with higher order dispersion expanded up to 12^{th} order in some cases. At the same time, the physics underlying the development of SC from ultrashort pulses can be well interpreted when considering a simplified form of the generalised NLSE. In particular, we write here the generalised NLSE in the dimensionless form that additionally accounts for the TOD and Raman effects

$$i\frac{\partial\psi}{\partial\xi} + \frac{\partial^2\psi}{\partial\tau^2} + i\delta\frac{\partial^3\psi}{\partial\tau^3} + |\psi|^2\psi + \rho\psi\,(h_R * |\psi|^2) = 0, \tag{1.62}$$

where the dimensionless variables correspond to $\psi = A\sqrt{\gamma L_{\text{D}}}$, $\xi = z(1-f_{\text{R}})/L_D$, $\tau = T\sqrt{2(1-f_{\text{R}})}/T_0$. The TOD and Raman effects are included here through the dimensionless parameters $\delta = \sqrt{2(1-f_R)}\beta_3/3T_0|\beta_2|$ and $\rho = f_{\text{R}}/(1-f_{\text{R}})$, the operator $(*)$ stands for the convolution. This particular dimensionless form of the generalised NLSE will be used in the subsequent analysis in Chapter 4.

[18] The fission distance can be well approximated as $L_{\text{fiss}} \approx L_D/N$.

1.5/ Machine learning methods in fibre optics: Overview

As was noted in the general introduction to this thesis, nowadays, methods and tools of ML have a tremendous impact on how we conduct research in various fields of science [2]. The rapid growth of the scientific community's interest in ML approaches in recent decades is driven by the power of ML to perform classification, pattern recognition, optimisation and prediction of the complex system dynamics, as well as many other tasks that are at the heart of the scientific method [3, 4]. These developments have quickly penetrated the fields of optics and photonics, where ML algorithms have found practical applications in the design of optical components and new photonic materials [11], in imaging and sensing applications, to name only a few [13, 64].

A particularly intriguing direction for the application of ML approaches has been found in the field of ultrafast photonics, where optical systems require systematic control and optimisation [14]. One of the major directions in this domain is the optimization of ultrafast fibre lasers, where the dynamics is governed by a complex interplay between nonlinear, dispersive and dissipative processes [65]. The need for ML algorithms arises naturally due to the number of control parameters (degrees of freedom) in such systems that have to be optimised simultaneously to reach the desired dynamical regime [66], as well as the sensitivity of the optical circuits to noise [67]. In a particular example of fibre lasers, one can reach qualitatively different dynamical regimes (e.g. continuous wave, mode-locking, noise-like pulse generation etc.) by controlling the polarization, pump power, spectral filters and losses [68,69]. In this case, conventional greedy search algorithms over the available system parameters become computationally expensive for optimization purposes. Contrary to that, ML search algorithms, including genetic and evolutionary algorithms, turn out to be an efficient solution to establish the desired operational regime of ultrafast optical fibre systems [70–72]. In particular, the real-time optimisation and self-tuning of the fibre laser can be implemented using a feedback loop via a genetic algorithm [73], which results in a remarkable improvement over linear parameter tuning.

Another novel application of ML in the field of ultrafast nonlinear fibre optics is the prediction of nonlinear dynamics and pattern identification from noisy and incomplete measurements. It was shown that the use of neural networks can enable us to find a link between the local intensity maxima of the chaotic temporal field and features of the noisy broadband spectra [74], a fact that immediately found applications in the analysis of noisy SC generation [75] and noise-driven MI [76]. More recently it was demonstrated how the pre-trained feed-forward neural networks can be efficiently used to speed-up the integration of the generalised NLSE to simulate the SC generation process [77].

The application of search algorithms and neural networks to optimise, control and predict the ultrafast dynamics of optical fibres has resulted in significant progress in the field [14]. At the same time, the model-free approaches mentioned above usually provide us with black-box type models and solutions that are mostly applicable over the range of parameters used in the training process, which significantly limits their generalisability. In contrast, several qualitatively different ML approaches have recently been proposed to discover underlying physical principles and hidden low-dimensional dynamical patterns directly from data [16, 17, 78–81]. In particular, these novel data-driven methods are able to automatically discover governing differential equation models, identify dominant physical processes, and perform model-based prediction and control of the state of physical systems by analysing dynamical data (e.g., spatio-temporal evolution) [8].

1.5.1/ State-of-the-art data-driven methods in fibre optics

We would now like to focus on data-driven methods that are designed to aid in the physical interpretation of dynamics and even potentially make data-driven discoveries. A comprehensive overview of data-driven methods for science and engineering, as well as practical implementation examples, are presented in the book "Data-driven science & engineering" by Steven L. Brunton and J. Nathan Kutz [82], and we describe specific details of techniques relevant to this thesis in the following chapters. Currently, the applications of such methods in the field of nonlinear fibre optics are rather scarce, in contrast to the model-free approaches discussed in Section 1.5. Nevertheless, it is conceivable that these novel data-driven techniques and algorithms will soon become one of the central methodologies for optimisation, control and discovery in complex fibre optics systems [14].

The goal of this branch of ML approaches can be either to extract interpretable and generalisable models from data (e.g., differential equation models) or to develop data-driven algorithms that can complement conventional theoretical analysis [82]. In ML, it is common to distinguish two main classes of algorithms known as supervised and unsupervised learning [3]. Supervised learning algorithms require the labeled data input. This implies that examples of the desired model output for a particular input are already given in the training dataset. In unsupervised learning algorithms, no training labels are given, so the algorithm itself must find patterns in the data to determine how to cluster and classify new data. Supervised learning is usually associated with regression-like problems, while unsupervised learning is associated with the clustering and dimensionality reduction techniques.

Figure 1.4 gives an overview of several supervised and unsupervised data-driven techniques that have already been applied to the field of nonlinear fibre optics.

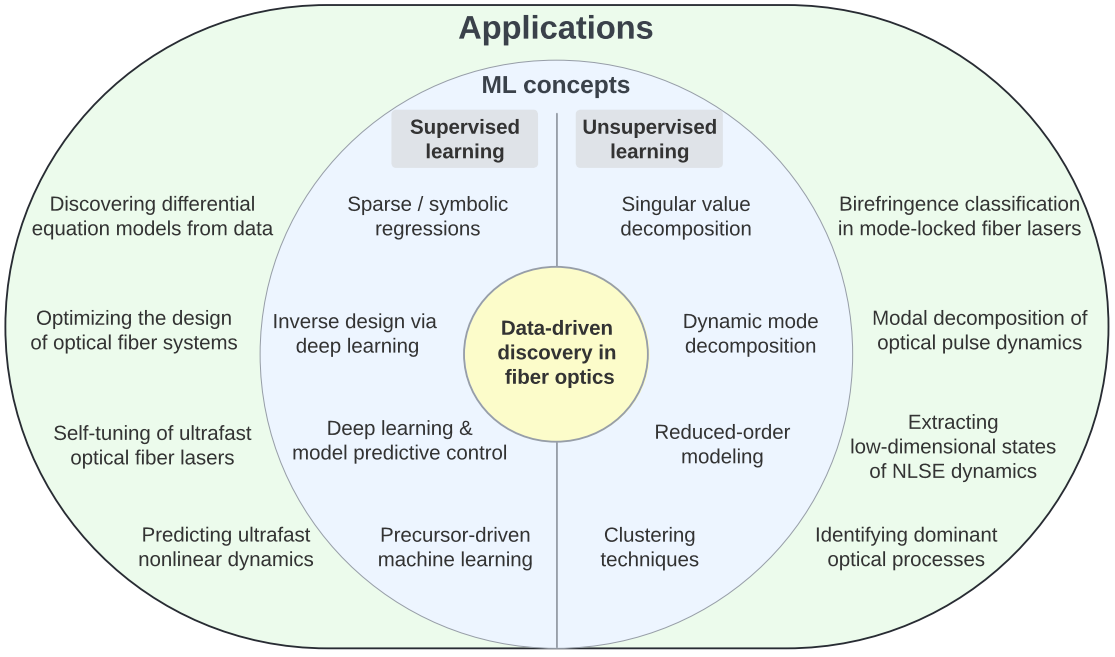


Figure 1.4: Several supervised and unsupervised data-driven methods and their applications in fibre optics.

The emerging applications of supervised data-driven techniques in the field of nonlinear fibre optics enable the self-tuning of ultrafast fibre lasers via deep learning with model predictive control [83]. This provides a data-driven alternative to the optimization methods for fibre lasers discussed in Section 1.5. The ability of this approach to robustly predict the birefringence and future state of the laser allows the high performance of a mode-locked fibre laser to be maintained even in the presence of chaotic perturbations in the cavity. Another data-driven strategy aiming at prediction of chaotic extreme pulses in a Kerr resonator based on the precursor-driven ML and reservoir computing has recently been demonstrated [84, 85].

One of the important practical applications of novel data-driven techniques consists in the inverse design of various optical fibre systems. Even though the concept of inverse design is not very recent [86] and has been extensively explored in photonics applications [87], it has been only recently shown to be enhanced via deep-learning and specifically tailored for the optical fibre systems [88–90].

While supervised data-driven approaches are proving to be very successful in a variety of practical applications, the generalisability of unsupervised techniques can be very useful in the exploration of the physical principles underlying the dynamics. For example, dynamics-mode decomposition [79] and reduced-order modeling [82] can be exploited to retrieve low-dimensional modes of NLSE dynamics in different propagation scenarios [91,92]. Singular value decomposition combined with the sparse sensing turns out to be valuable in classification of birefringence in mode-locked fibre lasers [67].

Recently, a novel approach has been proposed to identify local dominant physical processes directly from dynamical data in the framework of unsupervised clustering [17, 18, 93–95]. The ability to automatically separate regions of spatio-temporal field evolution where certain dispersive and nonlinear effects dominate the propagation is crucial to understanding complex NLSE dynamics and can accomplish conventional theoretical methods. This particular data-driven approach called the data-driven dominant balance [17] will be one of the central topics of this thesis. In Chapters 3 and 4 for the first time we apply this technique to various nonlinear fibre optics propagation scenarios, including noise-driven MI, optical wave breaking and soliton fission induced by TOD and Raman effects. We provide a detailed algorithmic framework, physical interpretation of the method, and discuss its possible future applications.

An even more general and ambitious approach in data-driven science is the discovery of differential equation models from data [16, 78, 96–98]. The capability to extract interpretable dynamical models from numerical data and experimental measurements is at the forefront of modern science and has the potential to impact various fields also outside of physics, including neuroscience, biology, and geosciences [82]. In this regard, the developed numerical methods and the availability of experimental data in nonlinear fibre optics make this field an ideal testbed for this novel approach. Moreover, data-driven discovery of interpretable and generalisable models can be very promising in certain regimes of nonlinear fibre optics propagation where the governing differential equations are only partially known. In this thesis we focus on one of the recently proposed technique called the sparse identification of nonlinear dynamics (SINDy) [16]. The first application of SINDy in the field of nonlinear fibre optics is presented in Chapter 2, where we demonstrate how the differential equation model of the FWM dynamics can be directly extracted from dynamical data.

2

Data-driven discovery of four-wave mixing dynamics in nonlinear fibre optics

2.1/ Introduction

As highlighted in Section 1.5 of Chapter 1, machine learning (ML) techniques based on statistical approaches and numerical algorithms have proven to be a powerful tool for analysis and prediction of dynamics of complex systems in various areas of science. From a fundamental point of view, it is of particular interest to use data-driven approaches to study nonlinear systems when the structure of the underlying differential equation models can be partially or even completely unknown. Specifically, a number of inverse-problem-like algorithms have recently been developed to identify the underlying mathematical structure of differential equation models governing the system dynamics based only on analysis of data generated by the system [16, 78, 80, 99, 100]. The particular approach known as the sparse identification of nonlinear dynamics (SINDy) developed by Steven L. Brunton, Joshua L. Proctor and J. Nathan Kutz has already found extensive use in various fields, including mechanics, hydrodynamics and plasma physics [16]. Its implementations in optics, however, have been more limited, but it has found applications in the field of telecommunication networks [101].

The SINDy algorithm is based on the well-known observation that the dynamics of many physical systems can be described by a set of coupled differential equations, with the right-hand side (RHS) part consisting of only a few terms. This sparsity of the RHS in the equation space allows SINDy to discover forms of the governing differential equations by analyzing data series. Built on various numerical strategies, SINDy provides an algorithmic approach to determine the active RHS terms from a wide library of potential candidate functions for both systems of ordinary and partial differential equations [96, 102–104]. The feasibility of SINDy was demonstrated on several canonical physical models manifesting nonlinear and chaotic dynamics, including the Lorenz system, the case of a fluid wake behind a cylinder and a nonlinear pendulum [16].

In this chapter, the SINDy technique is applied to the canonical problem of optical four wave mixing (FWM) in the nonlinear Schrödinger equation (NLSE) system [22, 105–107]. These dynamics can be described by the coupled differential equations for the relative amplitude and phase of frequency components evolving with distance in an optical fibre [108, 109]. Being the dominant process in various physical systems, including cold

atoms, plasma physics and hydrodynamics [50, 110–112] FWM is one of the key physical effects that can be used to assess the validity of the SINDy algorithm. It is also of interest to test the performance of SINDy in the case of noisy data, which is an important step towards the identification of differential equation models from experimental data [103,113].

2.2/ Degenerate four wave mixing in optical fibres

In this section, we present two different formalisms to describe degenerate FWM dynamics in an optical fibre: the coupled amplitude equations and the Hamiltonian model. The derived models will be used in further analysis of the FWM dynamics via SINDy algorithm. In addition, we discuss in detail the phase space representation of the FWM dynamics.

2.2.1/ Hamiltonian formalism

The degenerate FWM in optical fibres was briefly discussed in Section 1.3.2, where it was linked to modulation instability (MI) dynamics. A starting point in the description of FWM dynamics is the NLSE that we write here in the following dimensionless form

$$i\frac{\partial A}{\partial \xi} + \frac{1}{2}\frac{\partial^2 A}{\partial \tau^2} + |A|^2 A = 0, \tag{2.1}$$

where $\xi = z/L_{NL}$, $\tau = T/\sqrt{|\beta_2|L_{NL}}$, and $A(\xi,\tau) = A(z,T)/\sqrt{P_0}$[1]. Degenerate FWM describes the interaction of three optical components with distinct frequencies and can be studied by inserting the following ansatz into Eq. 2.1

$$A(\xi,\tau) = A_0(\xi) + A_1(\xi)\exp(i\Omega\tau) + A_{-1}(\xi)\exp(-i\Omega\tau), \tag{2.2}$$

where $A_0(\xi)$ is the field at the pump frequency and $A_{\pm1}(\xi)$ are frequency sidebands detuned by $\pm\Omega$. This yields the following set of coupled complex amplitude equations that describe the energy exchange between three discrete frequency components[2]

$$-i\frac{dA_0}{d\xi} = \left(|A_0|^2 + 2\Big[|A_{-1}|^2 + |A_1|^2\Big]\right)A_0 + 2A_{-1}A_1A_0^*, \tag{2.3a}$$

$$-i\frac{dA_{-1}}{d\xi} = \left(-\frac{1}{2}\Omega^2 + |A_{-1}|^2 + 2\Big[|A_0|^2 + |A_1|^2\Big]\right)A_{-1} + A_1^*A_0^2, \tag{2.3b}$$

$$-i\frac{dA_1}{d\xi} = \left(-\frac{1}{2}\Omega^2 + |A_1|^2 + 2\Big[|A_{-1}|^2 + |A_0|^2\Big]\right)A_1 + A_{-1}^*A_0^2, \tag{2.3c}$$

Equations 2.3 describe the dynamical energy exchange between the pump A_0 and two equidistant sidebands $A_{\pm1}$. At the same time, solving the NLSE numerically with the initial conditions of this form (see Eq. 2.2) results in cascaded FWM process, where additional spectral lines are generated at the expense of the pump [114]. The full mathematical description of the cascaded FWM turns out to be rather cumbersome, but restricting the analysis to only three interacting spectral components results in rather simple "ideal FWM" model that captures the essential physics of the system.

[1] This particular dimensionless form of the NLSE is used throughout this chapter.

[2] The normalised frequency used in this chapter is defined as $\Omega = \omega\sqrt{|\beta_2|L_{NL}}$.

The set of governing Eqs. 2.3 can be simplified by describing the FWM dynamics using the Hamiltonian formalism. As was demonstrated in [108,109], the Hamiltonian of Eq. 2.3 can be written in terms of two real-valued physical variables, namely, the relative sideband intensity η and phase ϕ

$$H(\eta,\phi) = 2\eta(1-\eta)\cos\phi + (\Omega^2+1)\eta - \frac{3}{2}\eta^2, \tag{2.4}$$

where

$$\eta = \frac{|A_0(\xi)|^2}{P_0} = \frac{|A_0(\xi)|^2}{|A_{-1}(\xi)|^2 + |A_0(\xi)|^2 + |A_1(\xi)|^2}, \tag{2.5a}$$

$$\phi = \arg\left[A_{-1}(\xi)\right] - 2\arg\left[A_0(\xi)\right] + \arg\left[A_1(\xi)\right]. \tag{2.5b}$$

Using the relations $d\eta/d\xi = \partial H/\partial\phi$ and $d\phi/d\xi = -\partial H/\partial\eta$, the dynamics of degenerate FWM can be described by the set of only two self-consistent equations

$$\frac{d\eta}{d\xi} = -2\eta\left(\left[1-\eta\right]^2 - \kappa^2\right)^{1/2}\sin\phi, \tag{2.6a}$$

$$\frac{d\phi}{d\xi} = -\left(\Omega^2+1\right) + 3\eta - 2\frac{1+2\eta^2-3\eta-\kappa^2}{\left(\left[1-\eta\right]^2-\kappa^2\right)^{1/2}}\cos\phi, \tag{2.6b}$$

where $\kappa = \kappa(\xi) = (|A_{-1}(\xi)|^2 - |A_1(\xi)|^2)/P_0$ describes the asymmetry between the sidebands. Variable κ along with $H(\eta,\phi)$ are known to be the system's invariants implying that these quantities are conserved throughout the propagation. Assuming the case of the equal sidebands ($\kappa = 0$) one can further simplify the set of governing equations

$$\frac{d\eta}{d\xi} = 2\eta\left(\eta-1\right)\sin\phi, \tag{2.7a}$$

$$\frac{d\phi}{d\xi} = -\left(\Omega^2+1\right) - 2\cos\phi + 3\eta + 4\eta\cos\phi. \tag{2.7b}$$

This is equivalent to setting $A_{-1}(\xi) = A_1(\xi)$ in Eq. 2.3 resulting in

$$-i\frac{dA_0}{d\xi} = \left(|A_0|^2 + 4|A_1|^2\right)A_0 + 2A_1^2A_0^*, \tag{2.8a}$$

$$-i\frac{dA_1}{d\xi} = \left(-\frac{1}{2}\Omega^2 + 3|A_1|^2 + 2|A_0|^2\right)A_1 + A_1^*A_0^2. \tag{2.8b}$$

An advantage of the Hamiltonian formulation presented here is that it better captures the physical aspects of the degenerate FWM dynamics when compared to the NLSE and the coupled amplitude equations. The Hamiltonian model can provide additional information about the separation of the dynamics into different regimes and can help derive simple analytical expressions for the initial conditions leading to stationary solutions [108, 109]. Moreover, it can be quite helpful in any attempts to observe the well-known Fermi-Pasta-Ulam-Tsingou periodic recurrence [21,115] under laboratory conditions [116–118]. Indeed, another advantage of Eqs. 2.7 is that the physical variables η and ϕ can readily be measured experimentally [119]. The state-of-the-art experimental setups enabling the observation of ideal FWM dynamics and the Fermi-Pasta-Ulam-Tsingou recurrence in the MI are presented in the PhD theses of Guillaume Vanderhegen [120] and Anastasiia Sheveleva [121]. We refer readers to these theses also for an extensive literature review covering both theoretical and experimental works on the topic.

2.2.2/ Phase space representation of dynamics

Equations 2.7 can be solved numerically for given initial conditions (η_0, ϕ_0) and sideband detuning Ω. In the case, when $|A_{\pm 1}| \ll |A_0|$ the phase-matching condition corresponding to the maximum FWM gain is $\Omega = \sqrt{2}$. This value is used in the following simulations in this chapter. Figure 2.1 shows the intensity evolution of $|A(\xi, \tau)|^2$ for two different sets of initial conditions: (a) $\eta_0 = 0.95$ and $\phi_0 = 0$, (b) $\eta_0 = 0.95$ and $\phi_0 = \pi$. It can be seen that the interaction of three optical field components associated with the degenerate FWM results in the existence of two different dynamical regimes [see Figs. 2.1(a) and (b)].

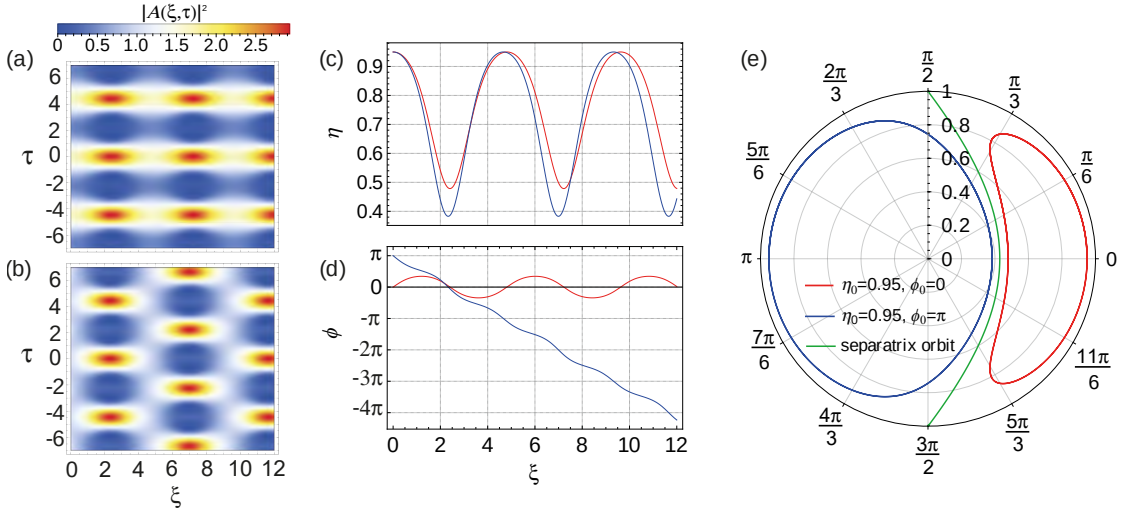


Figure 2.1: Spatio-temporal dynamics of degenerate FWM represented by $|A(\xi, \tau)|^2$ for the initial relative amplitude $\eta_0 = 0.95$ and different initial relative phases: **(a)** $\phi_0 = 0$, **(b)** $\phi_0 = \pi$. The influence of the initial phase shift results in the transverse phase shift apparent in **(b)**. **(c)** and **(d)** show the corresponding dynamics of $\eta(\xi)$ and $\phi(\xi)$ as the functions of propagation distance ξ. **(e)** displays the associated phase space dynamics. Red curves in subfigures **(c-e)** show the dynamics associated with initial condition $\phi_0 = 0$ [subfigure **(a)**] and the blue curves with initial condition $\phi_0 = \pi$ [subfigure **(b)**]. The green curve in **(e)** represents the trajectory of the system's separatrix (see text). The vertical axis in **(e)** shows the range of $\eta \in [0, 1]$, while angles ϕ are shown around the circle.

In both situations we can observe the periodic spatio-temporal evolution, however the initial conditions corresponding to the π phase shift between the pump and the sidebands [Fig. 2.1(b)] result in the transverse phase shift in the recurrence pattern. Figures 2.1(c) and (d) show the spatial evolution of the relative intensity η and phase ϕ for two sets of initial conditions (red curves: $\eta_0 = 0.95$ and $\phi_0 = 0$, blue curves: $\eta_0 = 0.95$ and $\phi_0 = \pi$). This clearly illustrates the periodic energy exchange between the pump and the sidebands occurring during the propagation of the optical field.

Using the same color code, Fig. 2.1(e) shows the corresponding phase space portraits of the two trajectories in the $\eta(\xi) - \phi(\xi)$ polar coordinates. Note that this is fully consistent with the phase space representation of the dynamics used in various fields of physics. Recurrence patterns become apparent here as the FWM dynamics forms closed trajectories in the associated phase space. In terms of the two new variables $\eta(\xi)$ and $\phi(\xi)$, the dynamics is analogous to a one-dimensional nonlinear oscillator, and its phase space is

divided into two dynamical regions via separatrix - the homoclinic trajectory obtained for $(\eta_0, \phi_0) \rightarrow (1, \pm\pi/2)$ [108] [green curve in Fig. 2.1(e)].

To better illustrate distinct dynamical regimes arising in the FWM dynamics, Fig. 2.2(a) shows phase space trajectories corresponding to 6 different sets of initial conditions. The trajectories on the RHS of the separatrix were obtained by numerically solving Eqs. 2.7 for $\eta_0 = 0.8, 0.9, 0.95$ with $\phi = 0$ (red curves). It can be seen how the periodic trajectories approach the separatrix orbit as the initial relative amplitude η_0 increases. The trajectories on the left-hand side (LHS) of the separatrix were obtained for $\eta_0 = 0.4, 0.8, 0.95$ with $\phi = \pi$ (blue curves). Two ensembles of the phase space trajectories show quite different spatio-temporal dynamics, as we highlighted in Figs. 2.1(a) and (b). Specifically, the LHS trajectories show the transverse phase shift in the recurrence pattern.

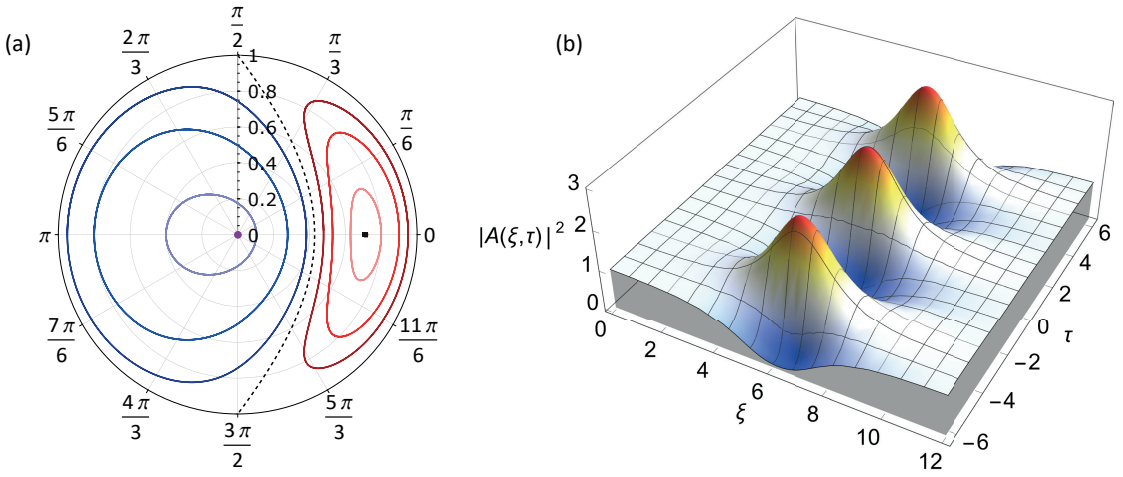


Figure 2.2: **(a)** Phase space trajectories of the ideal FWM dynamics obtained for different initial conditions. The degenerate FWM process described by the Hamiltonian model leads to spatio-temporal dynamics demonstrating the well-known Fermi-Pasta-Ulam-Tsingou recurrence patterns. Red lines represent the RHS trajectories obtained for initial conditions $\eta_0 = 0.8, 0.9, 0.95$ with $\phi = 0$. Blue lines represent the LHS trajectories obtained for initial conditions $\eta_0 = 0.4, 0.8, 0.95$ with $\phi = \pi$. The black rectangular and purple circular points indicate the phase space position of the stationary solution of the Hamiltonian system for $\Omega = \sqrt{2}$ (see text). The black-dashed line shows the separatrix orbit. **(b)** Spatio-temporal dynamics of the separatrix represented by $|A(\xi, \tau)|^2$. The dynamics were obtained for the initial relative amplitude $\eta_0 = 1 - \epsilon$ with $\epsilon = 10^{-5}$ and initial relative phase $\phi_0 = \pi/2$.

Another important dynamical aspect is the existence of stationary solutions in the degenerate FWM model, which can also be seen in Fig. 2.2(a). The black rectangular point on the RHS corresponds to one of the eigensolutions of the Hamiltonian system (see Eq. 2.7). In particular, for $\Omega = \sqrt{2}$ considered here, the initial conditions $\eta_e = (3+\Omega^2)/7 = 5/7$ and $\phi_e = 0$ lead to the stationary solution of the system, where the dynamics do not experience any spatial dependence. Thus, the corresponding phase space dynamics are represented by the fixed point on the RHS part with respect to the separatrix. The fixed point at the origin of the phase space (highlighted in purple) corresponds to another eigensolution of the Hamiltonian system with $\eta_e = 0$ [109]. Finally, Fig. 2.2(b) shows the spatio-temporal dynamics of the separatrix obtained for $(\eta_0, \phi_0) = (1 - \epsilon, \pm\pi/2)$ with $\epsilon = 10^{-5}$ [108]. In fact, in the more general case of NLSE dynamics, the trajectory of the separatrix is related

to the Akhmediev breather (AB) solution [21].

The degenerate FWM process described by the Hamiltonian model leads to spatio-temporal dynamics demonstrating the well-known Fermi-Pasta-Ulam-Tsingou recurrence patterns. We can see from the above that these dynamics are very sensitive to the initial conditions and experience different physical regimes depending on the choice of initial conditions. In this regard, the degenerate FWM model can be an ideal testbed for the SINDy algorithm, which is specifically designed to study nonlinear systems. So far, we have only considered the ideal FWM dynamics in the absence of noise, which is usually present in experiments. Evaluating the validity of SINDy on noisy dynamical data is crucial for the further development of the algorithm and especially for its practical applications. The application of SINDy to noisy degenerate FWM dynamics is presented in Section 2.4.

2.3/ Methodology of SINDy

In this section, the methodology of SINDy is outlined [16]. In particular, we focus on the fundamentals of symbolic regression and sparse sensing and how they can be used to extract the form of governing differential equations from data. The construction of the candidate nonlinear functions, optimization methods and some aspects of overfitting of the returned differential equation models are also discussed.

2.3.1/ Extracting differential equation models from data

The starting point of SINDy is the fact that the dynamics of many physical systems can be described by a set of coupled differential equations of the form

$$\frac{d}{d\zeta}\mathbf{x}(\zeta) = \mathbf{f}\left[\mathbf{x}(\zeta)\right], \tag{2.9}$$

where $\mathbf{f}$ is the vector field function governing the evolution of the state vector $\mathbf{x} = [x_1(\zeta); x_2(\zeta); ...; x_n(\zeta)]$ with respect to the variable ζ. The state vector $\mathbf{x}$ thus contains n elements associated with physical variables involved in the dynamics (e.g., displacement, angle, temperature, amplitude, phase, etc.). The variable ζ is usually associated with time or a propagation coordinate, defined in relation to the considered physical problem.

In most cases, given that a form of the vector field function $\mathbf{f}$ is known, one can directly integrate Eqs. 2.9 to predict the evolution of the state vector $\mathbf{x}$. However, for many complex physical systems, the structure of $\mathbf{f}$ might be unknown (or partially unknown), so that one only has an access to observations of the evolution of physical variables obtained from experiments. These measured data representing, for example, the spatial (or temporal) evolution of the state vector can be written in the following matrix form

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}^T(\zeta_1) \\ \mathbf{x}^T(\zeta_2) \\ \vdots \\ \mathbf{x}^T(\zeta_m) \end{pmatrix} = \begin{pmatrix} x_1(\zeta_1) & x_2(\zeta_1) & \dots & x_n(\zeta_1) \\ x_1(\zeta_2) & x_2(\zeta_2) & \dots & x_n(\zeta_2) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(\zeta_m) & x_2(\zeta_m) & \dots & x_n(\zeta_m) \end{pmatrix}, \tag{2.10}$$

sampling n physical variables at m discrete values of ζ. The dimensionality n of the state vector is linked to the considered physical system and/or experimental conditions. In addition, the dataset may contain multiple matrices (see Eq. 2.10) corresponding to various initial conditions of a state vector $\mathbf{x}(\zeta = 0) = [x_1(\zeta = 0); x_2(\zeta = 0); ...; x_n(\zeta = 0)]$. From Eq. 2.10 we can directly estimate the matrix of derivatives by means of numerical differentiation

$$\dot{\mathbf{X}} = \begin{pmatrix} \dot{\mathbf{x}}^T(\zeta_1) \\ \dot{\mathbf{x}}^T(\zeta_2) \\ \vdots \\ \dot{\mathbf{x}}^T(\zeta_m) \end{pmatrix} = \begin{pmatrix} \dot{x}_1(\zeta_1) & \dot{x}_2(\zeta_1) & \dots & \dot{x}_n(\zeta_1) \\ \dot{x}_1(\zeta_2) & \dot{x}_2(\zeta_2) & \dots & \dot{x}_n(\zeta_2) \\ \vdots & \vdots & \ddots & \vdots \\ \dot{x}_1(\zeta_m) & \dot{x}_2(\zeta_m) & \dots & \dot{x}_n(\zeta_m) \end{pmatrix}, \tag{2.11}$$

where the dot stands for derivative with respect to ζ. Equation 2.11 thus represents the LHS of Eq. 2.9. The problem of extracting the form of differential equations from data consists in finding the best suitable RHS functions in Eq. 2.9 that acts on the state vector $\mathbf{x}$ to balance the equality.

2.3.2/ Library of candidate functions

To find the best suitable functional representation of $\mathbf{f}$, we can construct a large library of candidate functions $\boldsymbol{\Theta}$. Based on physical knowledge, among these candidates can be polynomial, periodic and any other possible functions of variables $x_1(\zeta)$, $x_2(\zeta)$, ..., $x_n(\zeta)$. Mathematically, a library of candidate functions can be represented in the matrix form $\boldsymbol{\Theta}(\mathbf{X}) = [\theta_1(\mathbf{X}), \theta_2(\mathbf{X}), \ldots, \theta_p(\mathbf{X})]$ with dimensions $m \times p$ with p being the total number of candidate RHS functions acting on the columns of $\mathbf{X}$. Then, the matrix of derivatives 2.11 can be estimated in the following way

$$\dot{\mathbf{X}} = \boldsymbol{\Theta}(\mathbf{X})\mathbf{C}, \tag{2.12}$$

where matrix $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_n]$ of dimensions $p \times n$ is a row vector of coefficients standing before the corresponding candidate terms in library $\boldsymbol{\Theta}$. Equation 2.12 introduces a regression problem that can be solved to find the best approximation of the matrix $\mathbf{C}$. After solving the regression problem for $\mathbf{C}$, one can directly construct the set of differential equations approximating the evolution of the physical variables

$$\dot{\mathbf{x}}_k(\zeta) = \mathbf{f}_k\big[\mathbf{x}(\zeta)\big] = \boldsymbol{\Theta}(\mathbf{x}^T)\mathbf{c}_k, \tag{2.13}$$

with $\boldsymbol{\Theta}(\mathbf{x}^T)$ denoting a vector of symbolic functions of the state vector elements. Equations 2.10-2.13 introduce a framework to extract the differential equation models from data via solving a regression problem. In [122] the power of a similar approach was demonstrated on the example of coupled Rössler oscillators. However, solving the conventional regression problem results in an output model that includes all the candidate terms in $\boldsymbol{\Theta}$ (the output coefficient matrix is dense) which limits the physical interpretability of the results. To overcome this limitation, the SINDy algorithm proposed in [16] utilizes a sparsity promoting criterion when solving the regression problem to return sparse and physically-interpretable models from data.

2.3.3/ Sparse regression problem

The motivation for formulating the search for differential equation models best representing the dynamics within the framework of the sparse regression problem is based on the well-known observation that many physical systems can be described by differential equations with a small number of RHS terms [16]. In other words, the vector field $\mathbf{f}$ in Eq. 2.9 is usually sparse in the equation space.

Therefore, SINDy searches for a solution to the regression problem (see Eq. 2.12) that provides a good fit to the input data while keeping the smallest number of terms in the RHS part. Algorithmically, a sparse approximation of $\mathbf{c}_k$ in Eq. 2.13 can be found by employing the least absolute shrinkage and selection operator (LASSO) regression [123]

$$\mathbf{c}_k = \arg\min_{\mathbf{c}'_k} \left\| \dot{\mathbf{X}}_{:k} - \boldsymbol{\Theta}(\mathbf{X})\mathbf{c}'_k \right\|_2 + \alpha \left\| \mathbf{c}'_k \right\|_1, \tag{2.14}$$

where $\dot{\mathbf{X}}_{:k}$ is the k-th column of $\dot{\mathbf{X}}$ and α denotes the LASSO l_1 hyperparameter[3]. The notation $\|.\|_2$ and $\|.\|_1$ stands for the l_2 and l_1 norms of a matrix, respectively.

[3]Note that hereafter the variable α denotes the LASSO l_1 hyperparameter (not to be confused with the linear loss coefficient introduced in Chapter 1).

2.3.4/ Optimization methods

In order to increase the robustness of the algorithm, the sequential thresholded least squares (STLS) algorithm was implemented in the original SINDy paper [16], and this was shown to outperform the conventional LASSO regression. Algorithm 1 shows the main steps of STLS optimization that uses a regression while iteratively discarding the terms with coefficients smaller than a given threshold λ_t.

Algorithm 1 Sequential thresholded least squares optimization for the SINDy algorithm

- Estimate the LHS matrix $\dot{\mathbf{X}}$
- Initialize candidate RHS functions $\boldsymbol{\Theta}$
- Initialize the hyperparameter α_2 and the threshold value λ_t
- Initialize a number of iterations $iters$ needed for convergence
- Allocate a matrix of coefficients $\mathbf{C}$

Perform an initial guess for $\mathbf{C}$ using the least squares approximation
for $i = 1, ..., iters$ **do**
1. Find indices $\mathbf{k}_i$ of coefficients in $\mathbf{C}$ above the threshold: $\mathbf{k}_i = (\text{abs}(\mathbf{C}) \geq \lambda_t)$
2. Threshold the remaining coefficients $\mathbf{C}[\sim \mathbf{k}_i] = 0$
3. Update $\mathbf{C}$ via regression $\mathbf{C}[\mathbf{k}_i] = \arg\min_{\mathbf{C}'} \left\| \dot{\mathbf{X}} - \boldsymbol{\Theta}(\mathbf{X})[\mathbf{k}_i]\mathbf{C}' \right\|_2 + \alpha_2 \left\| \mathbf{C}' \right\|_2$

end for
Return $\mathbf{C}$

To perform the initial guess for coefficient matrix $\mathbf{C}$, one can use the least squares solution

$$\mathbf{C} = (\boldsymbol{\Theta}^T\boldsymbol{\Theta})^{-1}\boldsymbol{\Theta}^T\dot{\mathbf{X}}. \tag{2.15}$$

Next, the algorithm sequentially thresholds the library terms with coefficients smaller than a given threshold λ_t and updates the matrix of coefficients $\mathbf{C}$ via ridge regression with the hyperparameter α_2. The output of the SINDy algorithm via STLS contains a sparse matrix of coefficients $\mathbf{C}$ that is subsequently used to construct the differential equation model approximating the dynamics

$$\dot{\mathbf{x}}(\zeta) = \mathbf{f}\left[\mathbf{x}(\zeta)\right] = \mathbf{C}^T\left[\boldsymbol{\Theta}(\mathbf{x}^T(\zeta))\right]^T. \tag{2.16}$$

The SINDy algorithm with STLS identifies differential equations for each element of the state vector $\mathbf{x}$, while trying to keep the minimum number of the RHS terms required for the convergence.

Note that there are many different optimization methods employed in the framework of SINDy algorithm that are based on l_1 and l_2 norms [102, 103, 113]. The development of SINDy continues extending the approach for the boundary value problems [124], as well as enhancing the technique via deep generative modeling [125] and reinforcement learning [126].

2.4/ Results

The following section presents the results of applying the SINDy technique to the case of ideal FWM dynamics in an optical fibre, discussed in detail in Section 2.2. The technique is first applied to the Hamiltonian FWM model (see Eq. 2.7) in the ideal noise-free case. The performance of SINDy is then evaluated for input noisy data that mimics the experimental conditions. Next, the results of applying SINDy to coupled amplitude equations (see Eqs. 2.8) are presented.

2.4.1/ Sparse identification of ideal FWM dynamics

The main steps involved in the application of the SINDy algorithm to ideal FWM dynamics represented by the Hamiltonian model are highlighted in Fig. 2.3.

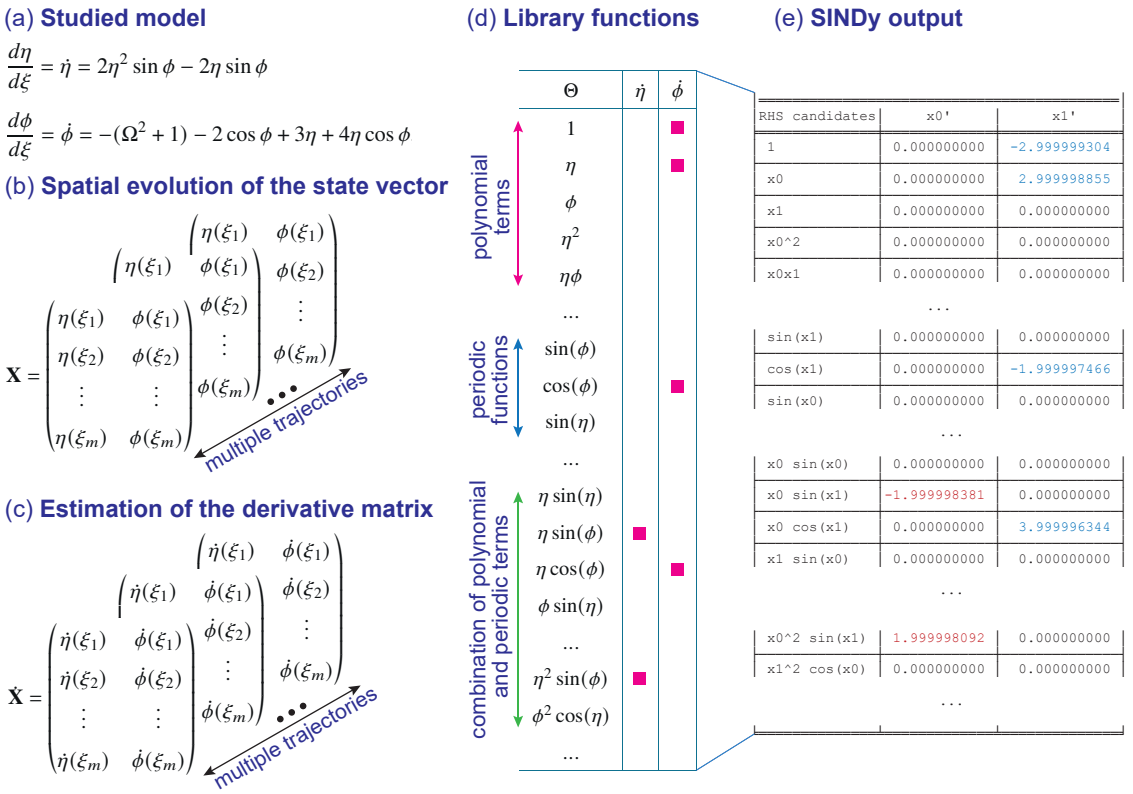


Figure 2.3: Flowchart of the SINDy algorithm applied to the Hamiltonian model of ideal FWM dynamics in an optical fibre. **(a)** shows the considered ideal FWM system in Hamiltonian formalism. **(b)** Multiple state vectors representing the $\eta(\xi)$ and $\phi(\xi)$ trajectories for different initial conditions. **(c)** Estimated derivative matrices. **(d)** A library of RHS candidate functions (the red rectangles highlight those associated with the ideal FWM system). **(e)** SINDy algorithm output showing the sparse coefficients estiamted for each candidate function where variables `x0` and `x1` correspond to η and ϕ respectively. The only nonzero returned coefficients for both variables are highlighted in red and blue. The threshold hyperparameter $\lambda_t = 0.5$ was used here as an input to the STLS optimizer (see the text for details).

Figure 2.3(a) shows the considered Hamiltonian model (see Eq. 2.7) the dynamics of which is simulated numerically for various initial conditions (η_0, ϕ_0) spanning the phase space. The input data to the SINDy algorithm contains multiple trajectories for η and ϕ representing the spatial evolution of the state vector [see Fig. 2.3(b)]. We then numerically estimate matrices of derivatives representing the LHS part of Eq. 2.12 [see Fig. 2.3(c)].

In the next step, the RHS part is assumed to be unknown and we construct a large library of RHS candidate terms. For the example given here, the library consists of η and ϕ polynomials and their combinations (up to the third order): $1, \eta, \phi, \eta^2, \eta\phi, \ldots, \eta\phi^2, \eta^3, \phi^3$, trigonometric functions of both variables: $\sin\eta, \sin\phi, \cos\eta, \cos\phi$, and combinations of polynomials and trigonometric functions: $\eta\sin\eta, \eta\sin\phi, \phi\cos\eta, \ldots, \eta^2\sin\eta, \phi^2\sin\phi, \ldots, \phi^3\cos\phi$ forming a total of 32 RHS candidate functions [see subfigure (d)]. The specific choice of RHS functions is based on the expected evolution of the system exhibiting the characteristic periodic Fermi-Pasta-Ulam-Tsingou recurrence dynamics [21, 22, 115].

The output of the SINDy algorithm is shown in Fig. 2.3(e), displaying the obtained coefficients associated with different RHS terms of the retrieved differential equation model for both variables η and ϕ. The input data contained 20 amplitude and phase trajectories $\eta(\xi)$ and $\phi(\xi)$ experiencing the dynamics on both side of the system's separatrix. These trajectories correspond to 20 different initial conditions (η_0, ϕ_0) randomly distributed throughout the phase space, such that $\eta_0 \in [0, 1]$ and $\phi \in [0, 2\pi]$. Specifically, the black rectangular dots in Fig. 2.4(c) show the initial conditions (η_0, ϕ_0) used to construct the input dataset for the SINDy algorithm.

To obtain the above results, Eq. 2.7 was solved numerically using the standard `odeint` from the `SciPy` library [127] with a spatial grid of 12000 points spanning from $\xi = 0$ to $\xi = 12$, which results in trajectories for η and ϕ typically containing 1 to 4 dynamical cycles depending on initial conditions.

Applying the SINDy algorithm with STLS optimization to these generated noise-free data, results in the output differential equation model with only a small number of active RHS terms from the library shown in Fig. 2.3(d). The algorithm successfully identifies the Hamiltonian FWM model (see Eq. 2.7) with 6 active RHS terms: 2 out of 32 RHS terms turn out to be active for $\dot{\eta}$ and 4 out of 32 RHS terms turn out to be active for $\dot{\phi}$ [as shown in Fig. 2.3(d) and (e)]. Moreover, from Fig. 2.3(e) we can see that the returned coefficients are correct within an accuracy of $\sim 10^{-6}$.

For completeness, Fig. 2.4 shows the dynamics of the ideal FWM process reconstructed using SINDy. The performance of the returned model is assessed for two different trajectories corresponding to the initial conditions $\eta_0 = 0.95$, $\phi_0 = 0$ and $\eta_0 = 0.95$, $\phi_0 = \pi$ (highlighted in red and blue, respectively)[4] exhibiting dynamics on different sides of the system's separatrix [shown by the black-dashed line in Fig. 2.4(c)]. Figures 2.4 (a) and (b) show the comparison of the ideal FWM dynamic simulated numerical using Eq. 2.7 (solid curves) with the model returned by SINDy (circles).

In the noise-free case, the excellent agreement between the retrieved and the ideal Hamiltonian models is perhaps not surprising in view of SINDy's well-established capability to analyse chaotic systems such as the Lorenz model [16]. Nevertheless, this example highlights several important details of sparse identification of nonlinear dynamical systems.

[4]Note that these trajectories were not used in the input dataset.

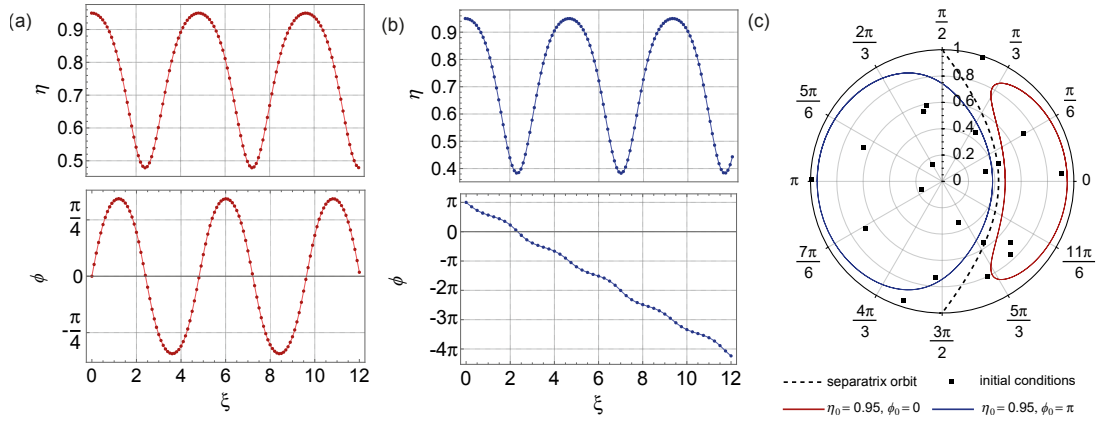


Figure 2.4: Comparison of ideal FWM dynamics with the model reconstructed by SINDy algorithm based on the analysis of the noise-free dynamical data. **(a)** Trajectories of $\eta(\xi)$ and $\phi(\xi)$ obtained for initial conditions $\eta_0 = 0.95$, $\phi_0 = 0$ from the output SINDy model (circles) and the Hamiltonian FWM model (solid line). **(b)** Trajectories of $\eta(\xi)$ and $\phi(\xi)$ obtained for initial conditions $\eta_0 = 0.95$, $\phi_0 = \pi$ from the output SINDy model (circles) and the Hamiltonian FWM model (solid line). Using the same color code **(c)** shows the two trajectories plotted in $\eta - \phi$ polar coordinates. The black squares represent the randomly chosen initial conditions in the phase space that were used to produce the input dataset. The ξ-evolution and phase space plots use red and blue colored lines for the dynamics in the right and left parts of the separatrix, respectively (the separatrix is shown as a black dashed line). Note that the phase space trajectories reconstructed using SINDy are not shown here, since they are indistinguishable from the ideal FWM dynamics in this case.

An important result here is that we clearly see how the technique can work well when the RHS candidate function library contains periodic and more complex functions, and not just simple polynomials. This is an important result because even though the SINDy algorithm allows inclusion of any functions in the candidate library $\boldsymbol{\Theta}$, the identification of the simplest and accurate models from data is still a challenge [102].

It can also be seen from the results that in the case of complex dynamical systems that can exhibit different physical regimes, it is essential to use an input dataset containing many evolutions of the state vector corresponding to different initial conditions. This allows SINDy to explore different states of the dynamical system. Additional analysis shows that for the ideal FWM model studied here in the noise-free case, SINDy can successfully extract the model coefficients with essentially the same accuracy even for an input dataset of only 5 trajectories for $\eta(\xi)$ and $\phi(\xi)$.

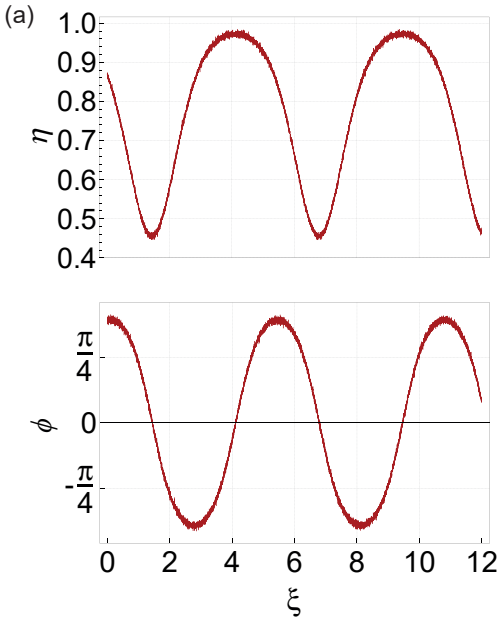
We also observed that including fewer trajectories in the input dataset leads to unreliable identification of active RHS terms, sometimes due to overfitting or obtaining oversimplified models. Indeed, including too few trajectories in this case may not allow the algorithm to fully explore the dynamics on both sides of the system's separatrix. As an example, we can consider an input dataset containing only trajectories corresponding to initial conditions leading to stationary solutions of the Hamiltonian system [see black rectangular and purple points in Fig. 2.2(a)]. In this case, the SINDy algorithm's search for the simplest differential equation models that reflect evolution is likely to result in oversimplified models that ignore the dynamics of some state vector variables [96].

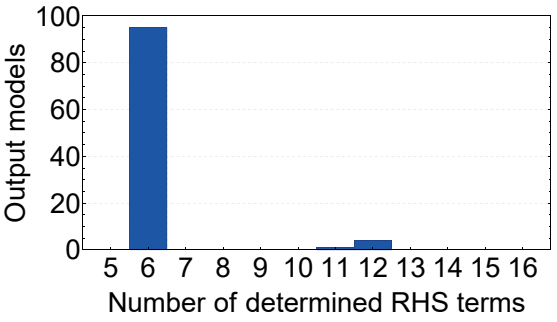
2.4.2/ Discovery of the ideal FWM model from noisy data

A more rigorous test involves discovering the underlying models from noisy data, as this paves the way for applying the technique to experimental measurements. The challenge, however, lies in the fact that sparse regression is potentially sensitive to noise, which requires adapting SINDy for this case [100, 113, 128]. One approach to address this is by applying the SINDy algorithm separately to random subsets ("bootstraps") of a given input dataset, resulting in several different returned models, each associated with its own terms and coefficients [113]. Statistical analysis of these different output models then allows us to estimate the mean values and uncertainties of the sparse coefficients.

In our analysis of the FWM dynamical data with noise, we employ a similar approach. However, instead of analyzing bootstrap samples from a single dataset [113], we consider an ensemble of input data based on scanning over different initial conditions. This approach is more typical for experiments in optics, where large datasets [119] can readily be measured. Specifically, we treat 2000 simulated trajectories for random initial conditions (η_0, ϕ_0) on both sides of the separatrix, and after computing each trajectory, we impose random multiplicative Gaussian noise with a relative noise factor δ (expressed in percentage) on the computed trajectories. We then combine the trajectories into 100 sets of 20, which are separately analyzed by SINDy, yielding 100 output models, each with their own terms and coefficients. Therefore, we can estimate the uncertainty in the model terms coefficients.

Figure 2.5(a) shows an example of $\eta(\xi)$ and $\phi(\xi)$ trajectories for $\delta = 2.5\%$ to illustrate the level of noise on the data.





(c)

RHS terms	coefs. vals.	estimated mean	estimated error
$\eta^2 \sin\phi$	−2	−1.9992	$\pm 1.3 \times 10^{-3}$
$\eta \sin\phi$	2	1.9989	$\pm 1.7 \times 10^{-3}$
$\eta \cos\phi$	4	3.9991	$\pm 2.0 \times 10^{-3}$
1	−3	−2.9998	$\pm 0.9 \times 10^{-3}$
η	3	2.9997	$\pm 1.7 \times 10^{-3}$
$\cos\phi$	−2	−1.9996	$\pm 1.1 \times 10^{-3}$

Figure 2.5: **(a)** shows an example of input $\eta(\xi)$ and $\phi(\xi)$ trajectories corresponding to initial conditions $\eta_0 = 0.87, \phi = 1.2$ with imposed $\delta = 2.5\%$ Gaussian noise. **(b)** Histogram displaying the number of active RHS terms in the output SINDy models. **(c)** shows the computed mean and standard deviation for each of the non-zero coefficients of the 6-term models. The threshold value for the STLS optimiser in this case was chosen to be $\lambda_t = 0.5$.

For each set of 20 trajectories, SINDy returns a differential equation model with a number of active RHS terms. To interpret the results, the histogram in Fig. 2.5(b) shows the number of active RHS terms detected for 100 output models. In this case, 95 out of 100 models contain only 6 active RHS terms, while up to 12 active RHS terms were found for the remaining 5 models. Models with a higher number of terms are rare and have non-identical active RHS terms, so we can expect these models to be associated with overfitting. In contrast, models containing only 6 terms were found to have similar active RHS terms while their coefficients may differ slightly. In addition to being the most frequently returned model in this case, it contains the fewest number of terms, making it the preferred model from a physical point of view.

In Fig. 2.5(c) we list the estimated mean and standard deviation for each of the active RHS terms. Comparison with the expected values from the Hamiltonian model shows that all the estimated mean values of the returned coefficients are within 1 standard deviation, while their standard deviations are of the order of $\sim 10^{-3}$.

To further assess the performance, we can now propagate the identified mean model and compare the predicted evolution with the ideal FWM dynamics (see Fig. 2.6). For this test, we chose initial conditions $\eta_0 = 0.63, \phi = \pi/3$ [subfigure (a)] and $\eta_0 = 0.58, \phi = \pi/3$ [subfigure (b)], resulting in two trajectories lying close to the separatrix, but on the opposite sides of it. The red solid curves in all plots show the dynamics of the ideal FWM process for the given initial conditions, while the blue dashed curves represent the dynamics reconstructed by the SINDy model with the mean values of coefficients. In this case the results are again visually indistinguishable.

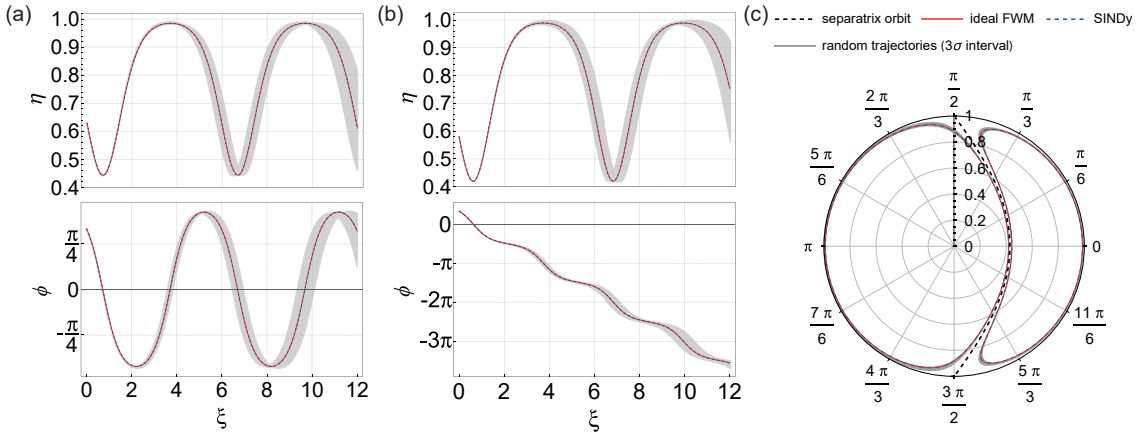
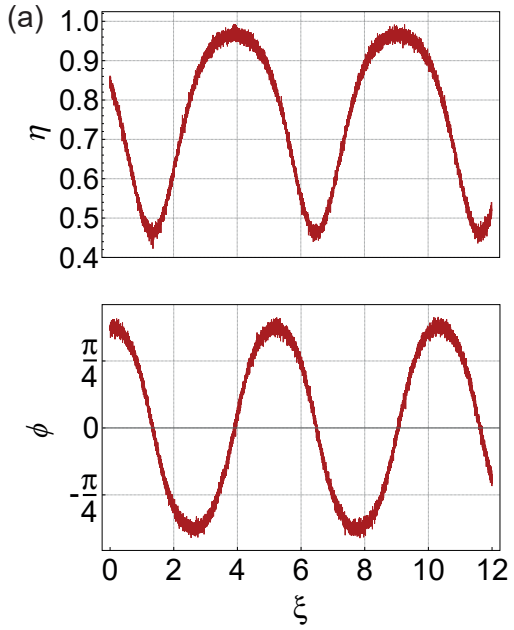
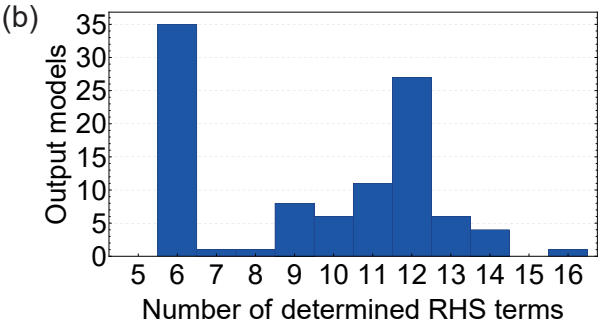


Figure 2.6: Comparison of the ideal FWM dynamics (red solid curves) with the mean SINDy model reconstructed from the input $\eta(\xi)$ and $\phi(\xi)$ trajectories with imposed $\delta = 2.5\%$ Gaussian noise (blue dashed curves): **(a)** for initial conditions $\eta_0 = 0.63, \phi = \pi/3$ and **(b)** $\eta_0 = 0.58, \phi = \pi/3$. The grey curves show the trajectories $\eta(\xi)$ and $\phi(\xi)$ obtained while randomly sampling the coefficients of the retrieved mean SINDy model over a range of 3 standard deviations. **(c)** displays the corresponding results plotted in phase space (the separatrix is shown as a black dashed line).

Since the coefficients of the identified mean model have statistical uncertainty limits, we can validate the results by predicting the trajectories $\eta(\xi)$ and $\phi(\xi)$ while randomly varying the model coefficients within 3 standard deviations. This result is represented by ensembles of gray curves in Fig. 2.6 and illustrates the mean model uncertainty intervals.

The results presented in Figs. 2.7 and 2.8 show the same statistical analysis performed on the input 100 sets of 20 trajectories for $\eta(\xi)$ and $\phi(\xi)$, but this time with imposed $\delta = 5\%$ Gaussian noise.





(c)

RHS terms	coefs. vals.	estimated mean	estimated error
$\eta^2 \sin\phi$	-2	-1.9978	$\pm 2.8 \times 10^{-3}$
$\eta \sin\phi$	2	1.9971	$\pm 3.8 \times 10^{-3}$
$\eta \cos\phi$	4	3.9950	$\pm 4.9 \times 10^{-3}$
1	-3	-2.9989	$\pm 1.5 \times 10^{-3}$
η	3	2.9973	$\pm 3.4 \times 10^{-3}$
$\cos\phi$	-2	-1.9977	$\pm 2.4 \times 10^{-3}$

Figure 2.7: **(a)** shows an example of input $\eta(\xi)$ and $\phi(\xi)$ trajectories corresponding to initial conditions $\eta_0 = 0.84, \phi = 1.16$ with imposed $\delta = 5\%$ Gaussian noise. **(b)** Histogram displaying the number of active RHS terms in the output SINDy models. **(c)** shows the computed mean and standard deviation for each of the non-zero coefficients of the 6-term models. The threshold value for the STLS optimiser in this case was chosen to be $\lambda_t = 0.5$.

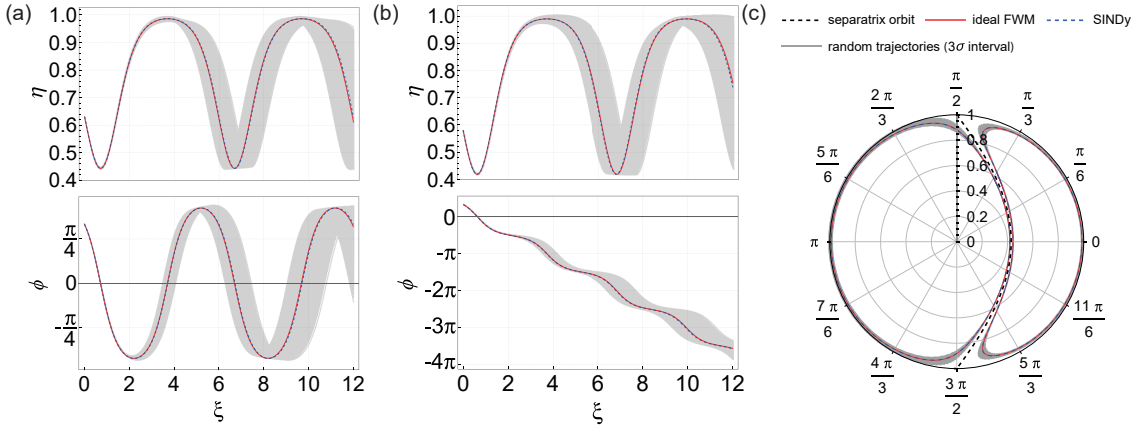


Figure 2.8: Comparison of ideal FWM dynamics (red solid curves) with the mean SINDy model reconstructed from the input $\eta(\xi)$ and $\phi(\xi)$ trajectories with imposed $\delta = 5\%$ Gaussian noise (blue dashed curves): **(a)** for initial conditions $\eta_0 = 0.63, \phi = \pi/3$ and **(b)** $\eta_0 = 0.58, \phi = \pi/3$. The grey curves show the trajectories $\eta(\xi)$ and $\phi(\xi)$ obtained by randomly sampling the coefficients of the retrieved SINDy model over a range of 3 standard deviations. **(c)** displays the corresponding results plotted in phase space (the separatrix is shown as a black dashed line).

As expected, at higher noise levels [see Figs. 2.7(a) and (b)] we get a much wider range of returned models, shown in the histogram in Fig. 2.7(b). Indeed, different output models contain up to 16 active RHS terms, but at the same time we can again observe a clear peak corresponding to a model with 6 active RHS terms (the same sparsity pattern was identified for 35 out of 100 output models). The table in Fig. 2.7(c) displays the results of statistical analysis of these 35 models output SINDy models. The resulting mean coefficient values still accurately describe the ideal FWM model, and the standard deviations are still at $\sim 10^{-3}$, but about twice as large as in the case of $\delta = 2.5\%$ noise level.

This larger uncertainty associated with the returned sparse coefficients is highlighted in Fig. 2.8, where we again plot an ensemble of grey curves corresponding to the model predictions when randomly varying the model coefficients within their 3 standard deviations. Despite this larger uncertainty limits, the mean SINDy model still provides a good fit to the ideal FWM dynamics.

At even higher noise levels, reliable model identification becomes more difficult. Our analysis shows that at $\delta = 7.5\%$ the 6-terms model is still identified most frequently, but the accuracy of the mean coefficient values is $\sim 10^{-2}$ compared to the ideal FWM dynamics. While the technique can still lead us to physically consistent results in the case of moderately to highly noisy data, it becomes clear that the noise level is crucial when trying to identify parsimonious models from data.

Another interesting detail that was observed when analysing SINDy output models is that, at different noise levels, the overfitted models with more than 6 active RHS terms usually contain sparse coefficients with large values of standard deviation (in some cases exceeding the mean values of the corresponding coefficients). Finally, for the input data with imposed $\delta = 10\%$ Gaussian noise, the histogram representing the number of active RHS terms in the output models becomes almost uniform, so that reliable model identification is not possible in this case. These results motivate the development of robust SINDy algorithms for identification of parsimonious differential equation models from data in high-noise regimes [113, 129].

2.4.3/ Application of SINDy to coupled amplitude equations

Another important aspect that determines the success of the SINDy algorithm is the choice of the right coordinates and physical variables when representing the dynamics [102]. Indeed, the FWM process considered here has a relatively simple structure in the Hamiltonian formalism, involving only 6 active RHS terms. The SINDy algorithm can also be applied directly to the set of coupled amplitude equations (see Eq. 2.8). By expanding the complex field amplitudes into real and imaginary parts: $A_0(\xi) = a_0(\xi) + ib_0(\xi)$ and $A_1(\xi) = A_{-1}(\xi) = a_1(\xi) + ib_1(\xi)$, Eq. 2.8 can be written in the following way

$$\frac{da_0}{d\xi} = -\Big(a_0^2 + b_0^2 + 2a_1^2 + 6b_1^2\Big)b_0 - 4a_0a_1b_1, \tag{2.17a}$$

$$\frac{db_0}{d\xi} = \Big(a_0^2 + b_0^2 + 6a_1^2 + 2b_1^2\Big)a_0 + 4b_0a_1b_1, \tag{2.17b}$$

$$\frac{da_1}{d\xi} = \frac{\Omega^2}{2}b_1 - \Big(a_0^2 + 3\Big[b_0^2 + a_1^2 + b_1^2\Big]\Big)b_1 - 2a_0b_0a_1, \tag{2.17c}$$

$$\frac{db_1}{d\xi} = -\frac{\Omega^2}{2}a_1 + \Big(b_0^2 + 3\Big[a_0^2 + a_1^2 + b_1^2\Big]\Big)a_1 + 2a_0b_0b_1, \tag{2.17d}$$

which yields 4 coupled differential equations over real-valued functions with a total of 22 active RHS terms. Coupled amplitude equations are, of course, very common in physics and can be used in the description of various phenomena.

In order to apply SINDy to this system, we first construct a library of RHS candidate functions $\boldsymbol{\Theta}$. A reasonable choice of RHS candidate functions here are polynomials of the involved variables $1, a_0, b_0, a_1, b_1, ..., a_0^2, a_0 b_0, a_0 a_1, ...b_1^3 a_1, b_1^4$, which we extend here to fourth order, giving a total of 70 RHS candidate terms. Obviously, this represents a much more complex scenario compared to the Hamiltonian system with only 6 active terms and a library of 32 RHS candidates. Using a similar approach as described earlier, we first apply SINDy to noise-free data. In this case, SINDy successfully identified all the correct dynamical terms with an accuracy of approximately $\sim 10^{-5}$, without any overfitting.

It has to be noted, however, that a larger number of candidate terms in the system increases the impact of noise. In particular, achieving an accuracy of the order of $\sim 10^{-3}$ for the 22 terms coefficients was possible only when the noise level was reduced by an order of magnitude ($\delta = 0.25\%$) compared to the results for the Hamiltonian system. Table 2.1 lists the results of the SINDy algorithm model identification for the case of $\delta = 1\%$ imposed Gaussian noise using the same statistical approach as before. We again simulate 100 sets of 20 trajectories for random initial conditions $[a_0(\xi = 0), b_0(\xi = 0), a_1(\xi = 0), b_1(\xi = 0)]$, impose Gaussian noise on the resulting trajectories, and use this data as input to SINDy. The algorithm successfully returns the correct model, identifying exactly 22 RHS terms, but the accuracy of the mean coefficient is noticeably lower than in the case of the Hamiltonian system[5].

Figure 2.9 shows a comparison between ideal FWM dynamics and the one reconstructed by SINDy using the mean values of the returned coefficients (see Table 2.1). The red solid curves represent the result obtained by numerically solving the Hamiltonian model equations (see Eq. 2.7) for initial conditions $\eta_0 = 0.63, \phi = \pi/3$. The blue dashed curves are reconstructed from the predicted evolutions of variables $a_0(\xi), b_0(\xi), a_1(\xi), b_1(\xi)$ by evaluating $\eta(\xi) = |a_0(\xi) + i b_0(\xi)|^2$ and $\phi(\xi) = 2\arg\left[a_1(\xi) + i b_1(\xi)\right] - 2\arg\left[a_0(\xi) + i b_0(\xi)\right]$.

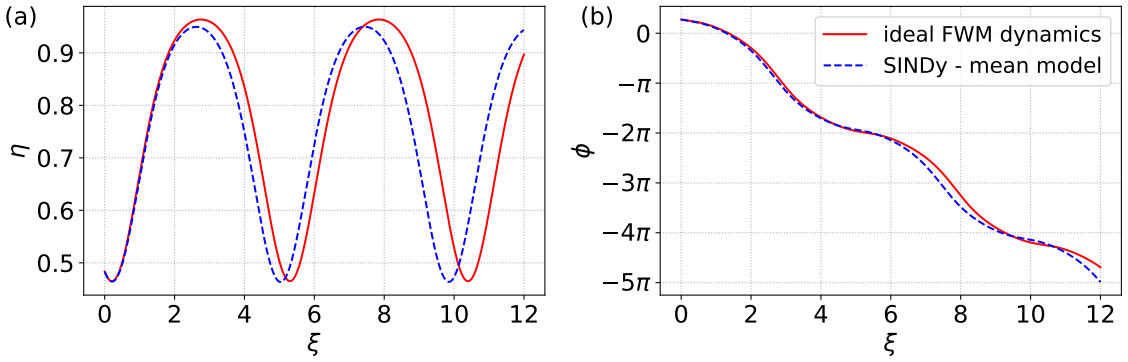


Figure 2.9: Comparison of ideal FWM dynamics (red solid curves) with the mean SINDy model reconstructed from the input $\eta(\xi) = |a_0(\xi) + i b_0(\xi)|^2$ and $\phi(\xi) = 2\arg\left[a_1(\xi) + i b_1(\xi)\right] - 2\arg\left[a_0(\xi) + i b_0(\xi)\right]$ trajectories with imposed $\delta = 1\%$ Gaussian noise (blue dashed curves) for initial conditions $\eta_0 = 0.63, \phi = \pi/3$.

[5]For simplicity, we do not show the estimated standard deviations here, which are typically of the order of 10^{-3} to 10^{-2} in this case.

Table 2.1: Application of SINDy to the coupled amplitude equations (see Eqs. 2.17) with a $\delta = 1\%$ Gaussian noise imposed on input data. First and second columns show the RHS terms and correct values of coefficients of the model (see Eqs. 2.17), respectively. Last 4 columns display mean values of the coefficients for active RHS terms identified through the statistical analysis of 100 returned SINDy models (see text). Only the identified active RHS terms are shown.

Active RHS terms	True coefficients	$\dot{a}_0$	$\dot{b}_0$	$\dot{a}_1$	$\dot{b}_1$
a_1	-1	–	–	–	-0.9180
b_1	1	–	–	0.9135	–
...	...	...	...	...	...
a_0^3	1	–	0.9995	–	–
$a_0^2 b_0$	-1	-0.9995	–	–	–
$a_0^2 a_1$	3	–	–	–	2.9341
$a_0^2 b_1$	-1	–	–	-0.9474	–
$a_0 b_0^2$	1	–	0.9991	–	–
$a_0 b_0 a_1$	-2	–	–	-1.9842	–
$a_0 b_0 b_1$	2	–	–	–	1.9846
$a_0 a_1^2$	6	–	5.9971	–	–
$a_0 a_1 b_1$	-4	-3.9972	–	–	–
$a_0 b_1^2$	2	–	1.9994	–	–
b_0^3	-1	-0.9993	–	–	–
$b_0^2 a_1$	1	–	–	–	0.9498
$b_0^2 b_1$	-3	–	–	-2.9309	–
$b_0 a_1^2$	-2	-1.9997	–	–	–
$b_0 a_1 b_1$	4	–	3.9975	–	–
$b_0 b_1^2$	-6	-5.9964	–	–	–
a_1^3	3	–	–	–	2.9139
$a_1^2 b_1$	-3	–	–	-2.9094	–
$a_1 b_1^2$	3	–	–	–	2.9145
b_1^3	-3	–	–	-2.9085	–
...	...	...	...	...	...

Inaccuracy in the returned coefficients leads to a discrepancy between the ideal FWM dynamics and the dynamics predicted by SINDy for large propagation distances. While the amplitude modulation depth is recovered with reasonable accuracy, there is a clear discrepancy in the periodicity of the amplitude and phase trajectories due to the poor estimation of the coefficients of the linear terms proportional to $\Omega^2/2$ (see Eq. 2.17).

2.5/ Summary, discussion & conclusion

There are several conclusions to be drawn from the results presented in this chapter. Firstly, we have shown using numerical simulations that data-driven discovery using sparse regression (SINDy algorithm) can successfully identify the governing differential equation model of ideal nonlinear FWM in optical fibres. The fact that this is possible in noise-free conditions is expected based on earlier studies [16], but our results also show that the SINDy algorithm yields successful results at noise levels of 5%, which are likely to be recorded in experiments. This result is particularly valuable because FWM is a central process not only in nonlinear optics but also in many other domains of physics, thus, the ability to SINDy correctly reconstruct these dynamics is an important validation step for the technique.

Secondly, we have discussed a convenient means of interpreting the results of SINDy in the presence of noise through the analysis of an ensemble of input dynamical data for different initial conditions. This approach involves inspection of a histogram distribution of the number of terms associated with multiple returned models, followed by computation of mean and uncertainty in the associated term coefficients. This allows us to readily compute the potential predictive accuracy of the models returned, by computation of the model's dynamics within the uncertainty limits. We anticipate that this approach to visualising model uncertainty may prove highly useful in determining the validity ranges of SINDy models derived from experimental measurements. Our results suggest that exploring the dynamics occurring under different initial conditions is essential for reliable identification of governing models, since many nonlinear physical systems can exhibit different dynamical regimes depending on the choice of initial conditions. In this regard, there are of course many improvements that one can consider such as combining an ensemble over initial conditions with internal data bootstrapping within each dataset. In addition, our analysis here has not implemented any specific preprocessing step to improve the numerical calculation of derivatives, and this is also a natural area of future work [130].

Finally, as a general conclusion, it is clear that sparse regression using SINDy promises to be a very powerful technique amongst the toolbox of methods available to researchers in nonlinear optics. Of course, the overall objective is to be able to analyse a partially-understood system with the aim of determining a governing model, and in this case a key element is the need to develop strategies to distinguish between different models that may be returned. We anticipate that the results presented here may point to further research in this direction and represent another advance in demonstrating the feasibility of data-driven discovery in nonlinear optics.

3

Data-driven dominant balance analysis of modulation instability

3.1/ Introduction

Being highly sensitive to initial and boundary conditions, nonlinear physical systems may exhibit different dynamical regimes during their evolution. In certain cases, it is possible to observe distinct stages of this evolution where a particular physical process (or interplay of processes) dominates the dynamics. The ability to separate these dynamical regions is essential, not only for purposes of qualitative interpretation of the dynamics, but also because mathematically, it allows one to significantly simplify the form of governing differential equations in each region. These reduced models can then be used to derive analytical solutions, optimize the numerical schemes and deepen the understanding of the underlying physics. Over the years, the use of conventional scaling analysis [131,132], asymptotic analysis [133], linear stability analysis [22], and perturbation theory methods [56,60] have helped uncover dominant physical mechanisms sometimes hidden due to the cumbersome structure of the generalised governing differential equations. In this regard, it is of great interest to develop data-driven techniques that are designed to complement conventional methods with the help of machine learning (ML) tools.

From a dynamical point of view, a dominant balance regime can be interpreted as a situation where only a subset of terms of the full differential equation model governs the dynamics at a given stage of the propagation (or equally, in some local region of the dynamics) [17, 134]. Recently, it has been suggested that the problem of identifying local dynamical regions can be solved using unsupervised clustering [17, 93]. The goal of any clustering algorithm can be to detect groups (clusters) of similar samples in the data or to determine the distribution of the data in the input data space [3]. The fundamental problem behind the data-driven search for dominant balances is to find a suitable parameter space in which clustering can be performed most efficiently.

In many areas of science where the dynamics of a system can be described by differential equation models, the dimensions of the desired parameter space can be related to the terms of the governing differential equation [17,18,93,94]. In other words, the cluster search can be performed in the "equation space", where each axis corresponds to a particular term in the governing differential equation. The identified clusters can subsequently be found to be dominated by the interaction of certain physical processes through "ad hoc" analyses based

on physical intuition [93] or through additional model selection algorithms [17, 18, 94].

Unsupervised dominant balance identification has been proven to be especially relevant when applied to the dynamics of nonlinear or chaotic systems, where the observed evolution is sometimes not well understood due to the complex interplay of several leading effects. In this case, ML algorithms can substantially enrich the conventional theoretical analysis by providing the data-driven interpretation of the physics. The data-driven dominant balance technique has been previously applied to a wide range of problems in the fields of geophysics and ocean sciences [93,135], active particle systems [136], turbulence and neuroscience [17, 18]. Moreover, the first application of the dominant balance technique to the field of nonlinear optics was also presented [17], where the dynamics of the supercontinuum (SC) generation process was analysed. The dominant balance algorithm was shown to correctly identify the ejection of the soliton driven by the balance between nonlinear and dispersive effects, as well as the appearance of the broad dispersive waves due to the higher-order dispersion. However, the algorithm implemented in [17] did not appear to be able to isolate the effect of Raman scattering on the initial stage of propagation. Trying to understand this observation and to apply the dominant balance approach to diverse propagation cases in optical fibres was one of the initial motivations of this thesis.

In nonlinear optics, the nonlinear Schrödinger equation (NLSE) is a fundamental model describing, with a given degree of approximation, a wide class of nonlinear field propagation scenarios in optical fibres [22]. As was highlighted in Chapter 1, the NLSE is known to be a fully integrable system, and its exact solutions can be constructed using the inverse scattering transform method [29]. This fact has given rise to a plethora of theoretical works studying the appearing periodic and localized solitonic structures as well as their interactions [36,40,45,46,137,138]. These findings were followed by the development of numerical simulation schemes [139] and experimental works that resulted in the first observations of the predicted nonlinear structures in optical fibres and water tanks [51, 52, 140–144]. Despite this great progress, the field remains to be an active area of fundamental theoretical and experimental research. One of the open problems that attracts a lot of attention is the spatio-temporal evolution of noise-driven modulation instability (MI), where the initial growth of the frequency sidebands is driven by small field perturbations [36] (see Section 1.3 of Chapter 1). It is just recently that the concept of "integrable turbulence" was introduced [38, 39, 145–147] enabling the theoretical investigation of the nonlinear stage [47] and statistics of random field in noise-driven MI [48].

Being well aware of the increasing number of papers focused on the rigorous physical description of MI and the development of new techniques to deepen the understanding of the emergence of localised nonlinear structures, in this chapter, we describe a fundamentally new approach to analyse the MI of the NLSE. First, the dominant balance approach is applied to interpret known analytic solutions. In particular, we consider the solitons on finite background (SFB) solutions associated with the MI dynamics (see Section 1.3.4 of Chapter 1). For these spatio-temporal dynamics, it is shown how the dominant balance approach can distinguish background regions of dominant nonlinear propagation from regions where nonlinearity and dispersion interact to drive localization. This is especially important in showing how data-driven dominant balance can provide complementary insights into the dynamics. Following the results found for the analytical SFB solutions, numerical simulations are used subsequently to study the more complex propagation case of noise-driven chaotic MI.

3.2/ Methodology of data-driven dominant balance

The concept of the data-driven dominant balance search that we present here aims to automate the identification of key interacting physical processes associated with different spatio-temporal (or similarly spatio-spectral)[1] regions of evolution. The particular approach developed by Jared L. Callaham, James V. Koch, Bingni W. Brunton, J. Nathan Kutz and Steven L. Brunton first introduced the unsupervised ML learning framework for the dominant balance identification. Recently, slightly different strategies have been proposed to fully automate this unsupervised search for dominant balance regimes [18,95]. Based on these works, below we show the most general framework of the data-driven dominant balance algorithm. This is outlined in flowchart form in Fig. 3.1.

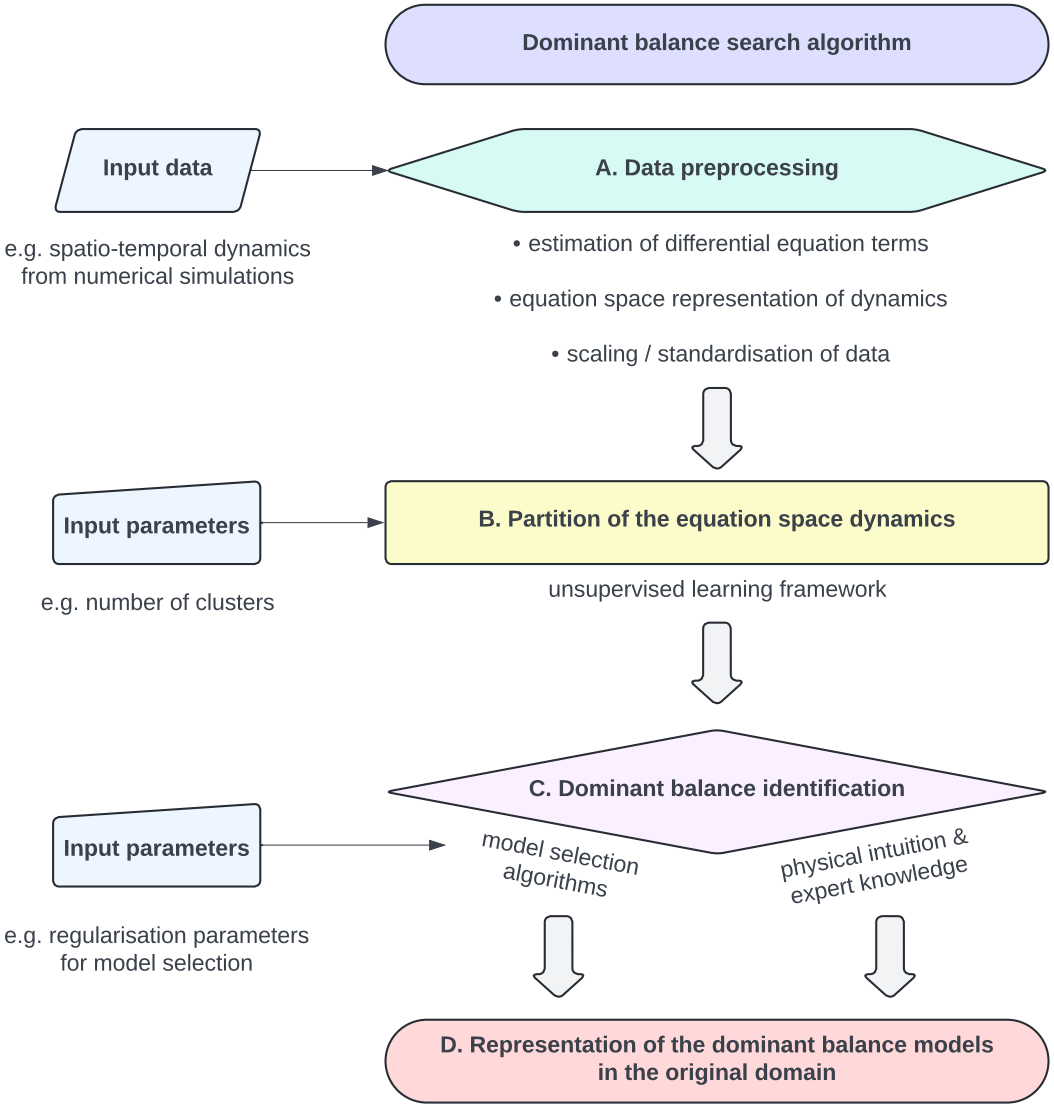


Figure 3.1: A flowchart of the general framework for implementing the dominant balance search algorithms.

[1]From now on in this chapter we refer to the analysis of spatio-temporal dynamics, although the same can be applied to the analysis of spatio-spectral domain.

The first step of the algorithm consists in loading dynamical data, which in our case represent the spatio-temporal evolution[2] of the field $\psi(\xi, \tau)$. The input dynamical data can be real or complex-valued and obtained from numerical simulations or from experimental measurements. In optics, the field $\psi(\xi, \tau)$ is complex-valued. The next steps of the dominant balance algorithm involve data preprocessing steps, an unsupervised clustering step that helps to partition the dynamical data into clusters and the dominant balance search steps, where each of the identified clusters are associated with particular dominating interacting physical processes (see flowchart steps A, B and C in Fig. 3.1). In the last step (see flowchart D in Fig. 3.1) the identified dominant balance models are mapped back to the original spatio-temporal domain for direct comparison with the input data. In the following subsections we describe each step of the dominant balance algorithm in detail.

3.2.1/ Equation space representation of dynamics

As discussed in the introduction to this chapter, in order to isolate the dominant physical regimes, it is important to perform the clustering of the dynamics in the appropriate parameter space. The novelty of the data-driven dominant balance approach consists in the unsupervised analysis of the equation space the axes of which are associated with the terms of the governing equation [17, 93]. This choice of the parameter space is not unique, but it was shown to be quite efficient in separation of distinct dynamical regimes. Below we elaborate on the proposed concept of equation space with a focus on the physical interpretation of equation space dynamics.

First, we write the general equation describing the spatio-temporal dynamics in the following form:

$$\sum_{k=1}^{K} f_k(\psi, \psi_\xi, \psi_\tau ..., \psi^2, \psi\psi_\xi, \psi\psi_\tau, ..., \psi_{\xi\xi}, \psi_{\tau\tau}, ...) = 0, \tag{3.1}$$

where f_k represents a variety of functions of the field ψ and its derivatives. K stands here for the total number of differential equation terms involved in the dynamics. The general form of a governing equation (see Eq. 3.1) emphasizes the balance relation that has to be satisfied throughout the spatio-temporal domain. Namely, in all the local dynamical regions the terms of the governing equation must sum to zero.

It is important to mention that in the framework of the dominant balance search, there could be different ways of representing the dynamics. For example, one can consider the integral or integro-differential form of the governing equation that satisfy the balance relation. The particular choice of the parameter space may vary from one physical problem to another. However, the equation space representation we discuss here is quite advantageous since the dominant balance of a few terms in the governing equation has a relatively simple geometric interpretation, as will be shown in the following sections.

As was pointed out in [17], the dominant balance of a subset of S out of K terms of the governing equation, which takes place in some local region of the dynamics, can be represented by a cluster (or group of dynamical points) in the equation space that has a significantly reduced variance with respect to the other $K - S$ terms. In other words, in this local region S leading terms will dominate the dynamics, while the contribution of the remaining $K - S$ terms can be neglected.

[2]Here we use ξ and τ to represent some spatial and temporal variables, respectively, which will be defined later.

The spatial and temporal distribution of the field $\psi(\xi, \tau)$ can be described by the matrix

$$\begin{pmatrix} \psi(\xi_1,\tau_1) & \psi(\xi_1,\tau_2) & \dots & \psi(\xi_1,\tau_M) \\ \psi(\xi_2,\tau_1) & \psi(\xi_2,\tau_2) & \dots & \psi(\xi_2,\tau_M) \\ \vdots & \vdots & \ddots & \vdots \\ \psi(\xi_N,\tau_1) & \psi(\xi_N,\tau_2) & \dots & \psi(\xi_N,\tau_M) \end{pmatrix}, \tag{3.2}$$

where we introduce the discretization of the spatio-temporal domain with N and M being the total number of points in ξ and τ, respectively. The "equation space" representation mentioned above consists in mapping the spatio-temporal dynamics (see Eq. 3.2) into the parameter space where each axis is associated with one of K terms of the governing equation (see Eq. 3.1). For this, we introduce the equation space matrix

$$\mathbf{\Lambda} = \begin{pmatrix} f_1[\psi(\xi_1,\tau_1)] & f_2[\psi(\xi_1,\tau_1)] & \dots & f_K[\psi(\xi_1,\tau_1)] \\ f_1[\psi(\xi_1,\tau_2)] & f_2[\psi(\xi_1,\tau_2)] & \dots & f_K[\psi(\xi_1,\tau_2)] \\ \vdots & \vdots & \ddots & \vdots \\ f_1[\psi(\xi_1,\tau_M)] & f_2[\psi(\xi_1,\tau_M)] & \dots & f_K[\psi(\xi_1,\tau_M)] \\ f_1[\psi(\xi_2,\tau_1)] & f_2[\psi(\xi_2,\tau_1)] & \dots & f_K[\psi(\xi_2,\tau_1)] \\ \vdots & \vdots & \ddots & \vdots \\ f_1[\psi(\xi_N,\tau_M)] & f_2[\psi(\xi_N,\tau_M)] & \dots & f_K[\psi(\xi_N,\tau_M)] \end{pmatrix}, \tag{3.3}$$

with dimensions $NM \times K$. Note that the elements of the equation space matrix can be directly estimated from the field distribution $\psi(\xi, \tau)$ via numerical differentiation[3]. An important property of $\mathbf{\Lambda}$ is that the sum of all elements in each row must be approximately zero (note that when the data are obtained from numerical modelling, errors may occur not only in the numerical solution of differential equation, but also in the estimation of the derivatives). The above aspects are accounted for in the first step of the dominant balance search algorithm (see step A of the flowchart in Fig. 3.1).

In the next step, the columns of matrix $\mathbf{\Lambda}$ (i.e. the input features) are used as an input to the unsupervised clustering algorithm in order to retrieve patterns in equation space data (see block diagram B in Fig. 3.1). Specifically, the goal of this next step is to separate the equation space dynamics into distinct clusters (groups of dynamical points) representing different dynamical regimes. A detailed graphical illustration of these first steps of the algorithm will be given in Section 3.2.3.

3.2.2/ Gaussian mixture model framework

The problem of identifying clusters (groups of dynamical points) in the equation space that represent distinct dynamical regimes can be addressed by various unsupervised clustering techniques. This includes the use of K-means clustering [93], DBSCAN [18] and Gaussian mixture modeling (GMM) [149]. The power of the unsupervised search for the dominant balances was demonstrated on a number of canonical problems in the fields of turbulence and geophysical fluid dynamics, such as the turbulent boundary layer problem and the oceanic barotropic vorticity [17, 18, 93]. It is worth noting that considering these examples using different clustering and model selection strategies yielded qualitatively similar results. Indeed, this emphasizes the possibility of implementing various approaches to the

[3]For this purpose we use the `findiff` Python library allowing accurate derivative estimation [148].

dominant balance identification problem that can be specifically tailored for the application purposes.

Indeed, having a variety of existing unsupervised clustering algorithms, it is of great interest to explore various strategies for retrieving the hidden dynamical regimes from the equation space data. One approach that has been shown to perform well in the task of identifying dominant balances is based on the GMM clustering, where learnt models can be used to extract physical insights.

Unsupervised cluster analysis using GMM provides an initial partition of the equation space into separate regions by means of training probabilistic models that assume the data is composed of a mixture of Gaussian distributions with different mean and covariance. In the general case, a multivariate Gaussian distribution of the K-dimensional vector $\mathbf{x}$ is given by [3]

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{K/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right], \tag{3.4}$$

where $\boldsymbol{\mu}$ is a mean vector, $\boldsymbol{\Sigma}$ is a $K\times K$ covariance matrix, and $|\boldsymbol{\Sigma}|$ denotes the determinant of $\boldsymbol{\Sigma}$. Then, a mixture of Gaussians can be represented in the following way[4]

$$\sum_{i=1}^{n_c}\pi_i\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i), \tag{3.5}$$

where $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)$ is associated with one of the n_c GMM components with its own mixing coefficient π_i, mean vector $\boldsymbol{\mu}_i$, and covariance matrix $\boldsymbol{\Sigma}_i$. In our case, the vector $\mathbf{x}$ represents the rows of the matrix $\boldsymbol{\Lambda}$, which are related to the notion of "observations" in ML. In what follows, we model the equation space dynamics $\boldsymbol{\Lambda}$ as a superposition of the GMM components (see Eq. 3.5). The motivation for using GMM clustering in particular is based on the idea that the learnt covariance structure of GMM clusters can be useful for the physical interpretation of the results [17].

In the context of the partition of equation space into the distinct dynamical regimes (see flowchart B in Fig. 3.1), we can now use the GMM clustering with n_c components to model the equation space distribution. In this case, the covariance matrix corresponding to the i-th component of the learnt GMM model that describes the equation space dynamics has the following structure

$$\boldsymbol{\Sigma}_i = \begin{pmatrix} \Sigma_{f_1[\psi(\xi,\tau)]} & \Sigma_{f_1[\psi(\xi,\tau)];f_2[\psi(\xi,\tau)]} & \cdots & \Sigma_{f_1[\psi(\xi,\tau)];f_K[\psi(\xi,\tau)]} \\ \Sigma_{f_2[\psi(\xi,\tau)];f_1[\psi(\xi,\tau)]} & \Sigma_{f_2[\psi(\xi,\tau)]} & \cdots & \Sigma_{f_2[\psi(\xi,\tau)];f_K[\psi(\xi,\tau)]} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{f_K[\psi(\xi,\tau)];f_1[\psi(\xi,\tau)]} & \Sigma_{f_K[\psi(\xi,\tau)];f_2[\psi(\xi,\tau)]} & \cdots & \Sigma_{f_K[\psi(\xi,\tau)]} \end{pmatrix}, \tag{3.6}$$

where the diagonal elements $\Sigma_{f_k[\psi(\xi,\tau)]}$ have the meaning of the variance of the chosen Gaussian component with respect to the particular axis in the equation space (or, equivalently, with respect to the particular term f_k of the differential equation). Off-diagonal elements of a covariance matrix describe the correlation between the differential equation

[4]The estimation of the mentioned statistical moments of Gaussian distributions, as well as the GMM clustering are standard and are algorithmically implemented in many programming languages. For this reason, we do not show the details of their computation, but rather focus on their physical interpretation in the context of the dominant balance approach.

terms in the chosen component and define the orientation of the corresponding cluster in the equation space (note that the covariance matrix $\mathbf{\Sigma}_i$ is symmetric).

The information embedded in the structure of the covariance matrix of the GMM components turns out to be very valuable for the interpretation of the dominant physical models. In statistical terms, the covariance matrix of a multivariate Gaussian distribution is equivalent to the concept of variance in the one-dimensional case. A larger variance in a certain direction means that this direction has a greater influence on the results and vice versa. In the context of the equation space dynamics, if some diagonal elements in one of the GMM clusters have near-zero entries and the other terms have almost no correlation with them, it can be assumed that the influence of the corresponding terms of the differential equation is negligible for these dynamics. In other words, the dynamics associated with this cluster will be approximately dominated by differential equation terms with significantly non-zero covariance matrix elements. This can be very instructive in the manual recognition of dominant balances as well as in the implementation of model selection algorithms. An illustrative example of analysing the covariance structure of the GMM components is given in the following section when considering the Peregrine soliton (PS) dynamics (see Section 3.2.3).

The particular GMM algorithm that is employed in this thesis is the `GaussianMixture` from the Python package `scikit-learn` [150], as implemented in [149]. This algorithm is designed to softly partition the data into a given number of n_c regions, each of which is associated with a multivariate Gaussian distribution. The n_c is, thus, a hyperparameter to be selected depending on the governing equation structure. As there are no guarantees that one of the GMM clusters will fully capture the dynamics corresponding to the particular dominant balance regime, the hyperparameter n_c should normally be selected to be higher than the number of potential dominant balances. All the identified clusters can be grouped together in the next step of the algorithm to form the final dominant balance models (see block diagram C in Fig. 3.1).

In addition, despite the fact that the feature scaling is not mandatory when performing the GMM clustering on a dataset, it is generally a good practice to use standardisation in such cases. Indeed, the GMM implementation mentioned above uses the K-means strategies to set the starting values of the expectation-maximization algorithm and, thus, can be sensitive to feature scaling. To better partition the equation space into subsets of dynamic points, we use standardisation of the $\mathbf{\Lambda}$ matrix before clustering. The process consists of rescaling each feature so that its standard deviation is equal to 1 and its mean is 0 (see `StandardScaler` of `sklearn.preprocessing` [150]).

Finally, it is important to mention that, in general, the assumption that the input data will follow a GMM distribution is not justified. In our case, the input data are assumed to be generated as a mixture of Gaussians with zero kurtosis and skewness. As it is not always the case with the equation space dynamics, there is still some ambiguity in the selection of the initial number of GMM clusters n_c. In our analysis, we simply choose a number of clusters greater than the number of expected dominant balances, typically $n_c \leq 2^K$ to avoid overfitting, while visually controlling that the dynamics of the equation space is sufficiently partitioned. Of course, there are various strategies that can be implemented, such as using Bayesian inference criterion [3] or special clustering algorithms with physical constraints. And this naturally opens up directions for future developments.

3.2.3/ GMM clustering analysis of the Peregrine soliton dynamics

To better illustrate the idea of the dominant balance search, below we consider an example of the Peregrine soliton (PS) dynamics in optical fibres which is analysed in the corresponding equation space via unsupervised GMM clustering. These dynamics are governed by the NLSE that we write here in the following dimensionless form (see Eq. 1.55)

$$i\,\psi_\xi + \psi_{\tau\tau} + |\psi|^2\psi = 0, \tag{3.7}$$

where $\xi = z/L_{NL}$, $\tau = T/\sqrt{|\beta_2|L_{NL}/2}$, and $\psi(\xi,\tau) = A(z,T)/\sqrt{P_0}$ (see Section 1.3 of Chapter 1). The PS corresponds to the SFB family of NLSE solutions (see Eq. 1.57 in Chapter 1). In the limit $a \to 0.5$ the solution of Eq. 1.57 can be written as a linear combination of a constant plane wave background and rational solitonic solution

$$\psi_{\rm ps} = \left(1 - 4\frac{1+2i\xi}{1+4\tau^2+4\xi^2}\right)\exp(i\xi), \tag{3.8}$$

Figure 3.2(a) shows the intensity map representing the spatio-temporal dynamics of the PS (see Eq. 3.8).

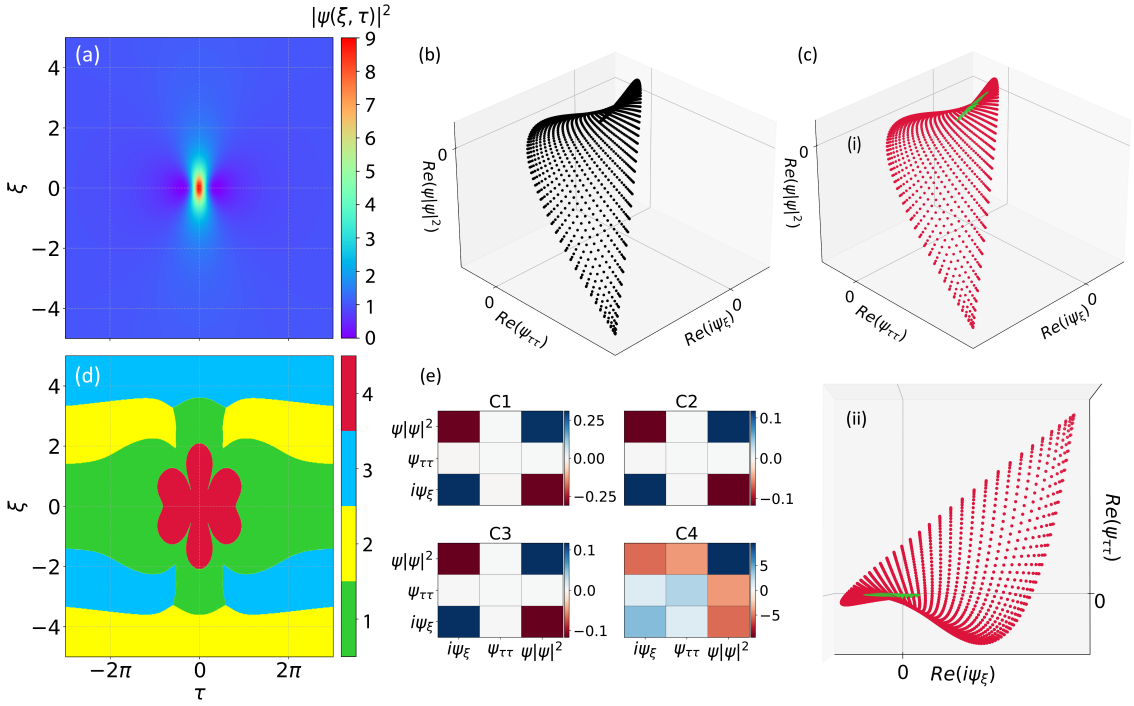


Figure 3.2: Unsupervised GMM clustering applied to the equation space dynamics of the PS. **(a)** Spatio-temporal dynamics represented by $|\psi(\xi,\tau)|^2$. **(b)** Equation space dynamics of the PS plotted in coordinates associated with real components of NLSE terms $i\psi_\xi, \psi_{\tau\tau}, \psi|\psi|^2$. **(c-i)** shows the four identified GMM clusters in the equation space represented by different colors. Note that yellow and light blue clusters (clusters C_2 and C_3, respectively) are not visually distinguishable due to their low variance and close overlap with the green cluster (cluster C_1). In **(d)**, using identical color code, the identified GMM clusters are mapped back to the original spatio-temporal domain. **(e)** displays the covariance matrices of the four identified GMM clusters. In **(c-ii)** one of the projections of clustered equation space is shown using the same color code.

Following the dominant balance search algorithm, we first represent the spatio-temporal NLSE dynamics in the associated equation space $\{i\,\psi_\xi, \psi_{\tau\tau}, |\psi|^2\psi\}$. Next we perform the GMM clustering and separate the equation space into $n_c = 4$ clusters exceeding the number of expected dominant balances (with $K = 3$, we consider 3 potential dominant balances, associated with dispersive $\{i\,\psi_\xi, \psi_{\tau\tau}\}$, nonlinear $\{i\,\psi_\xi, |\psi|^2\psi\}$ and full NLSE propagation $\{i\,\psi_\xi, \psi_{\tau\tau}, |\psi|^2\psi\}$).

Figure 3.2(c) shows the resulting equation space dynamics separated into four clusters, each associated with its own multivariate Gaussian distribution [note that yellow and light blue clusters (clusters C_2 and C_3, respectively) are not visually distinguishable due to their low variance and close overlap with the green cluster (cluster C_1)]. We mention also that here we only show the real parts of the differential equation terms when representing the equation space dynamics, but similar results are found for the imaginary parts. In Fig. 3.2(b) we see how these four identified equation space clusters partition the spatio-temporal map into distinct regions when mapped back to the original domain (highlighted in green, yellow, light blue, and red, respectively).

As mentioned in Section 3.2.2, the most important parameter of the returned GMM clusters is their covariance structure that can already guide us through the identification of dominant balance models. Figure 3.2(e) shows the learnt covariance matrices for each of the identified clusters, and we can readily distinguish some patterns in their structure. For example, one can point out the similarity in the covariance structure of the clusters C_1, C_2 and C_3, where we have a strong correlation between the propagation and Kerr nonlinear terms $\{i\,\psi_\xi, |\psi|^2\psi\}$. In the equation space [see Fig. 3.2(c)], these 3 clusters (highlighted in green, yellow, and light blue, respectively) are closely packed together and constrained in the same direction of the equation space. From this, one would expect them to belong to the same dynamical regime (or the same dominant balance). It can also be observed that the covariance structure of these 3 clusters does not show a significant contribution of the second order dispersion, so that they most likely belong to a dominant balance model associated with the dominance of nonlinearity.

A qualitatively different covariance structure is observed for the cluster C_4 (colored red), where we see the clear presence of the second order dispersion as well, so one would expect this particular cluster to be associated with the full NLSE dynamics. Indeed, when we look at the equation space again, we see that the cluster mentioned above is spread all over equation space without having a reduced variance with respect to any direction. As an additional remark, note that in the above example the initial number of clusters can be chosen arbitrarily, as discussed earlier. For simplicity of this illustrative example, $n_c = 4$ is chosen here.

We can see from this example how the learnt covariance structure of the GMM clusters can help us in finding different dynamical regimes in the evolution. At this stage, the expert knowledge can be used to perform the identification of the dominant balance models (see block diagram C in Fig. 3.1). This consists in manually grouping the identified clusters together, for example, based on their covariance structure to form the final dominant balance models [93]. However, this would require "a priori" knowledge and/or intuition about the different dynamical regimes in the considered system.

Inspection of the covariance structure of the clusters can assist in model identification, but there will still be some ambiguity in the procedure. For example, from Fig. 3.2(d), we can see that different GMM clusters can represent the dynamical patterns at quite different scales. Specifically, the scale of the covariance matrices for clusters C_1, C_2 and C_3 are

several orders of magnitude smaller compared to that of cluster C_4 [see the colorbars in Fig. 3.2(d)]. This means that the dynamics captured by some clusters can be much less important globally.

Another difficulty is that in certain cases it may not be clear how to assign a GMM cluster to a particular dominant balance based on its covariance structure. In particular, since the returned covariance matrices are dense, it is necessary to introduce certain criteria to assign a given cluster to a particular dominant balance model. All this motivates the development of algorithmic model selection strategies that can replace the use of human intuition in the task of identifying dominant physical interactions. In the following section we discuss one of the initially proposed approaches based on the sparse principal component analysis (SPCA) [17].

3.2.4/ Model selection: Sparse principal component analysis

SPCA extends conventional principal component analysis by introducing a sparsity criterion for the input features [151]. Conventional principal component analysis describes an orthogonal projection of the input data onto a linear space with lower dimensions, maximising the variance of the projected input data [3]. The principal component analysis technique is consistent with clustering methods based on the expectation-maximisation algorithm, such as GMM, and can be directly applied to the equation space data to identify lower-dimensional patterns represented by directions of maximum variance [8].

However, it is difficult to directly use principal component analysis to perform the dominant balance model selection because the obtained GMM covariance matrices are usually dense. This is essentially the same problem encountered when trying to manually assign a GMM cluster to a particular dominant balance based on its covariance structure, as discussed in Section 3.2.3. To overcome this difficulty, it was proposed to take into account only the sparse approximation corresponding to the first principal component, which contains most of the variance in each GMM cluster [17]. The SPCA may label features with small variances as insignificant. If a GMM cluster has entries in the covariance matrix that are close to zero for a particular term of the differential equation, that term can be automatically discarded using SPCA. This can be done by solving this problem in the framework of the least absolute shrinkage and selection operator (LASSO) regression [123] on the principal axes. The SPCA thus returns a sparse and interpretable model that can capture the leading terms governing the dynamics in each GMM cluster.

The implementation of the SPCA for the dominant balance model selection, however, introduces an additional hyperparameter α, known as the LASSO l_1 regularisation hyperparameter. As in various ML tasks, the hyperparameter α balances descriptiveness and sparsity of the identified models. Specifically, for very low values of α, one gets the dominant balance models that include all the differential equation terms. Contrary to that, for higher values of α the dynamics captured by a cluster will be represented by sparse dominant balance models that select just a few dominating terms. Finally, for really high α values, we eventually get "empty" models, where all the equation terms are neglected. From this, we can see that the algorithm requires an optimization step that finds the best hyperparameter value α_{opt} to distinguish the sparse and interpretable dynamical models without loss of the descriptiveness.

Algorithm 2 shown below lists the main steps of the dominant balance model selection procedure implemented in the framework of SPCA [149].

Algorithm 2 Dominant balance model selection algorithm via SPCA

- Initialize a vector of the LASSO regression hyperparameters α_p
- Allocate a zero vector of reconstruction errors ϵ_p

I. Construct the regularization path
for each of the LASSO regression hyperparameter α_p **do**
  for each GMM cluster C_i **do**
    1. Extract an array $\mathbf{\Lambda}_i$ corresponding to a subset of equation space matrix $\mathbf{\Lambda}$ associated with the cluster C_i
    2. Initialize a SPCA model with a given α_p and prescribed number of principal components to compute
    3. Perform the SPCA with the array $\mathbf{\Lambda}_i$ as an input data
    4. Label the active terms identified via SPCA
    5. Extract the feature array $\mathbf{\Lambda}_r$ associated with inactive terms from the SPCA
    6. Compute the reconstruction error for α_p: $\epsilon_p = \epsilon_p + \|\mathbf{\Lambda}_r\|_F$
  end for
end for
II. Find the optimal value α_{opt} by analysing the Pareto-type curve of the reconstruction error ϵ against α
III. For selected α_{opt} group the GMM clusters with the same active terms to form the final dominant balance models

The GMM clusters obtained in the previous step are used as the input to the model selection algorithm. Each cluster C_i, where $i \in [1, n_c]$, is associated with an array $\mathbf{\Lambda}_i$ representing a subset of the equation space matrix $\mathbf{\Lambda}$. The hyperparameter α is optimised with respect to the reconstruction error estimated as the Frobenius norm of inactive terms in a given cluster (see Eq. 3.10).

In the first step, we initialize a vector of LASSO regression hyperparameters α_p over which the optimization is performed. Typically, we select $\alpha \in [\alpha_{\min}, \alpha_{\max}]$ with $\alpha_{\min} = 10^{-2}$ and $\alpha_{\max} = 10^3$ found empirically, separated by 100 log-equidistant points. Now we can construct the regularization path to select the α_{opt}. For every α_p we loop through all the identified clusters and accumulate the reconstruction error of inactive terms. To this end, the SPCA is performed sequentially for each of the GMM clusters C_i. The optimisation task to be addressed is a principal component analysis problem with a l_1 penalty [150]

$$\hat{A}_i(\alpha), \hat{B}_i(\alpha) = \arg\min_{A,B} \frac{1}{2}\|\tilde{\mathbf{\Lambda}}_i^T - AB\|_F^2 + \alpha\|A\|_1, s.t \|B_{i:}\|_2 \leq 1, \tag{3.9}$$

where $\tilde{\mathbf{\Lambda}}_i$ represents the absolute matrix of the original equation space matrix $\mathbf{\Lambda}$ with the subtracted column-wise mean, $\tilde{\mathbf{\Lambda}}_i^T$ is a singular value decomposition (see `SparsePCA` in [150] for details). The SPCA automatically labels active and inactive terms in each cluster at a given α_p. From this, one can compute the reconstruction error for cluster C_i: $\epsilon = \|\mathbf{\Lambda}_r\|_F$, where $\mathbf{\Lambda}_r$ is a subset of $\mathbf{\Lambda}_i$ representing a feature array of inactive terms in cluster C_i found via SPCA with a given α_p. The notation $\|.\|_F$ stands for the Frobenius norm

$$\|A\|_F = \sum_{i,j} \sqrt{|A_{ij}|^2}. \tag{3.10}$$

Finally, the reconstruction error associated with a particular input α_p can be computed simply as a sum of reconstruction errors over all the GMM clusters (throughout the entire equation space). This entire procedure is shown in Algorithm 2(I).

As the output of the Algorithm 2(I), we get the $\alpha-$path represented as the reconstruction error ϵ against α. This forms the Pareto-type curve balancing sparsity and model accuracy. The selection of α_{opt} can be made by inspecting the obtained curve. For values of α close to $\alpha_{\min}$, the reconstruction error is usually relatively small, as most of the terms are used to describe the dynamics in each cluster. As we move towards higher α, SPCA starts to neglect the terms with small variance in the equation space (equivalently, with near zero entries in the covariance matrix) resulting in a slow increase of ϵ.

At some point, the reconstruction error starts to grow rapidly, leading to an inflection of the curve. Physically it means that at higher values of α, SPCA eventually gets rid of the differential equation terms playing essential role in the dynamics. Based on this, to select α_{opt} one may look at values of α just before the rapid growth of the reconstruction error [see step (II) of Algorithm 2]. The clusters with the same sparse structure (with the same identified active terms) can be grouped together and form final dominant balance models.

Let us now come back to the PS example and perform the dominant balance model selection using the SPCA algorithm (see Fig. 3.3)[5].

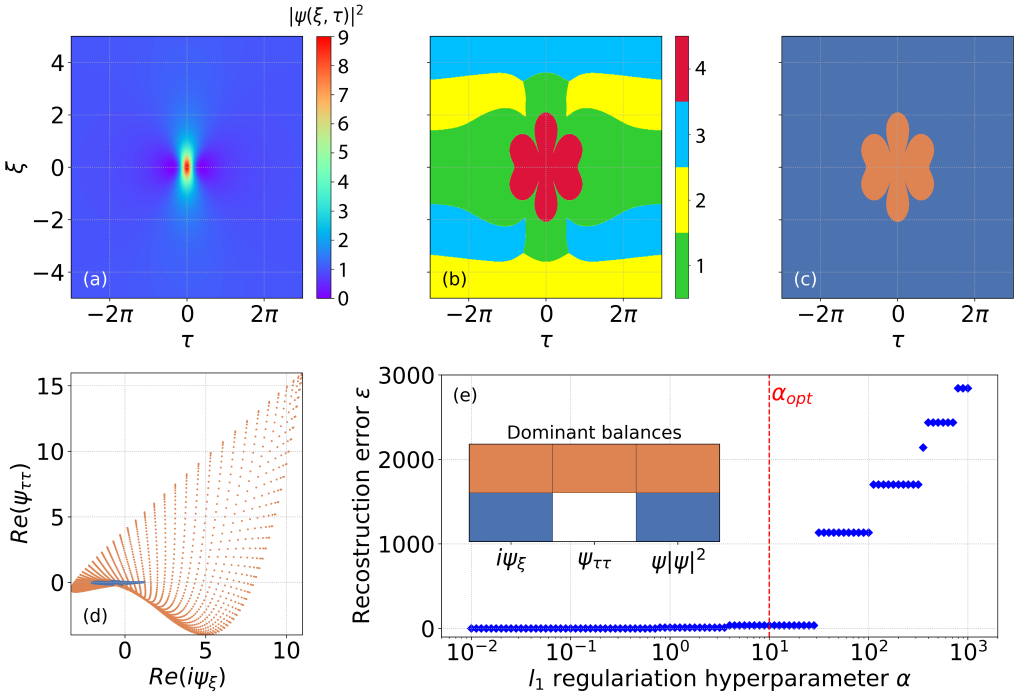


Figure 3.3: Model selection algorithm implemented in the framework of the SPCA illustrated on the PS example. **(a)** spatio-temporal dynamics represented by $|\psi(\xi,\tau)|^2$. **(b)** the identified on the first step four GMM clusters mapped to the original spatio-temporal domain (the same color code is used here as in Fig. 3.2). **(c)** shows the dominant balance regions found via the SPCA model selection algorithm. The two returned dominant balance models (blue and orange) are shown in subfigure **(e)**. **(d)** using identical color code, shows one of the projections of the identified dominant balance clusters in the equation space. **(e)** displays the computed $\alpha-$path via SPCA model selection algorithm with the highlighted value of α_{opt}.

Figures 3.3(a) and (b), respectively, show the spatio-temporal dynamics of the PS and the

[5]Note that we reproduce some elements of Fig 3.2 here in order to more fully explain the procedure.

corresponding spatio-temporal cluster map separated into local regions associated with four identified GMM clusters [see Fig. 3.2(c)]. Next, the four identified GMM clusters are used as the input to the SPCA model selection algorithm to compute the $\alpha-$path shown in Fig. 3.3(e). We can see that the obtained curve shows a "staircase" growth with increasing α. In fact, each of these "stairs" corresponds to different models with increasing sparsity. We can directly see that the reconstruction error ϵ experiences a jump after $\alpha \approx 30$ and starts to grow rapidly. On this basis, one can set $\alpha_{opt} = 10$, which corresponds to the sparsest model with a relatively low reconstruction error that appears just before the inflection point.

After setting $\alpha_{opt} = 10$, the algorithm returns two dominant balance models [see Fig. 3.3(c)]. The first model, represented by the blue region in Fig. 3.3(c), describes the dominant balance between propagation and nonlinear Kerr terms $\{i\psi_\xi, |\psi|^2\psi\}$ and is associated with the plane wave background. We can now see how the three GMM clusters (green, yellow and light blue) were grouped together to form the nonlinear dominant balance model [see the comparison of Figs. 3.3(c) and (d)]. This demonstrates the robustness of the algorithm with respect to the redundant clusters that represent essentially the same dynamical regime.

In the illustrative example considered here, grouping these three clusters together can be quite straightforward based on their orientation and close overlap in the equation space [see Fig. 3.2(d)]. This result is consistent with the previously discussed idea of the covariance structure of the returned GMM clusters. Indeed, these clusters have quite similar structure of the covariance matrices with near zero entries with respect to the dispersive term [see Fig. 3.2(e)]. That is in contrast with the cluster C_4 that shows a clear correlation between all the terms in its covariance matrix. We can see how this cluster was subsequently assigned to the full NLSE dominant balance [orange region in Fig. 3.3(c)] that separates the localised solitonic structure from the background. The physical interpretation of the identified local dominant balance models for the SFB and noise-driven MI dynamics will be discussed in detail in Section 3.3.

Before discussing the physical aspects of the returned dominant balance models, it is instructive to take another look at the equation space dynamics. Figure 3.3(d) shows one of the projections of the NLSE equation space, where the dynamical points are plotted in $\mathrm{Re}(i\psi_\xi) - \mathrm{Re}(\psi_{\tau\tau})$ coordinates. The equation space initially separated into four GMM clusters now contains only two dominant balance regions (orange and blue, respectively).

The results shown clearly illustrate the geometric interpretation of the dominant balance. We can see how the nonlinear dominant balance cluster (blue) is restricted in the equation space to have near zero variance with respect to the dispersive term $\psi_{\tau\tau}$. In other words, the dynamical points associated with the dominant nonlinear regime are approximately confined with respect to the direction $i\,\psi_\xi + |\psi|^2\psi \approx 0$ in the three-dimensional equation space. Contrary to that, the full NLSE dynamics [see orange dominant balance cluster in Fig. 3.3(d)] is associated with dynamical points distributed on a plane in the equation space with no reduced variance with respect to any directions. This is because geometrically the full NLSE ($i\,\psi_\xi + \psi_{\tau\tau} + |\psi|^2\psi = 0$) represents a plane in the corresponding equation space. While being quite intuitive in the case of the NLSE with only three differential equation terms, the dominant balance equation space interpretation is quite complex in the case of the generalised NLSE. The aspects of the generalised NLSE dynamics analysed with dominant balance approach are presented in Chapter 4 of this thesis.

The dominant balance model selection via SPCA finalizes the dominant balance search

algorithm (see Fig. 3.1). It is important to emphasise that the proposed approach is quite different from an alternative method such as global thresholding. In a simplistic manner, one can try to establish local dominant balance regions by imposing a hard threshold on differential equation terms (i.e. when the absolute contribution of a term is below the threshold, it is neglected). The aim of the data-driven dominant balance approach presented here is to find dominant physical models based on the local relative importance of differential equation terms. In many physical systems, the scale of significance varies locally across the domain, so the threshold approach will not work. As will be shown in Section 3.3 the presented dominant balance algorithm searches for spatio-temporal regions of similar dynamical regimes, while accounting for the local relative importance of terms. In the simplest cases, where the scale of dynamics does not change significantly across the domain, the thresholding approach is, in fact, a limiting case of the proposed algorithm.

3.3/ Results

In this section, the data-driven dominant balance approach is used to analyse various aspects of the MI of NLSE. The approach is first applied to the analytical SFB solutions describing various localised and periodic structures associated with MI propagation (see Section 1.3 in Chapter 1). The dominant balance search algorithm is then applied to the case of chaotic noise-driven MI dynamics.

3.3.1/ Dominant balance analysis of SFB dynamics

The SFB (see Eq. 1.57) describe a broad family of exact NLSE solutions that are intimately related to MI dynamics. The MI in optical fibres is also associated with the emergence of various complex localised structures that have been explored in many previous studies, including higher-order NLSE solutions [152], collisions of breathers and solitons [138], ghost interactions [142], etc. The data-driven dominant balance approach, utilized alongside established methods like the inverse scattering transform [39, 139, 153], can enrich our understanding and representation of the complex MI dynamics.

We first provide an interpretation of the dominant balance analysis of the Peregrine soliton (PS) dynamics performed in Section 3.2. Figure 3.4(a-i) again shows the spatio-temporal dynamics of PS represented by $|\psi(\xi,\tau)|^2$, while Fig. 3.4(a-ii) shows two dominant balance models identified in the spatio-temporal domain: the one associated with the dominant balance between propagation and nonlinear Kerr terms $\{i\psi_\xi, |\psi|^2\psi\}$ (blue) and the one associated with the full NLSE dynamics (orange). From the comparison of Figs. 3.4(a-i) and 3.4(a-ii), it becomes apparent that strong spatio-temporal localization around ($\xi = 0, \tau = 0$) emerges due to the combined effects of nonlinearity and dispersion within the NLSE (orange region). In fact, this is the full NLSE dynamics that drives the spatio-temporal compression. In contrast, the surrounding low intensity region is found to be dominated by nonlinearity (blue). This result demonstrates well how the dominant balance technique can successfully separate two distinct physical regimes involved in dynamics.

It is instructive to also look at the results of the dominant balance identification in the associated equation space. Figure 3.4(b) shows the equation space dynamics divided into two groups of dynamical points associated with nonlinear and full NLSE dominant balance models (blue and orange, respectively)[6]. From the projections shown in Figs. 3.4(b-ii) and (b-iii), we can directly see how these found dominant balance models are distributed in the equation space. The nonlinear dominant balance (blue) is confined in the equation space with almost zero variance with respect to the second-order dispersion [it lies on the line $\mathrm{Re}[\psi_{\tau\tau}] \approx 0$ in Fig. 3.4(b-iii)]. At the same time, this blue cluster forms a straight line with a non-zero slope, showing the correlation between $i\psi_\xi$ and $|\psi|^2\psi$ in Fig. 3.4(b-ii).

Although not explicitly shown in Figs. 3.4(b-ii) and (b-iii), we can recognise that the blue cluster is quite dense and contains more dynamical points than the orange cluster [see Fig. 3.4(a-ii)]. Contrary to that, the dynamical points assigned to the orange cluster (full NLSE dynamics) are scattered throughout the equation space without having a reduced variance with respect to any directions. It can also be noted that these dynamical points are much less dense, resulting in fewer points in the spatio-temporal map [see Fig. 3.4(a-ii)].

[6]Note that the initial GMM clustering step of the dominant balance search algorithm is not shown here. Figure 3.4(b) only displays the final dominant balance models identified in the equation space.

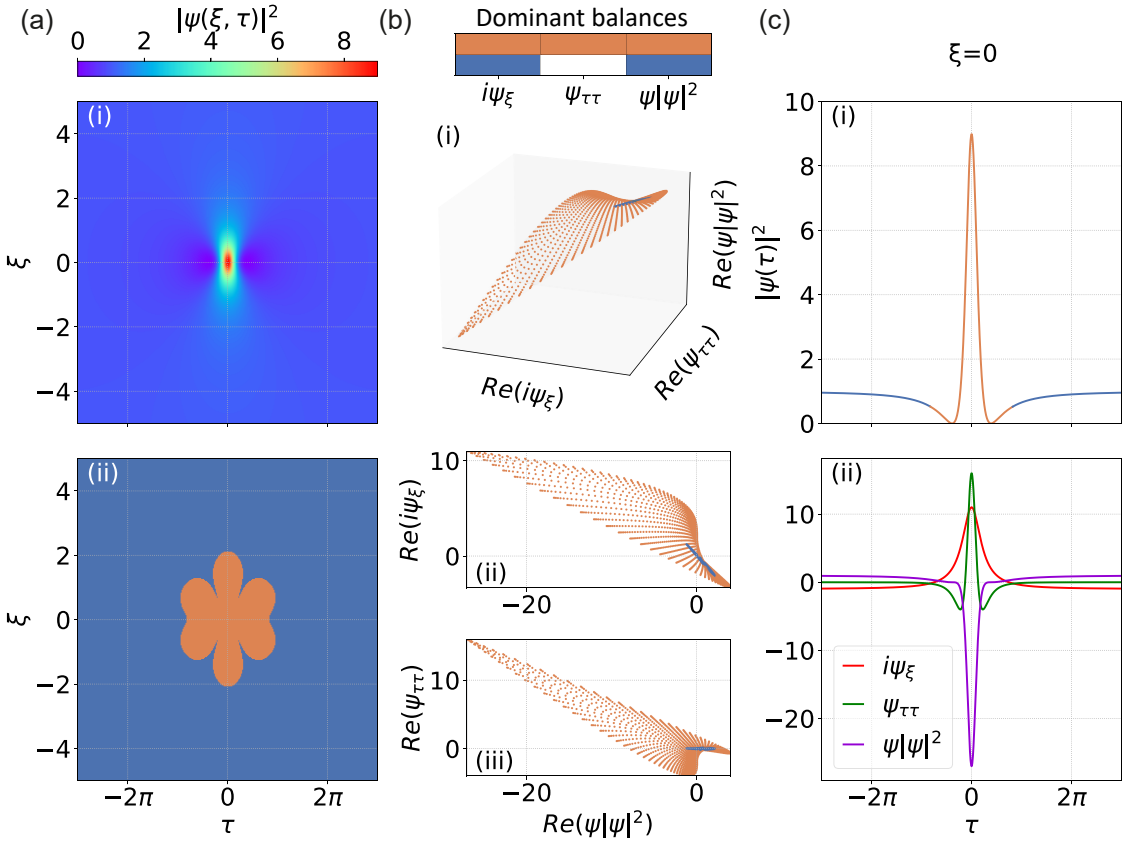


Figure 3.4: Application of the dominant balance approach to the PS dynamics. **(a-i)** Spatio-temporal dynamics represented by $|\psi(\xi,\tau)|^2$. **(a-ii)** Segmented spatio-temporal map showing two identified dominant balance regions: dominated by nonlinearity (blue) and associated with the full NLSE dynamics (orange). **(b)** using the same color code plots the identified dominant balance models in the equation space coordinates associated: (i) with real components $\{i\psi_\xi, \psi_{\tau\tau}, \psi|\psi|^2\}$; (ii) with real components $\{\psi|\psi|^2, i\psi_\xi\}$; (iii) with real components $\{\psi|\psi|^2, \psi_{\tau\tau}\}$. In **(c)**, using identical color code, (i) shows the intensity profile at $\xi = 0$; (ii) shows the individual contributions of the NLSE terms at $\xi = 0$ (see text for details).

To further illustrate the results, Fig. 3.4(c-i) additionally shows the temporal envelope of the PS at $\xi = 0$, where the colors represent the two dominant balance regions as before. At this point of maximum compression of the PS (at $\xi = 0$), the terms of the differential equation $i\psi_\xi$, $\psi_{\tau\tau}$, and $|\psi|^2\psi$ are real-valued functions of τ and can be directly plotted together to illustrate their relative contributions [see Fig. 3.4(c-ii)][7]. We can clearly see from Fig. 3.4(c-i) that the relative contributions of differential equation terms vary significantly across τ. Specifically, in the central part of the temporal envelope, all NLSE terms contribute comparably to the dynamics driving the localization of the PS. Contrary to that, the temporal wings of the PS see the contribution of propagation and nonlinear terms, while the second-order dispersion contribution is effectively zero. We see how this particular region of the temporal envelope is "balanced" by $i\psi_\xi$ and $|\psi|^2\psi$ terms, leading to the dominance of nonlinearity in the background (blue region).

[7]Note that by definition all three contributions must sum to zero at each temporal point as long as the equation $i\psi_\xi + \psi_{\tau\tau} + |\psi|^2\psi = 0$ is satisfied.

Next, we consider the spatio-temporal dynamics of the Akhmediev breather (AB) (another exact analytic SFB solution) that is known to be associated with the initial stage of the MI propagation (see Section 1.3 in Chapter 1). Setting the parameter of the SFB solution to be $a < 1/2$ in Eq. 1.57 results in the τ-periodic spatio-temporal dynamics with ω_m and b playing the role of modulation frequency and exponential growth rate, respectively.

Figure 3.5(a-i) shows the spatio-temporal AB dynamics obtained for $a = 1/4$ (see Eq. 1.57), while Fig. 3.5(a-ii) again shows the two identified dominant balance models in the spatio-temporal domain as before. The blue regions indicating the dominance of nonlinearity are clearly separated, as it was in the case of PS dynamics considered earlier. At the same time, we also see how the contribution of all NLSE terms (orange) leads to the spatio-temporal localisation of the AB. Figures 3.5 (c-i) and (c-ii) again display the relative contributions of the NLSE terms at $\xi = 0$. This clearly illustrates the role of different NLSE terms in driving the dynamics. In the case of the AB solution, it is particularly interesting how the dynamics of the lower amplitude regions between the localised peaks is associated with the balance between all the NLSE terms.

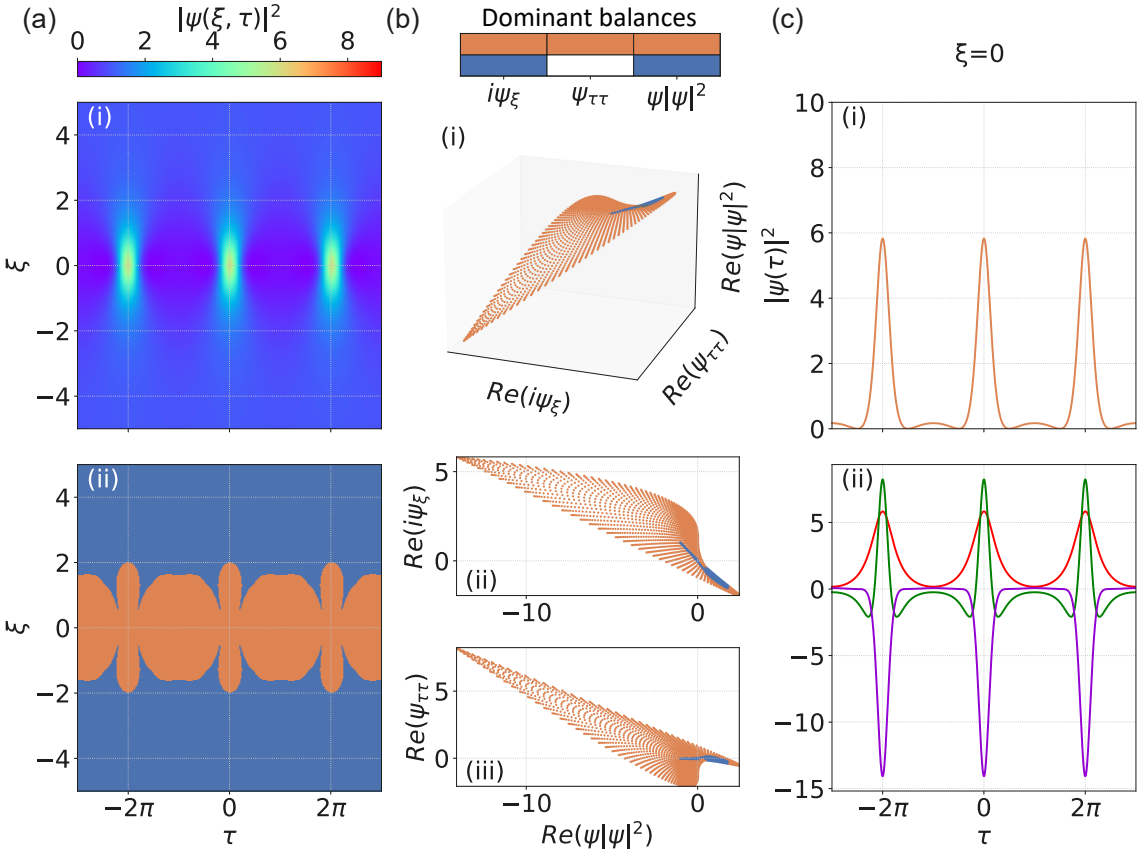


Figure 3.5: Application of the dominant balance approach to the AB dynamics. **(a-i)** Spatio-temporal dynamics represented by $|\psi(\xi,\tau)|^2$. (a-ii) Color segmentation of the spatio-temporal map showing two identified dominant balance regions: dominated by nonlinear and propagation terms (blue) and associated with full NLSE dynamics (orange). **(b)** using the same color scheme plots the identified dominant balance models in the equation space coordinates associated: (i) with real components $\{i\psi_\xi, \psi_{\tau\tau}, \psi|\psi|^2\}$; (ii) with real components $\{\psi|\psi|^2, i\psi_\xi\}$; (iii) with real components $\{\psi|\psi|^2, \psi_{\tau\tau}\}$. In **(c)**, using identical color code, **(i)** shows the intensity profile at $\xi = 0$; **(ii)** shows the individual contributions of the NLSE terms at $\xi = 0$, as indicated in the legend of Fig. 3.4.

Finally, we consider the spatio-temporal dynamics of Kuznetsov-Ma soliton (KM) [see Fig. 1.2(c)] representing the ξ-periodic SFB solution that can be obtained by setting $a > 1/2$ in Eq. 1.57. The results of the dominant balance analysis of the KM dynamics obtained for $a = 1$ are presented in Fig. 3.6 and have similar interpretation. From the comparison of Figs. 3.6 and 3.5, it can be seen how the dominant balance approach can readily distinguish the difference in the periodicity of AB and KM solutions.

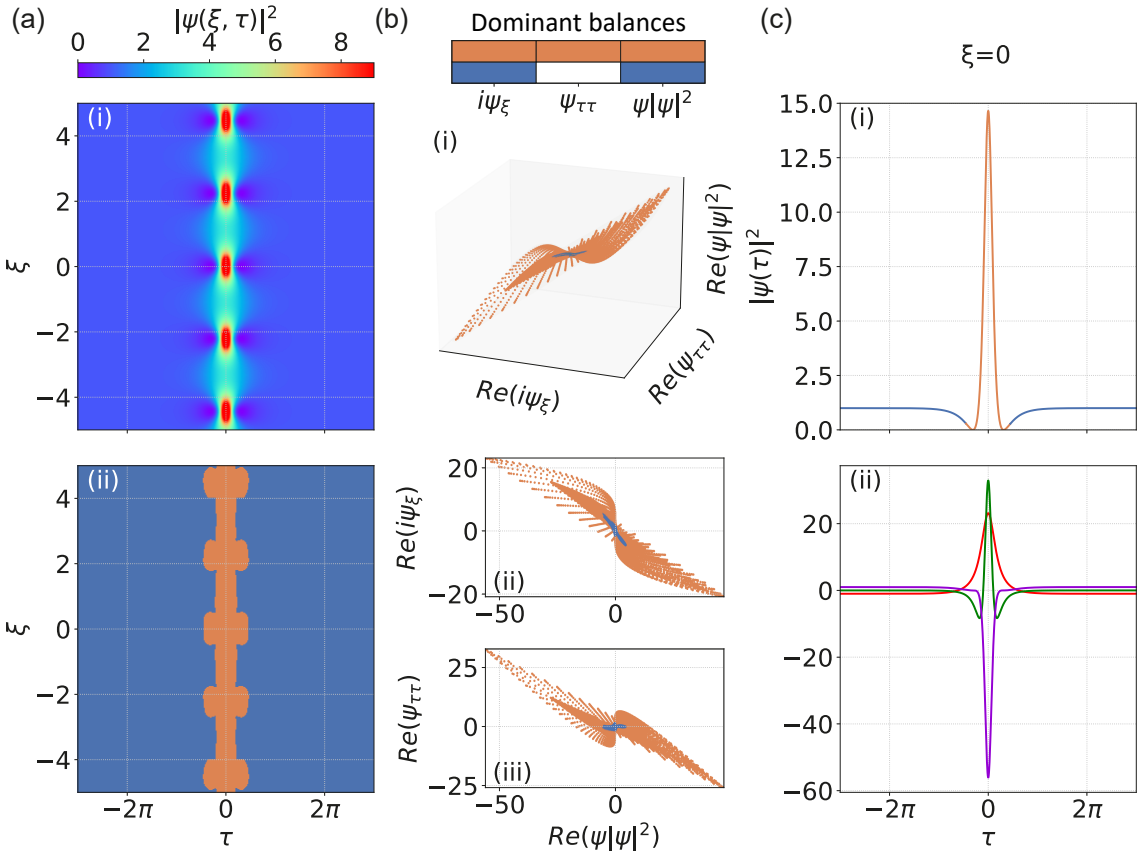


Figure 3.6: Application of the dominant balance approach to the KM dynamics. **(a-i)** Spatio-temporal dynamics represented by $|\psi(\xi,\tau)|^2$. **(a-ii)** Color segmentation of the spatio-temporal map showing two identified dominant balance regions: dominated by nonlinear and propagation terms (blue) and associated with full NLSE dynamics (orange). **(b)** using the same color scheme plots the identified dominant balance models in the equation space coordinates associated: **(i)** with real components $\{i\psi_\xi, \psi_{\tau\tau}, \psi|\psi|^2\}$; (ii) with real components $\{\psi|\psi|^2, i\psi_\xi\}$; (iii) with real components $\{\psi|\psi|^2, \psi_{\tau\tau}\}$. In **(c)**, using identical color code, **(i)** shows the intensity profile at $\xi = 0$; **(ii)** shows the individual contributs of the NLSE terms at $\xi = 0$, as indicated in the legend of Fig. 3.4.

The ability of the dominant balance algorithm to clearly separate the spatio-temporal dynamics of the SFB is crucial for any further attempts to apply the technique to the case of the noise-driven MI (see Section 3.3.2). The identified dominant balance dynamics of the SFB are somewhat expected from the structure of the analytical solutions. Yet it clearly illustrates the capacity of the data-driven dominant balance approach to account for the local relative contributions of the differential equation terms. Specifically, we can see how the two physical processes occurring at different dynamical scales (namely, the soliton localization and the plane wave background) can be readily identified.

3.3.2/ Dominant balance analysis of noise-driven MI

We now proceed to apply the dominant balance approach to interpret the more complex dynamics of noise-driven MI. For this purpose, the NLSE is solved numerically using the split-step Fourier method for a continuous wave input with a superimposed broadband low-level noise. In particular, the common optical one-photon-per-mode noise model is used here [54]. The chaotic MI dynamics can be observed for almost any class of random amplitude and/or phase fluctuations [22] (see Section 1.3), however the one-photon-per-mode noise model is known to yield the MI propagation that reproduce well the experiments [49]. Figures 3.7 (a) and (b) show the results of the numerical simulations for both the spatio-temporal and spatio-spectral domains [43].

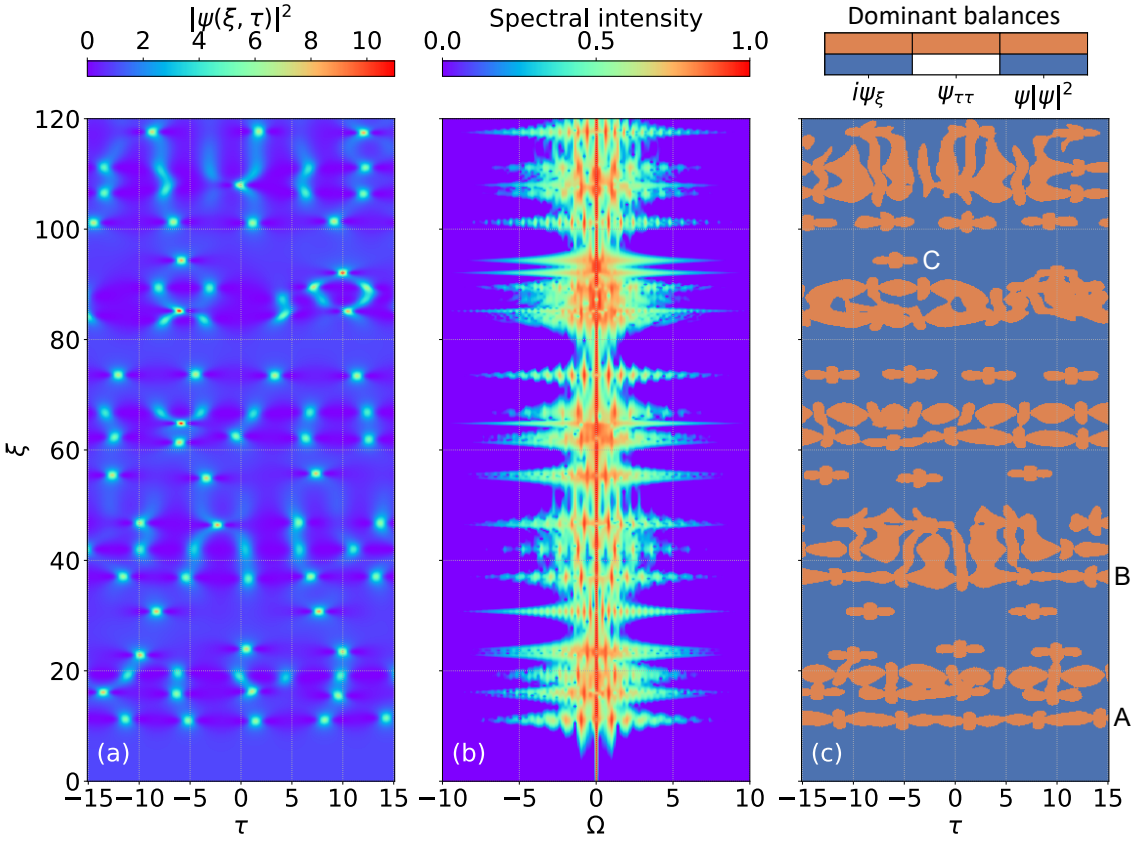


Figure 3.7: **(a)** Spatio-temporal evolution of noise-driven MI represented by $|\psi(\xi,\tau)|^2$. **(b)** Corresponding spectral dynamics. **(c)** Results of a dominant balance analysis that automatically identifies distinct regions of dominant balance in the spatio-temporal domain. The same color code is used here as in the previous figures in this section. Labels A, B, and C indicate several characteristic localized structures arising in the MI propagation (see text for details).

The results show how an injected continuous wave with superimposed random noise evolves into a series of localised structures in the spatio-temporal domain around $\xi \sim 10$. The initial stage of the propagation is associated with the formation of the characteristic MI sidebands centered at $\Omega = \pm 1$ in the spectral domain. This dynamics leads to the emergence of the AB characterized by a temporal periodicity of $\Delta\tau \approx 2\pi$. The later propagation is associated with the spontaneous emergence of various localised structures with

both temporal (transverse) and spatial (longitudinal) periodicity. The propagation is plotted up to $\xi = 120$. We can also see how the described temporal evolution is reflected in the frequency domain with chaotic broadening and narrowing of the spectrum. In fact, the random appearance of high-intensity ultrashort temporal peaks is associated with the broadening of the spectrum in such a way that a decrease in the temporal duration of localised peaks leads to an increase in the spectral bandwidth.

Figure 3.7(c) shows the results of the dominant balance approach applied to the spatio-temporal dynamics shown in Fig. 3.7(a). Again two different dominant balance models are identified by the algorithm: the one associated with the dominant nonlinear propagation (blue regions) and the one corresponding to the full NLSE dynamics (orange regions). This analysis enables us to identify the emerging localised structures that can be compared with the analytical SFB solutions analysed in Section 3.3.1. In particular, the dominant balance approach effectively identifies the AB structures with the temporal period of $\Delta\tau \approx 2\pi$ [for example, at $\xi \approx 11$ and $\xi \approx 38$, see labels A and B in Figure 3.7(c)]. It also reveals various ξ-periodic and PS-like rogue wave structures, such as the distinct feature in Fig. 3.7(c) at $\xi \approx 93$ [see label C]. Moreover, we can again see how the dominant balance approach associates the low intensity background region with the nonlinear dominant balance (for example, in the initial stage of the propagation $\xi < 10$). This visualization method distinctly highlights the well-known "nonlinear" stage of the MI associated with exponential gain derived from the linear stability analysis (see Section 1.3), and before any recurrence is observed [145, 153].

This result is particularly valuable because it demonstrates the ability of dominant balance to provide sometimes seemingly counterintuitive insights into the observed physics. Indeed, from a naive point of view one may assume that nonlinearity (blue dominant balance model) should dominate in spatio-temporal regions of relatively high intensity. Yet we see how the nonlinearity dominates the regions of relatively low intensity. This emphasizes the necessity to evaluate the relative contributions of dispersion and nonlinearity when interpreting NLSE dynamics. In particular, the absence of τ-structure in a plane wave background means that it practically does not see the dispersion, allowing only nonlinear self-focusing to dominate the background's evolution. The interaction between dispersion and nonlinearity begins only after the development of temporal structure from this initial nonlinear stage ($\xi > 10$).

Finally, from the algorithmic side, we can clearly see that by using the unsupervised GMM clustering along with the SPCA model selection procedure on the equation space dynamics instead of a simple threshold-based approaches, this technique can accurately identify evolving localised structures even in low-intensity regions. This result demonstrates the potential extension of this method to automatically detect emerging rogue wave structures [154].

3.4/ Summary, discussion & conclusion

These results clearly demonstrate the power of the dominant balance approach in studying the interaction between dispersion and nonlinearity in the context of MI dynamics. The combination of unsupervised GMM clustering, which can find hidden patterns in the equation space dynamics, and the SPCA model selection algorithm, which returns sparse, physically-interpretable models, can reveal dominant physical interactions even within turbulent NLSE dynamics. The algorithm can automate the process of finding dominant balance models in many complex optical systems and could, for example, be extended in the future to study the dynamics of ultrafast fibre lasers (see Section 1.5). Furthermore, it can be used to simplify complex numerical integration schemes by introducing adaptive simplified numerical strategies in local dominant balance regions.

However, there are challenges to be addressed before these further developments. First of all, the GMM + SPCA approach described above still requires some degree of expert supervision. In some cases it can be quite unclear how to select an optimal value of the l_1 regression hyperparameter α (see Section 3.2.4), especially with increasing number of differential equation terms K and GMM clusters n_c. Moreover, introducing a general α for all GMM clusters can be challenging as the algorithm tends to return almost "empty" dominant balance models for clusters with smaller values of the covariance matrix. As a result, the algorithm may partially ignore the local importance of the differential equation terms. Recently, it was proposed to perform the SPCA selection algorithm for each of the GMM clusters separately [136]. This may lead to better dominant balance model selection procedure, however at the expense of computational cost. Moreover, this approach again increases the degree of expert "supervision" over the results of analyses.

The results shown in this chapter were obtained by relying on physical intuition and theoretical knowledge when tuning the hyperparameters. To automatically select the best-fitting dominant balance models, a greedy search over a discrete number of GMM clusters n_c and continuous l_1 sparsity parameters α must be performed. This would result in a significant increase in computational complexity close to those corresponding to "black box" ML models, but may possibly eliminate the expert supervision in the dominant balance search. Finally, in order to perform such a search, it is necessary to introduce some metrics that can assess the performance of the retrieved dominant balance models. Some of the above issues will be addressed in Chapter 4 of this thesis.

4

Automated unsupervised identification of dominant physical processes in optical fibre pulse propagation

4.1/ Introduction

In this chapter, a fully unsupervised technique for dominant balance search is implemented to study various propagation scenarios in nonlinear fibre optics. In what follows, we first present an extension of the dominant balance search algorithm described in Chapter 3 that automates the model selection procedure. Then, the illustrative examples of both normal and anomalous dispersion regime optical pulse propagation are considered. In addition to identifying the dominant balance in the spatio-temporal domain, as has been done in previously published works, we demonstrate how the dominant balance maps to the corresponding spectral domain.

4.2/ Combinatorial model selection algorithm

As discussed in Chapter 3, the originally proposed dominant balance search algorithm [17] based on the combination of Gaussian mixture modeling (GMM) clustering and sparse principal component analysis (SPCA) model selection still requires a degree of expert supervision when tuning the hyperparameters (see Section 3.4). This may potentially limit the exploratory power of the approach. In machine learning (ML), SPCA turns out to be particularly relevant for problems in which the number of variables is comparable or even exceeds the number of observations [123]. In our case, the number of spatio-temporal points (observations) is much larger than the number of differential equation terms (variables) $K \ll NM$. Indeed, as was highlighted in Chapter 2, most physical systems can be described by differential equation models with just a few active terms, so that K is usually a small number. This motivates the idea of implementing a "combinatorial" search algorithm[1] that involves the use of specific criteria that can help to select the best local dominant balance model from a list of all possible candidate models.

[1] Combinatorial approaches typically involve finding an optimal solution from a finite (discrete) set of problem solutions [3].

Recently such an approach was proposed by Bryan E. Kaiser, Juan A. Saenz, Maike Sonnewald, and Daniel Livescu [18], where it was shown that the SPCA model selection procedure can be effectively replaced with a specific combinatorial search over candidate dominant balance models. In other words, for each local dynamical region associated with a GMM cluster, it is possible to select the best suitable dominant balance from the finite number of candidate dominant balance models. Consequently, this can potentially replace the SPCA model selection algorithm that requires tuning the continuous l_1 hyperparameter α (see Section 3.2.4).

Based on physical considerations, in what follows we restrict ourselves to the analysis of dynamics outside the steady-state regime and assume that the propagation term $i\psi_\xi$ is always active. Given that a dominant balance of just one term is not useful [18], the number of candidate dominant balance models equals to $2^{K-1}-1$. To illustrate this, Fig. 4.1 shows the dominant balance candidate models for the nonlinear Schrödinger equation (NLSE) and generalised NLSE models with $K = 4$ and 5 differential equation terms describing the propagation of ultrashort optical pulses in fibres (see Section 1.4). For example, in the case of NLSE with $K = 3$ we have only 3 candidate dominant balance models, while in the case of generalised NLSE with third order dispersion (TOD) and integral Raman terms we get a total of 15 candidate dominant balance models. From this, one can directly estimate that the complexity of the combinatorial search is $\mathcal{O}(2^K)$. Despite this, for relatively small K, the total number of candidate models remains small enough that the "best-fitting" dominant balance model can be efficiently identified. Specifically, the time complexity analysis reported in [18] showed that the combinatorial model selection algorithm can be successfully applied to governing differential equation with $K \leq 8$ terms.

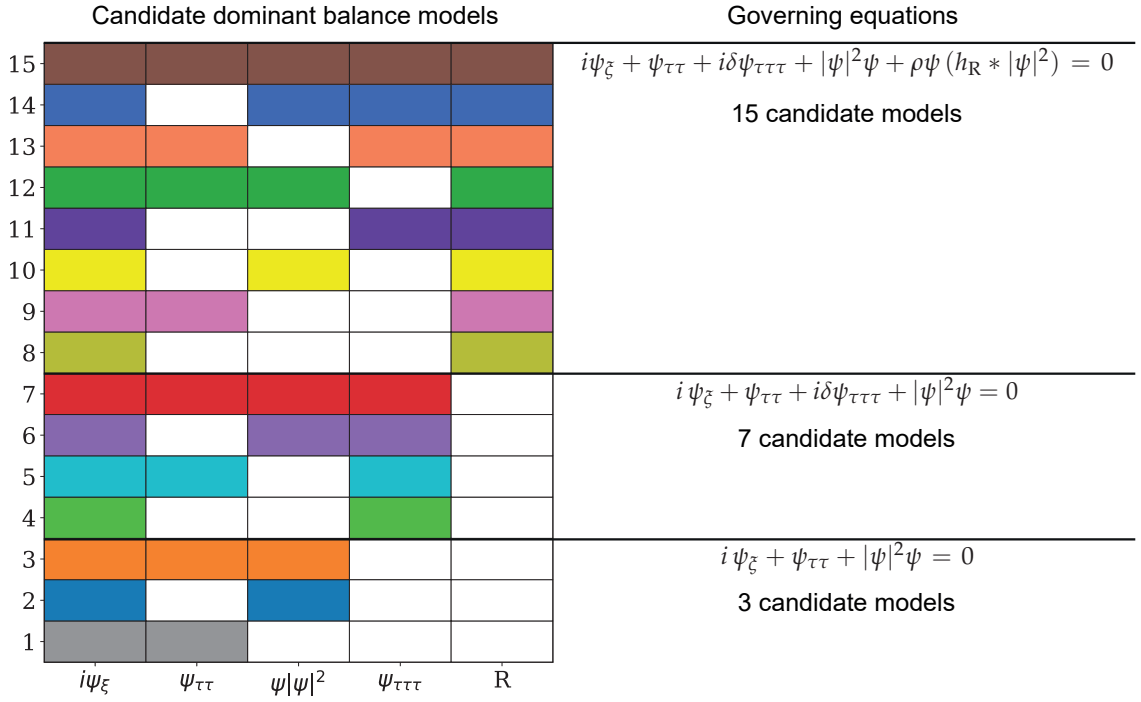


Figure 4.1: Candidate dominant balance models corresponding to the NLSE and the generalised NLSE propagation models with $K = 4$ and 5 differential equation terms (see Section 1.4). In each candidate dominant balance model, the colored boxes indicate active terms, while white boxes indicate inactive terms.

4.2.1/ Dominant balance search criteria

In this section we summarise the approach proposed in [18] that is based on the combinatorial search for the dominant balance model that best describes the dynamics in a given spatio-temporal region. To this end, we need to find the subset of S out of K terms of the governing differential equation that best describes the local dynamics, while assuming the contribution of the remaining $D = K - S$ terms negligible. To do that, one needs to formulate criteria of the dominant balance search. In [18] it was proposed to introduce two principal criteria describing the "best" local dominant balance:

1. The magnitude gap between the selected subset of S dominant terms and the subset of D remaining terms Γ must be maximised;

2. The magnitude gap within the selected subset of S dominant terms Ω must be minimised;

Now we proceed to define these criteria via a computable metric [18]. To this end, we first introduce the normalised equation space matrix $\hat{\mathbf{\Lambda}}$, each row of which $\hat{\mathbf{\Lambda}}_{j:}$ is normalised with respect to the smallest non-zero value in the corresponding row:

$$\hat{\mathbf{\Lambda}}_{j:} = \frac{|\mathbf{\Lambda}_{j:}|}{\min_{k\in\mathbf{F}}(|\Lambda_{jk}|)} , \tag{4.1}$$

where $j \in [1, ..., NM]$, $\mathbf{F} = \{1, ..., K\}$, and $\min_{k\in\mathbf{F}}(|\Lambda_{jk}|) \neq 0$. According to the normalisation, for each element of the normalised equation space matrix condition $\hat{\mathbf{\Lambda}}_{jk} \geq 1$ is satisfied. Next, for each of the candidate dominant balances, we can specify a subset of the selected terms $\mathbf{s}_j \subseteq \hat{\mathbf{\Lambda}}_{j:}$ of length S and a subset of remaining terms $\mathbf{d}_j \subset \hat{\mathbf{\Lambda}}_{j:}$ of length D at each spatio-temporal point.

The order of magnitude metric M describing the degree of dominance of a given subset of terms can be defined as follows, taking into account the aforementioned dominant balance criteria [18]

$$M_j = \frac{\Gamma_j}{1 + \Omega_j} \in [0, 1], \tag{4.2}$$

computed at each point of the spatio-temporal map $\psi(\xi_n, \tau_m)$, i.e. for each row of $\hat{\mathbf{\Lambda}}$, where

$$\Omega_j = \log_{10}[\max(\mathbf{s}_j)] - \log_{10}[\min(\mathbf{s}_j)] \in [0, \infty), \tag{4.3}$$

and

$$\Gamma_j = \begin{cases} \dfrac{\log_{10}[\min(\mathbf{s}_j) - \max(\mathbf{d}_j)]}{\log_{10}[\min(\mathbf{s}_j) + \max(\mathbf{d}_j)]}, & \text{if } \min(\mathbf{s}_j) > \max(\mathbf{d}_j) \\ 0, & \text{if } \min(\mathbf{s}_j) \leq \max(\mathbf{d}_j) \end{cases}, \tag{4.4}$$

with additional floor condition $\Gamma = 0$ for $\Gamma < 0$.

Figure 4.2 shows the behaviour of the introduced order of magnitude metrics. First of all, we can see how the introduced metric Γ (blue dashed curve) accounts for the dominant balance criteria (1). As the gap in magnitude between the smallest selected $\min(\mathbf{s}_j)$ and the largest neglected $\max(\mathbf{d}_j)$ terms increases, Γ_j grows rapidly, reaching $\Gamma_j \approx 0.91$ for an order-of-magnitude gap. In particular, Γ is essentially 1 for a gap of two orders of magnitude.

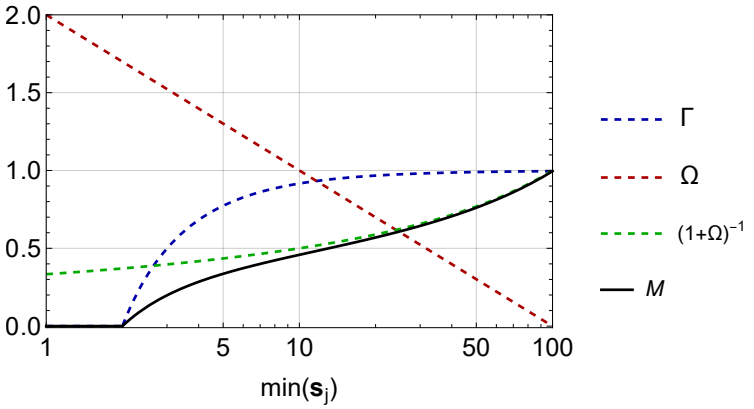


Figure 4.2: The black solid, blue dashed, red dashed and green dashed lines plot the order of magnitude M, Γ, Ω, and $(1+\Omega)^{-1}$ metrics, respectively (see Eqs. 4.2-4.4), computed for fixed $\max(\mathbf{d}_j) = 1$ and $\max(\mathbf{s}_j) = 100$ as a function of $\min(\mathbf{s}_j)$. These metrics are proposed in [18] to identify the "best performing" local dominant balance models.

At the same time, the differential equation terms of the selected dominant subset $\mathbf{s}_j$ should contribute comparably to the local dynamics. To this end, Γ is balanced by Ω that accounts for the dominant balance criteria (2), minimizing the order of magnitude difference in the selected subset. Red dashed curve plots Ω against $\min(\mathbf{s}_j)$ for $\max(\mathbf{s}_j) = 100$. It can be seen that Ω_j decreases linearly on the logarithmic scale as the magnitude gap in the selected subset decreases. This emphasizes the order of magnitude nature of this metric. More specifically, $\Omega = 2$ for a gap of two orders of magnitude and $\Omega \to 0$ as the standard deviation in the selected subset tends to 0.

Finally, the black curve in Fig. 4.2 plots $M_j = \Gamma_j/(1+\Omega_j)$ against $\min(\mathbf{s}_j)$. As both the $\Gamma_j \to 1$ and $\Omega_j \to 0$, $M_j \to 1$ maximizing the score for a given dominant balance. As Γ_j decreases and/or Ω_j increases, the full propagation model (which accounts for the contribution of all terms in a differential equation) will be more favoured, as there will be no clear dominance of a certain subset of terms.

Algorithm 3 presents the main steps of the combinatorial model selection algorithm that can be applied to the identified GMM clusters (see Section 3.2.2).

Algorithm 3 Combinatorial dominant balance model selection algorithm

- Initialize candidate dominant balance models (total of $2^{K-1}-1$ models)
- Compute the normalised equation space matrix $\hat{\mathbf{\Lambda}}$ (see Eq. 4.1)

I. Clusterwise combinatorial model selection
for each GMM cluster C_i **do**
1. Extract the array $\hat{\mathbf{\Lambda}}_i$, representing a subset of $\hat{\mathbf{\Lambda}}$ associated with the cluster C_i
for each candidate dominant balance model p **do**
2. Compute Γ_j^p, Ω_j^p and M_j^p at each spatio-temporal point of $\hat{\mathbf{\Lambda}}_i$, $j \in [1, ..., NM]$ (see Eqs. 4.2, 4.3, 4.4)
3. Estimate the cluster-averaged metric $\overline{M}_{C_i}^p = N_i^{-1} \sum_j^{N_i} M_j^p$
end for
4. Assign a cluster C_i to the dominant balance p with the highest $\overline{M}_{C_i}^p$ score
end for
II. Group the GMM clusters assigned to the same dominant balance together

For each of the GMM clusters C_i one can compute the cluster-averaged score associated with a candidate dominant balance p. If we denote $\hat{\mathbf{\Lambda}}_i$ a subset array of $\hat{\mathbf{\Lambda}}$ corresponding to the cluster C_i with N_i elements, then the cluster-averaged score M with respect to candidate dominant balance p can be computed as $\overline{M}^p_{C_i} = N_i^{-1} \sum_j^{N_i} M^p_j$. Based on this, each of the identified GMM cluster can be assigned to the dominant balance p from the list of candidate models (see Fig. 4.1) that obtained the highest $\overline{M}^p_{C_i}$. Finally, clusters assigned to the same candidate dominant balance can be grouped together to form the final dominant balance models.

The order of magnitude metric M [18] is certainly not unique and could be defined differently, however Eqs. 4.2-4.4 introduce a simple and physically-consistent algebraic framework for evaluating the success of the dominant balance search. With this, one can directly select the "best performing" local dominant balance model. For example, in the case of NLSE, the algorithm requires computing the order of magnitude metric M 3 times for each of the candidate dominant balances: dispersive $\{i\psi_\xi, \psi_{\tau\tau}\}$, nonlinear $\{i\psi_\xi, |\psi|^2\psi\}$, and full NLSE $\{i\psi_\xi, \psi_{\tau\tau}, |\psi|^2\psi\}$ (see Fig. 4.1). The candidate dominant balance model that obtains the highest M can then be chosen to best represent the local dynamics.

4.2.2/ Riemann wave propagation example

To provide an illustrative example of the dominant balance approach involving the GMM clustering and combinatorial model selection algorithm, below we consider an example of Riemann wave propagation in an optical fibre [155][2]. These dynamics are governed by the NLSE in the normal dispersion regime $i\psi_\xi - \psi_{\tau\tau} + |\psi|^2\psi = 0$ with specific initial condition corresponding to the Riemann wave

$$\psi(0,\tau) = N \exp\left(-\tau^2/4 - i\phi\right), \tag{4.5a}$$

$$\phi = \sqrt{2}N \int_{-\infty}^{\tau} \exp\left(-\tau'^2/4\right) d\tau' = \sqrt{2\pi}N[(1 + \mathrm{erf}(\tau/2)], \tag{4.5b}$$

where the input Gaussian pulse has an imposed chirp profile that is dependent on the pulse amplitude. Note that a different NLSE normalization is used here when compared to that used in Chapter 3. The dimensionless variables are defined as $\xi = z/L_\mathrm{D}$, and $\tau = t\sqrt{2}/T_0$, and $\psi = A\sqrt{\gamma L_\mathrm{D}}$. The notation erf stands for the error function and $N = \sqrt{L_D/L_{NL}} \approx 5.84$ was used in the simulations.

Figure 4.3(a) shows the spatio-temporal evolution of the input chirped Gaussian pulse (see Eq. 4.5) in the normal dispersion regime in an optical fibre leading to the envelope steepening and formation of the optical shock. The physical aspects of Riemann wave propagation along with an interpretation of the dominant balance analysis are given in Section 4.3.1; here we only use this example to illustrate the dominant balance algorithm.

Following the steps of the dominant balance algorithm (see Fig. 3.1 in Chapter 3), we first represent the dynamics in the associated equation space, and then perform the GMM clustering. The results of the GMM cluster analysis in the equation space are presented in Fig. 4.3(c), where we set $n_c = 10$ for the initial number of GMM clusters. Figure 4.3(b) shows the identified GMM clusters mapped back onto the original domain segmenting the spatio-temporal map. Note that the GMM clusters in Fig. 4.3 are displayed in different

[2]Note that in this section we present the analysis of the dominant balance procedure in detail and we do not place a strong emphasis on physical interpretation. We present summary results of the Riemann wave case again in Section 4.3.1 where we also discuss the physics in more detail.

colors for better visualisation, but are not yet assigned to any of the candidate dominant balance models. The white region in Fig. 4.3(b) indicates the low intensity region of the spatio-temporal map (below −30 dB in this case) which is not taken into account by the algorithm.

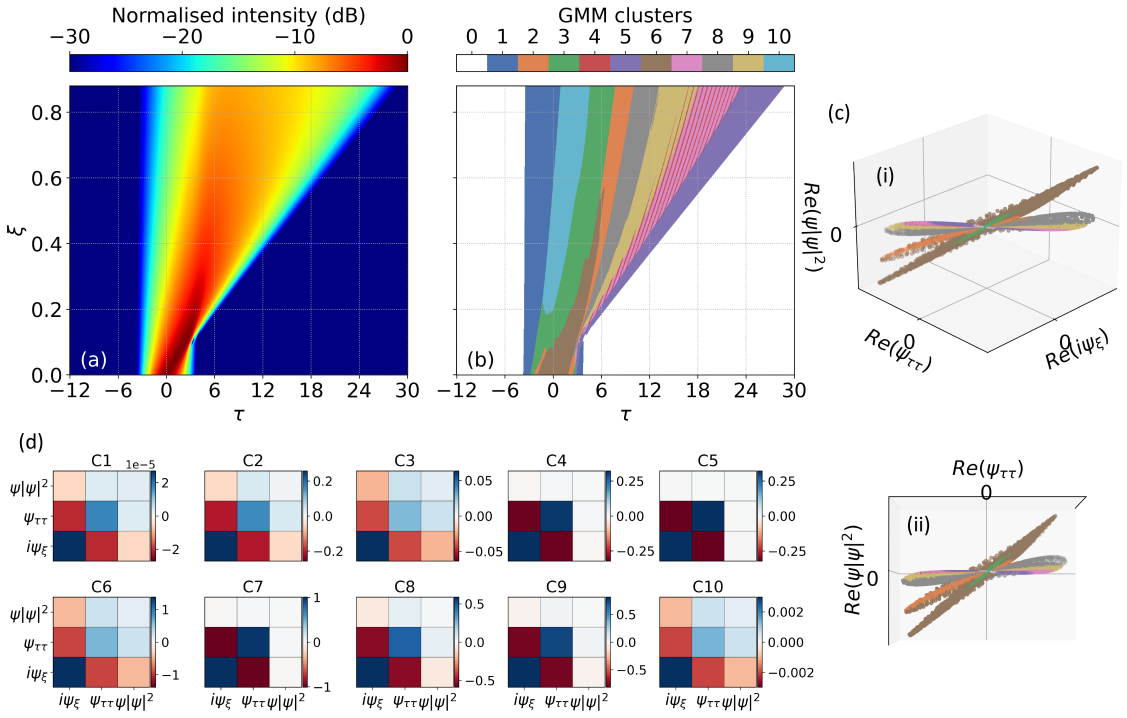


Figure 4.3: The unsupervised GMM clustering applied to the equation space dynamics of Riemann wave propagation. **(a)** Normalized spatio-temporal dynamics $|\psi(\xi,\tau)|^2$ plotted in dB. **(b)** The 10 identified GMM equation space clusters (represented by different colors) mapped onto the spatio-temporal domain. **(c-i)** using the same color code, shows the identified GMM clusters in the equation space plotted in coordinates associated with real parts of NLSE terms $\{i\psi_\xi, \psi_{\tau\tau}, \psi|\psi|^2\}$. Note that several clusters might not be well visually distinguishable due to their low variance and close overlap. In **(c-ii)** one of the projections of the segmented equation space is shown using the same color code. **(d)** displays the covariance matrices of the 10 identified GMM clusters.

As for the case of the PS dynamics (see Section 3.2.3), we can start by looking at the learnt covariance matrices of the GMM clusters [see Fig. 4.3(d)]. From this, one can directly see the dynamical patterns hidden in their structure. For example, clusters C_4, C_5, C_7 and C_9 (red, violet, pink and yellow, respectively) have a very similar covariance structure, showing a strong correlation between the propagation and dispersion terms $\{i\psi_\xi, \psi_{\tau\tau}\}$. At the same time, the covariance matrices have almost zero entries with respect to the nonlinear Kerr term. We can also see how these clusters are closely grouped in the equation space being restricted to a line with almost zero variance with respect to the Kerr term [red, violet, pink, and yellow clusters, respectively in Fig. 4.3(c)].

Clusters C_2, C_3, C_6, and C_{10} (orange, green, brown and light blue, respectively) show quite different dynamics, having a non-zero covariance between all terms in the NLSE. It can be seen how these clusters are also grouped together, while being orientated differently in the equation space showing no reduced variance with respect to any of the directions.

From the presented analysis, we can already gain some intuition regarding dominant balance models in the dynamics. In particular, clusters C_4, C_5, C_7 and C_9 (red, violet, pink and yellow, respectively) are likely to be associated with the dominant balance between the propagation term and the second-order dispersion. Clusters C_2, C_3, C_6 and C_{10} (orange, green, brown and light blue, respectively) can be attributed to the full NLSE dynamics based on their covariance structure.

Indeed, we can try to identify some dynamical patterns in the resulting GMM clusters "by ey", however there will still be some uncertainty in it. For example, it is not clear whether cluster C_8 should be associated with the dispersive or full NLSE dynamics based on its covariance structure. This uncertainty can be removed by implementing a model selection procedure at this step.

Figure 4.4 shows the results of the dominant balance analysis via combinatorial model selection algorithm (see Algorithm 3) for the Riemann wave propagation scenario.

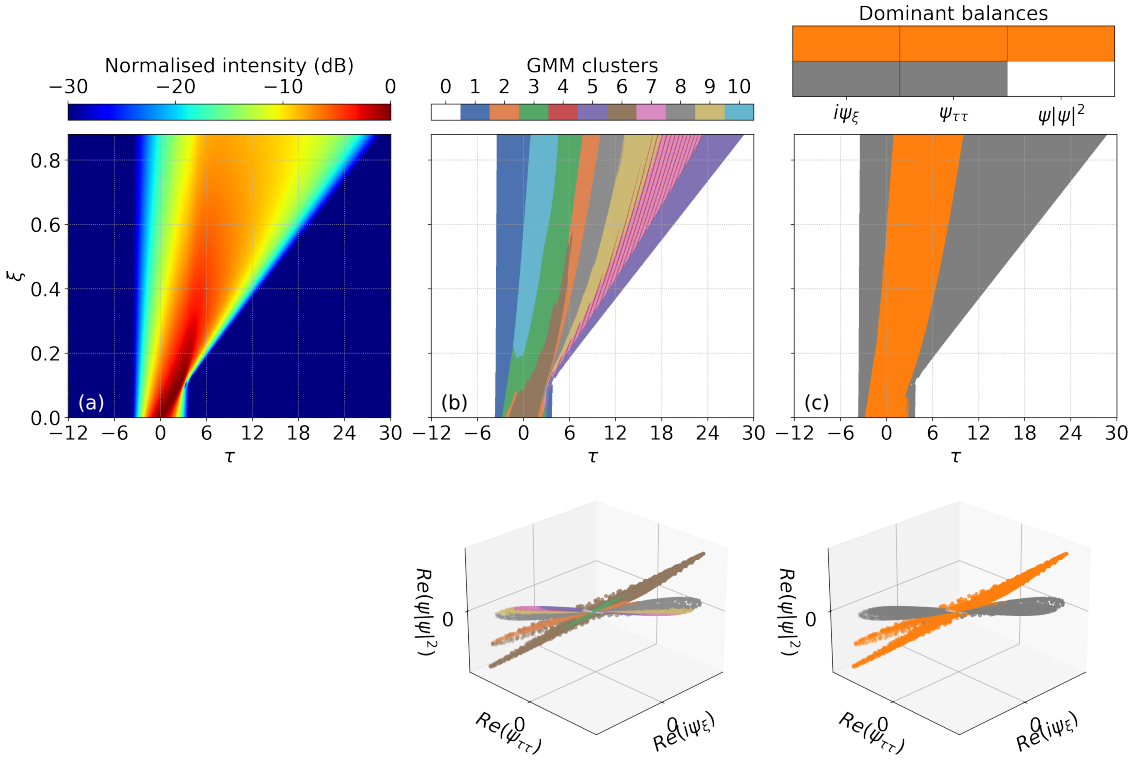


Figure 4.4: Combinatorial model selection algorithm applied to the GMM clusters identified for the case of Riemann wave propagation. **(a)** Normalized spatio-temporal dynamics represented by $|\psi(\xi,\tau)|^2$ (plotted in dB). **(b)** the 10 identified GMM equation space clusters (represented by different colors) mapped onto the spatio-temporal domain. **(c)** shows the dominant balance regions identified via combinatorial model selection algorithm (see Algorithm 3). The two dominant balance models are identified: the dispersive dominant balance between the propagation and dispersion terms $\{i\psi_\xi, \psi_{\tau\tau}\}$ (grey) and the full NLSE propagation model, where all the three terms $\{i\psi_\xi, \psi_{\tau\tau}, \psi|\psi|^2\}$ contribute comparably (orange). Bottom panel in subfigure **(b)** shows the identified equation space GMM clusters (see Fig. 4.3). Bottom panel in subfigure **(c)** using the same color code shows the identified groups of dynamical points associated with two dominant balance models plotted in the equation space (grey and orange, respectively).

From the bottom panels in Figs. 4.4(b) and (c) we can directly see how the algorithm automatically assigns different GMM clusters to the candidate dominant balance models. In particular, two dominant balance models were identified: dispersive dominant balance model $\{i\psi_\xi, \psi_{\tau\tau}\}$ (grey) and the full NLSE dominant balance model $\{i\psi_\xi, \psi_{\tau\tau}, |\psi|^2\psi\}$ (orange). The full NLSE dominant balance model is formed from clusters C_2, C_3, C_6 and C_{10}, while the remaining clusters are associated with the dominant balance between the propagation term and the second-order dispersion.

These results are consistent with the analysis of the GMM clusters covariance matrices. Indeed, we see how clusters C_4, C_5, C_7 and C_9 share similar covariance structure with almost zero correlation with respect to the Kerr nonlinearity. If we now look at Fig. 4.4(e), we can see a clear separation of the equation space dynamics into two distinct dynamical groups (grey and orange, respectively). This indicates the presence of two distinct physical regimes in the dynamics, which were successfully identified by the algorithm.

Although not strongly visually distinct, there is also a group of white dynamical points near the origin of the equation space in Fig. 4.4(e). These points correspond to the dynamics in the low-intensity region of the spatio-temporal map (below -30 dB) that were discarded.

Figure 4.4(c) shows the projections of the identified dominant balances to the original spatio-temporal domain. Together with the results of the dominant balance analysis, we can now see how we can interpret the observed physics. Specifically, the dominant balance analysis shows a clear separation of the dynamics into two regimes: (i) the leading edge of the pulse experiences the action of both nonlinearity and second order dispersion, and (ii) the broad lower-intensity background appearing at the later stage of the propagation is associated with the dispersive $\{i\psi_\xi, \psi_{\tau\tau}\}$ dominant balance.

4.2.3/ Uncertainties in the dominant balance models

As pointed out in the original work [17], the idea of a dominant balance in the presence of multiple physical processes is not necessarily well-defined outside of limiting asymptotic regimes. In fact, it is clear that all the terms of a governing model may have a non-zero contributions across the entire spatio-temporal domain. In [17], the "uncertainty" of identified dominant balance models was discussed from the perspective of the results returned by the GMM algorithm which is known to be a probabilistic clustering method. An alternative approach was proposed in [18], where the computed order of magnitude metric M averaged across the domain was used to estimate how well a particular dominant balance model represents the associated local dynamics. In what follows, considering the case of the Riemann wave propagation, we discuss how this uncertainty can be estimated using the dominant balance criteria introduced in this chapter.

The uncertainty of the dominant balance models is expected to be higher for GMM clusters located near a "border" dividing one dynamical regime from another (in the equation space or equivalently in the spatio-temporal domain). Therefore, in the considered example of Riemann wave propagation (see Fig. 4.4), special attention should be paid when associating the clusters C_1, C_8 and C_{10} with specific dominant balance models. Cluster C_1, however, does not play a significant role in the dynamics as it has a relatively low variance in the equation space and is describing the low-intensity region of the propagation[3]. In principle, in certain cases, we can neglect the dynamics associated with cluster C_1 by introducing an additional variance threshold. This is effectively embedded in the SPCA model selection

[3]Equivalently, it is centred near the origin of the equation space.

algorithm [17]. Indeed, as we select higher values of the l_1 regression hyperparameter, the SPCA algorithm will tend to neglect the dynamics of the cluster C_1 due to its small variance (several orders of magnitude smaller than for the most of the remaining clusters). In other words, the SPCA model selection algorithm will tend to return "empt" models for GMM clusters with relatively small variance as one increases the α hyperparameter.

In some specific propagation scenarios, GMM clusters with such "near-zero" dynamics are, indeed, not physically important and, thus, can be omitted. Nevertheless, it is sometimes undesirable to neglect the dynamics in substantially non-zero intensity regions, especially when the evolution involves dynamics at different scales. The combinatorial model selection algorithm, on the other hand, does not neglect GMM clusters with relatively low variance. This ensures that the dynamical information is not lost while the variance threshold can be set in the later step if necessary.

Unlike the C_1 cluster, the C_8 cluster has a rather large variance and must be accounted for. We can assess the accuracy of a given dominant balance model when describing the local dynamics based on the proposed metric M. This allows us to evaluate the efficiency of the dominant balance algorithm. The outputs of the combinatorial model selection algorithm for 4 GMM clusters C_1, C_5, C_8, and C_{10} are shown in Table 4.1.

Table 4.1: Output of the combinatorial model selection algorithm (see Algorithm 3) showing the averaged order-of-magnitude M metrics for 4 selected GMM equation space clusters learnt for the Riemann wave propagation example (columns 2–4). Columns 5 and 6 show the order of magnitude estimation of the associated covariance matrices and the number of dynamical points for each of the 4 selected GMM clusters, respectively.

GMM clusters	$\overline{M}_{C_i}^{nonl}$	$\overline{M}_{C_i}^{disp}$	$\overline{M}_{C_i}^{nlse}$	$\mathcal{O}[\mathrm{tr}(\mathbf{\Sigma}_{C_i})/3]$	N_i
C_1	0.00	**0.66**	0.58	$\sim 10^{-5}$	82373
C_5	0.00	**1.00**	0.26	$\sim 10^{-1}$	63257
C_8	0.00	**0.85**	0.50	$\sim 10^{-1}$	47770
C_{10}	0.00	0.10	**0.67**	$\sim 10^{-3}$	61687

Here, the second, third and fourth columns display the computed averaged order of magnitude metrics M for each of the candidate dominant balances ($\overline{M}_{C_i}^{nonl}$, $\overline{M}_{C_i}^{disp}$ and $\overline{M}_{C_i}^{nlse}$ stand for nonlinear, dispersive and full NLSE dominant balance models, respectively). The last two columns show the estimation of the order of magnitude of the covariance matrices $\mathcal{O}[\mathrm{tr}(\mathbf{\Sigma}_{C_i})/3]$ and the number of dynamical points N_i in each cluster C_i.[4]

We see how different GMM clusters receive different scores for belonging to one or another dominant balance model (see columns 2–4 of Table 4.1). None of the clusters are dominantly nonlinear. With high accuracy, the dynamics of clusters C_5 and C_8 are dispersion-dominated, and their variances in equation space are relatively large (see column 5 of Table 4.1). In contrast, the C_{10} cluster is dominated by the full NLSE, but its equation space variance is smaller. Finally, for the C_1 cluster, the estimates of $\overline{M}_{C_i}^{disp}$ and $\overline{M}_{C_i}^{nlse}$ are quite close to each other, with $\overline{M}_{C_i}^{disp}$ being slightly higher. Again we see that the C_1 cluster is the least important from the dynamical point of view due to its low variance. For more detailed analysis we can display the full spatio-temporal map showing the calculated M scores for all the GMM clusters (see Fig. 4.5).

[4]Notation $\mathrm{tr}(\mathbf{A})$ stands for the trace of a matrix $\mathbf{A}$), i.e. the sum of its diagonal elements.

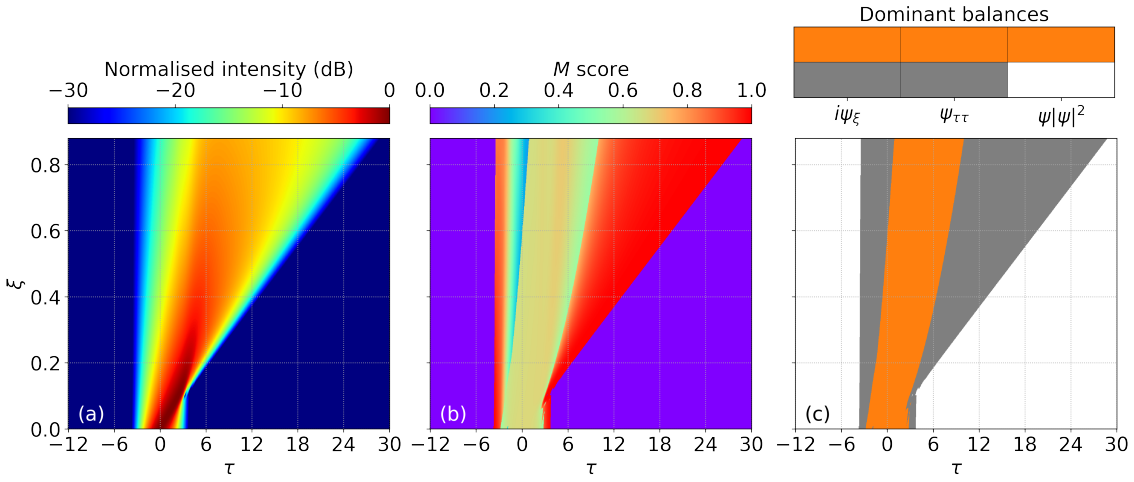


Figure 4.5: Estimation of the local dominant balance models uncertainty. **(a)** Normalized spatio-temporal dynamics $|\psi(\xi,\tau)|^2$ plotted in dB. **(b)** Spatio-temporal distribution of the order of magnitude score computed for the identified dominant balance models. **(c)** shows the dominant balance regions found via combinatorial model selection algorithm (see text for details).

It can be seen how the M score drops significantly near the boundary between the two dominant balance regions. The broad descending edge of the pulse, appearing at $\xi > 0.1$ and obtaining high M scores, clearly indicates the dominance of dispersion in this spatio-temporal region.

4.2.4/ Dominant balance search in spatio-spectral domain

The dynamics of ultrashort pulse propagation governed by the generalised NLSE (see Section 1.4 in Chapter 1) can be as well studied in the spatio-spectral domain. In principle, one can numerically solve the generalised NLSE directly in the frequency domain [156] and then follow the same steps of the dominant balance search algorithm (see Fig. 3.1). At the same time, it can be quite instructive to consider the dominant balance models simultaneously in the spatio-temporal and spatio-spectral domains. To this end, we can solve the generalised NLSE in the spatio-temporal domain and then compute the Fourier transform of each of the time-domain terms, while ensuring that all the estimated terms sum to zero across the spatio-spectral domain

$$\mathcal{F}\left[\sum_{k=1}^{K} f_k(\psi, \psi_\xi, \psi_\tau ..., \psi^2, \psi\psi_\xi, \psi\psi_\tau, ..., \psi_{\xi\xi}, \psi_{\tau\tau}, ...)\right] = 0. \tag{4.6}$$

In practice, errors can accumulate during numerical simulations and when estimating derivatives. Here we emphasize that, in fact, any possible dynamical model can be used as input to the dominant balance search algorithm under the condition that the model terms satisfy the equality (see Eq. 3.1). This emphasises the strength of the approach, as it turns out to be applicable to many models in different areas of science.

4.3/ Dominant balance analysis of optical pulse propagation scenarios

In this section, the described dominant balance search algorithm (see Section 4.2.1) is applied to various scenarios of pulse propagation in normal and anomalous dispersion regimes of optical fibres. For all the considered examples, we provide a physical interpretation of the results of dominant balance analysis in both the spatio-temporal and spatial-spectral domains.

4.3.1/ Riemann wave propagation

We begin by discussing the dynamics of the Riemann wave example considered in Section 4.2.2 and the physics revealed by the dominant balance search algorithm for this case. The initial conditions leading to the Riemann wave dynamics in optical fibre (see Eq. 4.5) are shown in Fig. 4.6(a). With the given input field, the initial stage of the propagation can be well described by the inviscid Burger's equation that we write here using the introduced dimensionless variables (see Section 4.2.2)

$$\frac{\partial|\psi|}{\partial\xi} \pm 3\sqrt{2}|\psi|\frac{\partial|\psi|}{\partial\tau} = 0. \tag{4.7}$$

Equation 4.7 is known to govern the propagation in which the input chirped pulse experiences the envelope steepening that leads to the formation of the characteristic shock associated with the so-called gradient catastrophe that occurs at distance ξ_s [155]. Using the method of characteristics, from Eq. 4.7 one can find the characteristic lines τ_{d_l} representing the energy flow directions [157]. For the case of the Gaussian input pulse considered here (see Eq. 4.5a), we get [155]

$$\tau_{d_l}(\xi,\tau) = \tau_l \pm 3\sqrt{2}N\exp\left(-\tau_l^2/4\right)\xi. \tag{4.8}$$

Figure 4.6(b) shows the characteristic lines governed by Eq. 4.8. The position of the shock ξ_s corresponds to the intersection of the characteristic lines and can be computed analytically using a simple formula. In the case of the Gaussian input pulse (see Eq. 4.5a), $\xi_s = \sqrt{e}/(3N) \approx 0.094$ for $N \approx 5.84$ considered here [green dashed line in Fig. 4.6(b)].

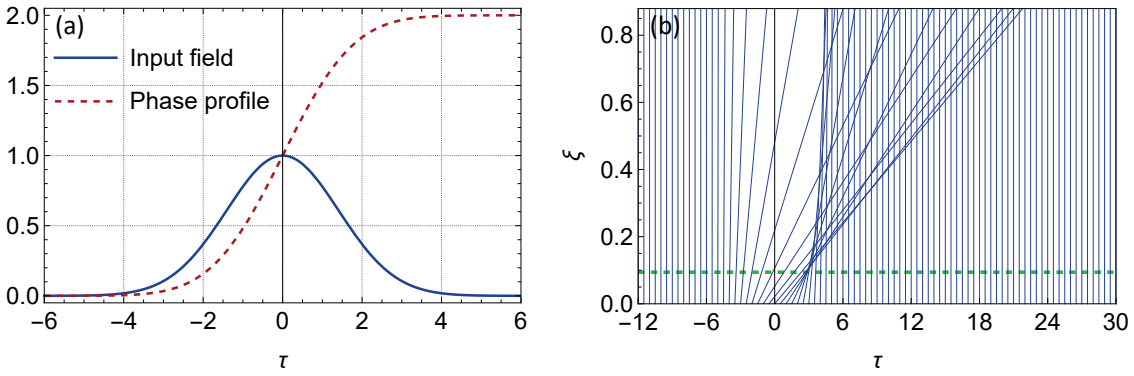


Figure 4.6: **(a)** Initial conditions corresponding to the Riemann wave propagation (see Eq. 4.5). Blue solid line shows the normalized input field amplitude $\psi(0,\tau)/N$. Red dashed line shows the normalized phase profile $\phi(0,\tau)/\sqrt{2\pi}N$. **(b)** displays the characteristic lines defined by Eq. 4.8 and representing the energy flow directions for the given initial conditions (see text for details).

Figures 4.7(a) and (b) show the results of the simulations of the NLSE for the Riemann wave initial conditions (see Eq. 4.5) in both spatio-temporal and spatio-spectral domains.

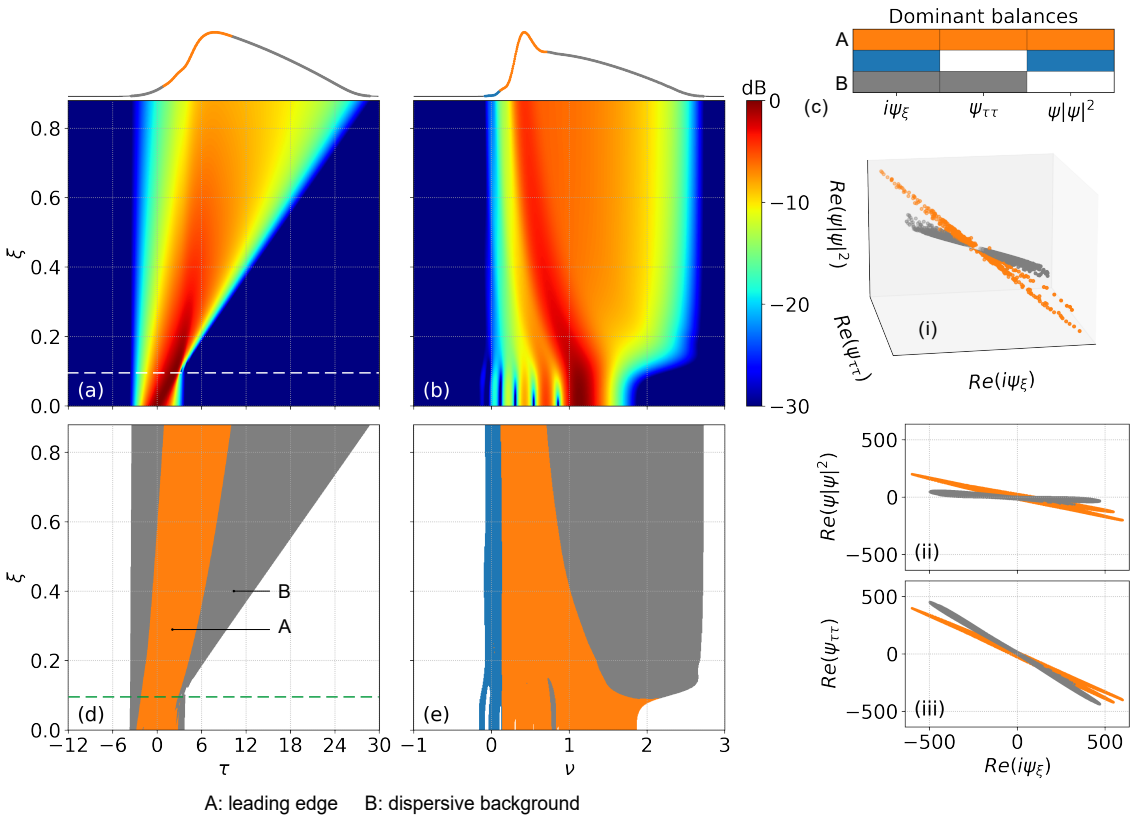


Figure 4.7: **(a)** spatio-temporal and **(b)** spatio-spectral intensity maps of Riemann wave propagation (in both cases the maps are normalised to the maximum intensity and plotted on the same logarithmic scale). The insets at the top show the corresponding output intensity profiles on a linear scale where the colors represent the dominant balance regions (see the inset frame in the top right corner of the figure). In **(c)** **(i)** shows the spatio-temporal evolution dynamics in equation space divided into two identified dominant balance regions. Two projections of the equation space are shown in subfigures **(ii)** and **(iii)**. **(d)** and **(e)** show the identified dominant balance models mapped onto the original domains for comparison with the intensity evolution maps. The same color code is used for the dominant balance models identified in the spatio-spectral domain, while meaning the dominant balances between the Fourier transformed temporal terms shown in the legend. **(d)** labels particular temporal features - A: the steepened pulse leading edge; B: the dispersive background. White and green dashed horizontal lines in subfigures **(a)** and **(d)** indicate the distance at which the shock occurs ($\xi_s \approx 0.094$).

From Fig. 4.7(a) we can see that the initial stage of the evolution is associated with the temporal steepening of the leading edge. At $\xi = \xi_s$ this steepening leads to the formation of the shock. Figures 4.7(d) and (e) show the results of the dominant balance analysis in spatio-temporal and spatio-spectral domains, respectively. In the spatio-temporal domain two different dynamical rigimes are identified: dispersive dominant balance $\{i\psi_\xi, \psi_{\tau\tau}\}$ (grey) and the full NLSE propagation model (orange). From Fig. 4.7(d) we can see how the central part of the pulse and the leading edge remain dominated by the full NLSE

dynamics throughout the propagation [see label A in Fig. 4.7(d)]. In fact, in both the temporal and spectral domains the dynamics before the shock is primarily governed by the NLSE[5]. Indeed, it is known that the shock develops in this case as the result of the simultaneous action of both nonlinearity and dispersion [155].

The later stage of the evolution appearing at $\xi > \xi_s$ shows the signatures of the dispersive propagation [see label B in Fig. 4.7(d)]. While the central part of the pulse envelope and the leading edge continue to be governed by the NLSE, we can also observe the formation of the broad, low intensity background in both temporal and spectral domains. This background is associated with the $\{i\psi_\xi, \psi_{\tau\tau}\}$ dominant balance (grey). The output field profiles plotted on the top of the Figs. 4.7(a) and (b) additionally illustrate the separation between the dominant balance local regions.

As discussed in Section 4.2.2, the presence of two dominant dynamical regimes in the case of Riemann wave propagation is apparent in the equation space representing the spatio-temporal dynamics, where the two distinct dynamical regimes (represented in grey and orange, respectively) are clearly separated [see Fig. 4.7(c)]. The equation space projections shown in Figs. 4.7(c) (ii) and (iii) clearly show the difference in the orientations of the dynamical points associated with the dispersive and full NLSE balance models. Specifically, the dynamical points associated with the dispersive dominant balance (grey) are oriented in the equation space to have near-zero variance with respect to the nonlinear term $|\psi|^2\psi$.

Note that the spatio-spectral field evolution contains an additional nonlinear dominant balance $\{\mathcal{F}(i\psi_\xi), \mathcal{F}(|\psi|^2\psi])\}$ in the vicinity of $\nu = 0$ [blue region in Fig. 4.7(e)]. This result is explained by the fact that the second-order dispersion is proportional to ν^2 which results in its negligible contribution in the local region around $\nu = 0$.

We can see from these results how even at relatively high powers of the input pulse ($N \approx 5.84$) considered here, the initial dynamics is primarily dominated by the full NLSE. In fact, it is the imposed chirp that makes the dispersive contribution to be comparable to the nonlinearity. This dynamics should be qualitatively different from the optical wave breaking propagation where the initial stage of the dynamics is known to be dominated by the SPM (see Section 1.2.2 in Chapter 1).

4.3.2/ Optical wave breaking

Optical wave breaking dynamics are observed in the normal dispersion regime in the case when the SPM dominates the group velocity dispersion (GVD) effect in the initial stage of the propagation ($L_D \gg L_{NL}$). Figures 4.8(a) and (b) show the results of the simulations of the NLSE in the normal dispersion regime $i\psi_\xi - \psi_{\tau\tau} + |\psi|^2\psi = 0$ for the initial conditions corresponding to the input Gaussian pulse

$$\psi(0, \tau) = N \exp(-\tau^2/4), \tag{4.9}$$

with $N = T_0\sqrt{\gamma P_0/|\beta_2|} = 30$. These initial conditions result in the optical wave breaking dynamics in which the second-order dispersion is known to play an important role in temporal pulse shaping at later stages of the propagation [22].

[5]Note that the dispersive dominant balance occupies the low intensity region in this initial stage and, thus, does not play an important role in dynamics.

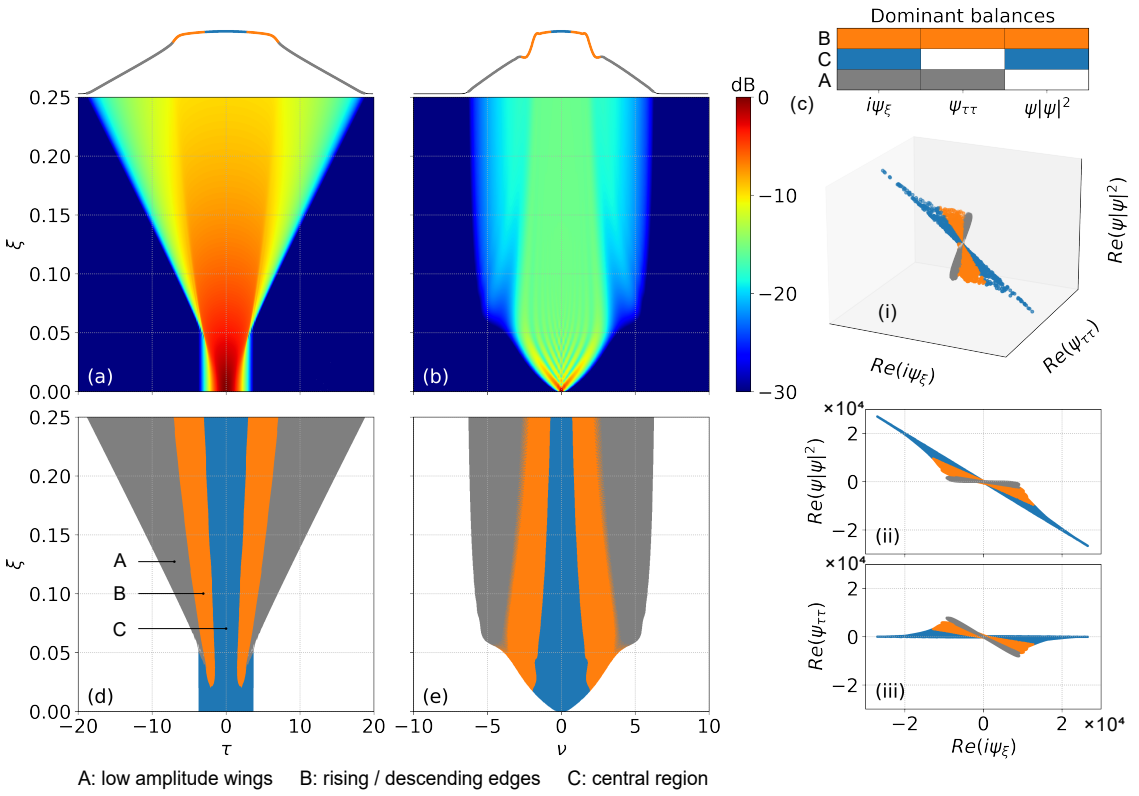


Figure 4.8: **(a)** spatio-temporal and **(b)** spatio-spectral intensity maps of optical wave breaking propagation (in both cases the maps are normalised to the maximum intensity and plotted on the same logarithmic scale). The insets at the top show the corresponding output intensity profiles on a linear scale where the colors represent the dominant balance regions (see the inset frame in the top right corner of the figure). In **(c)** **(i)** shows the spatio-temporal evolution dynamics in equation space divided into three dominant balance regions. Two projections of the equation space are shown in subfigures **(ii)** and **(iii)**. **(d)** and **(e)** show the identified dominant balance models mapped onto the original domains for comparison with the intensity evolution maps. The same color code is used for the dominant balance models identified in the spatio-spectral domain, while meaning the dominant balances between the Fourier transformed temporal terms shown in the legend. **(d)** shows labels for particular temporal features - A: low amplitude wings; B: rising/descending edges; C: highest intensity central region.

Figures 4.8(d) and (e) show the results of the dominant balance analysis in the spatio-temporal and spatio-spectral domains, respectively. Three local balance models are identified in this case: the dispersive dominant balance $\{i\psi_\xi, \psi_{\tau\tau}\}$ (grey), the nonlinear dominant balance $\{i\psi_\xi, |\psi|^2\psi\}$ (blue) and the full NLSE propagation model (orange). Figure 4.8(c) shows the equation space dynamics corresponding to the spatio-temporal evolution. The equation space projections shown in Figs. 4.8(c-ii) and (c-iii) again illustrate the geometric interpretation of the dominant balance approach. In particular, we can see how two interacting terms form a group of dynamical points falling on a line in the three-dimensional equation space having a reduced variance with respect to the direction of the negligible term (see the distribution of blue and grey dynamical points). In contrast, the full NLSE dynamics is associated with a group of dynamical points in a plane that do not have a reduced variance with respect to any axis of the equation space.

We can see how the results of the dominant balance analysis provide insight into the physics of optical wave breaking in both the spatio-temporal and spatio-spectral domains [see Figs. 4.8(d) and (e)]. In particular, it becomes apparent from the analysis of the temporal evolution that the regions of highest intensity are primarily dominated by nonlinearity [see label C in Fig. 4.8(d)], while the steepening trailing edges see the effect of both dispersion and nonlinearity [see label B in Fig. 4.8(d)]. On the later stage of propagation, the wave breaking leads to the emergence of the smaller amplitude temporal wings. These wings are found to be associated with the dispersive propagation [see label A in Fig. 4.8(d)].

Accordingly, the temporal dynamical regimes mentioned above are reflected in spectral domain shown in Fig. 4.8(b). Specifically, the central part of the spatio-spectral evolution is again associated with the nonlinear dominant balance (blue). It can also be seen that the appearance of spectral sidelobes coincides with the spatio-spectral regions where the dispersion dominates the propagation [see grey local regions in Fig. 4.8(e)]. An obvious visual similarity between these temporal and spectral pictures is expected in optical wave breaking propagation due to the approximately linear time-frequency mapping [158–160].

The dominant balance physics revealed in the case of optical wave breaking is indeed qualitatively different from the case of Riemann wave propagation (see Section 4.3.1). In the optical wave breaking case, the injected unchirped high-power Gaussian pulse experience a purely nonlinear propagation in the initial stage. This leads to the dominance of the nonlinearity in both time and frequency domains up to $\xi \approx 0.02$. It is only at the later stage that the full NLSE dynamics starts to dominate the evolution of the trailing edges of the flat-top pulse profiles.

To better illustrate the dominant balance search algorithm used in this chapter, in Appendix B we provide the Python code for the dominant balance analysis of the optical wave breaking propagation scenario considered above.

4.3.3/ Perturbed fundamental soliton propagation

The two examples discussed in Sections 4.3.1 and 4.3.2 were related to the normal dispersion propagation regime. A further three propagation scenarios presented in the following sections correspond to propagation in the anomalous dispersion regime.

The first example corresponds to the perturbed soliton propagation simulated numerically using the split-step Fourier method [22] for the following initial conditions $\psi(0,\tau) = N\,\mathrm{sech}(\tau/\sqrt{2})$ with $N = 1.5$. The simulation results and the corresponding dominant balance analysis shown in Fig. 4.9 allow us to study the effect of soliton perturbation.

Figures 4.9(a) and (b) show the spatio-temporal and spatio-spectral dynamics of the perturbed soliton, while Figs. 4.9(d) and (e) present the results of the dominant balance analysis for both domains. An oscillatory time evolution [see label A in Fig. 4.9(d)] leads to spectral compressions and expansions that is followed by the energy transfer from the oscillating time component to the dispersive low-amplitude background [see label B in Fig. 4.9(d)] arising in the presence of a perturbation [22]. It can be seen from Fig. 4.9(d) how the dominant balance algorithm separates this oscillating temporal part of the pulse from the broad dispersive background. Specifically, the central part of the temporal pulse envelope is associated with the full NLSE dynamics (orange), while the broad dispersive background is associated with $\{i\psi_\xi, \psi_{\tau\tau}\}$ dominant balance model (grey). This temporal dynamics is directly reflected in the frequency domain, where the oscillating spectral wings are also associated with the full NLSE dynamics [see Fig. 4.9(e)].

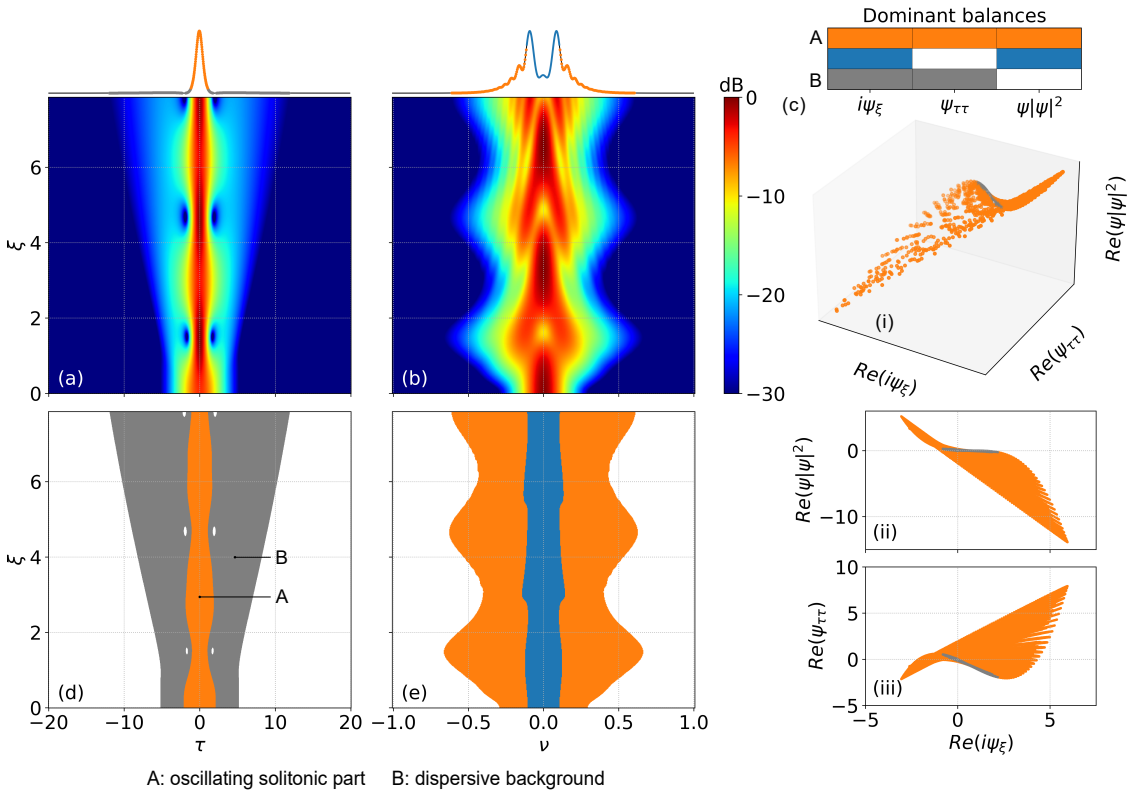


Figure 4.9: **(a)** spatio-temporal and **(b)** spatio-spectral intensity maps of perturbed soliton propagation (in both cases the maps are normalised to the maximum intensity and plotted on the same logarithmic scale). The insets at the top show the corresponding output intensity profiles on a linear scale where the colors represent the dominant balance regions (see the inset frame in the top right corner of the figure). In **(c)** **(i)** shows the spatio-temporal evolution dynamics in equation space divided into two dominant balance regions. Two projections of the equation space are shown in subfigures **(ii)** and **(iii)**. **(d)** and **(e)** show the identified dominant balance models mapped onto the original domains for comparison with the intensity evolution maps. The same color code is used for the dominant balance models identified in the spatio-spectral domain, while meaning the dominant balances between the Fourier transformed temporal terms shown in the legend. (d) labels particular temporal features - A: the oscillating solitonic component; B: the dispersive background.

It is noticeable that the dispersive dominant balance is not present in the frequency domain, while the central part of the pulse remain to be dominated by the nonlinearity (blue region). Indeed, the second order dispersion contribution is small in the vicinity of $\nu = 0$.

Finally, it should be noted that the equation space dynamics shown in Fig. 4.9(c) and reconstructed from the spatio-temporal pulse evolution (Fig. 4.9(a)) does not show a clear separation between the two dominant physical regimes in this case. Indeed, two groups of the dynamical points (grey and orange) overlap closely, which may result in slightly higher model uncertainty in this case.

4.3.4/ Soliton fission induced by third-order dispersion

Let us now consider the higher-order soliton propagation in the presence of third order dispersion (TOD). These dynamics is governed by the following generalised NLSE that we write here using the same dimensionless variables as earlier in this secion

$$i\psi_\xi + \psi_{\tau\tau} + i\delta\psi_{\tau\tau\tau} + |\psi|^2\psi = 0, \tag{4.10}$$

where the $i\delta\psi_{\tau\tau\tau}$ term accounts for the TOD effect. The following initial conditions were used for the numerical simulations using the split-step Fourier method: $\psi(0,\tau) = N\,\mathrm{sech}(\tau/\sqrt{2})$ with $N = 3$ and $\delta = \sqrt{2}\beta_3/3T_0|\beta_2| \approx 0.06$. Figure 4.10(a) shows the evolution of the input hyperbolic secant pulse leading to the soliton fission in the presence of the perturbation induced by TOD. Figure 4.10(b) shows the corresponding spectral evolution.

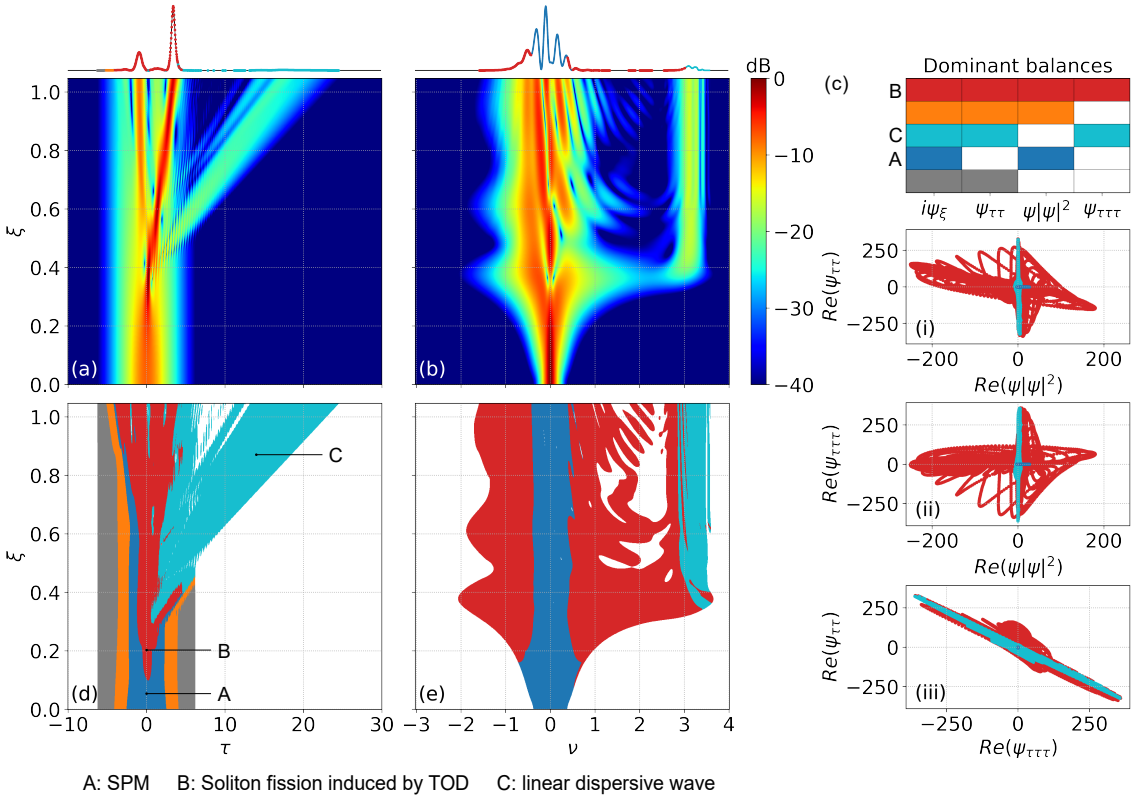


Figure 4.10: **(a)** spatio-temporal and **(b)** spatio-spectral intensity maps of the soliton fission dynamics induced by the TOD (in both cases the maps are normalised to the maximum intensity and plotted on the same logarithmic scale). The insets at the top show the corresponding output intensity profiles on a linear scale where the colors represent the dominant balance regions (see the inset frame in the top right corner of the figure). In **(c)** three projections of the equation space are shown in subfigures **(i)**, **(ii)** and **(iii)**. **(d)** and **(e)** show the identified dominant balance models mapped onto the original domains for comparison with the intensity evolution maps. The same color code is used for the dominant balance models identified in the spatio-spectral domain, while meaning the dominant balances between the Fourier transformed temporal terms shown in the legend. (d) labels particular temporal features - A: initial pulse compression; B: soliton fission dynamics induced by TOD; C: linear dispersive wave.

With $K = 4$ differential equation terms in Eq. 4.10, we have a total of 7 candidate dominant balance models (see Fig. 4.1), while only 5 dominant balance models are identified by the dominant balance search algorithm in this case. We can see from the results of the dominant balance analysis presented in Figs. 4.10(d) and (e) that the initial stage of the evolution is described by the nonlinear dominant balance model (blue region). Indeed, for $\xi < 0.2$, the propagation is associated with higher-order soliton compression, which leads to the appearance of a high-power temporal peak in the central part of the pulse [see label A in Fig. 4.10(d)]. These temporal dynamics results in nonlinear dominant balance governing the initial stage of the propagation in both the time and spectral domains.

At around $\xi = 0.1$ the contribution of the TOD becomes significant. The combined effect of the NLSE and the TOD (red region) leads to the soliton fission at $\xi \sim 0.4$ [see label B in Fig. 4.10(d)]. The later dynamics shows the ejection of the dispersive wave [see label C in Fig. 4.10(d)] that is associated with the $\{i\psi_\xi, \psi_{\tau\tau}, \psi_{\tau\tau\tau}\}$ dominant balance in both domains (light blue regions). This nonsolitonic radiation is, in fact, known to appear in higher-order soliton propagation in the presence of TOD [56] (see Section 1.4). At the same time, it can be seen how the full dynamics of the Eq. 4.10 drives the solitonic propagation in the temporal domain resulting in the dominance of the full propagation model in the spectral wings (red regions).

It is important to note that for the governing differential equation models with $K > 3$, it is not possible to visually represent the multi-dimensional equation space dynamics. Figure 4.10(c) shows three projections of the equation space highlighting the identified dominant balance models in the spatio-temporal domain. This clearly demonstrates a strong localization of the $\{i\psi_\xi, \psi_{\tau\tau}, \psi_{\tau\tau\tau}\}$ dominant balance in the equation space (a group of light blue dynamical points). The dominant balance between the second-order dispersion and propagation terms $\{i\psi_\xi, \psi_{\tau\tau}\}$ (grey), as well as the NLSE dominant balance (orange) describe the dynamics in local regions of a relatively low intensity of the spatio-temporal map. Moreover, the equation space clusters associated with these two dominant balances have significantly smaller variance, and thus these dynamics are of least importance [see Figure 4.10(c)]. Contrary to that, we see how the full propagation model (red regions) has a relatively large variance and is represented by a group of dynamical points scattered across the equation space.

4.3.5/ Soliton fission in the presence of TOD and Raman effects

As the final example we consider higher-order soliton propagation in the presence of both third order dispersion (TOD) and Raman effects. The governing generalised NLSE in this case is

$$i\psi_\xi + \psi_{\tau\tau} + i\delta\psi_{\tau\tau\tau} + |\psi|^2\psi + \rho\psi\,(h_\mathrm{R} * |\psi|^2) = 0, \tag{4.11}$$

where the dimensionless variables correspond to $\psi = A\sqrt{\gamma L_\mathrm{D}}$, $\xi = z(1 - f_\mathrm{R})/L_D$, $\tau = t\sqrt{2(1 - f_\mathrm{R})}/T_0$. The TOD and Raman effects are included through the dimensionless parameters δ and ρ, and the operator $(*)$ stands here for the convolution. Figures 4.11(a) and (b) show the results of the numerical simulations obtained for the input hyperbolic secant pulse with $N = 2$, which is defined in the following way using the introduced dimensionless variables: $\psi(0, \tau) = N\,\mathrm{sech}[\tau/\sqrt{2(1 - f_\mathrm{R})}]$. Dimensionless parameters $\delta = 0.05$ and $\rho = f_\mathrm{R}/(1 - f_\mathrm{R})$ with Raman fraction $f_\mathrm{R} = 0.34$ were used in the simulations.

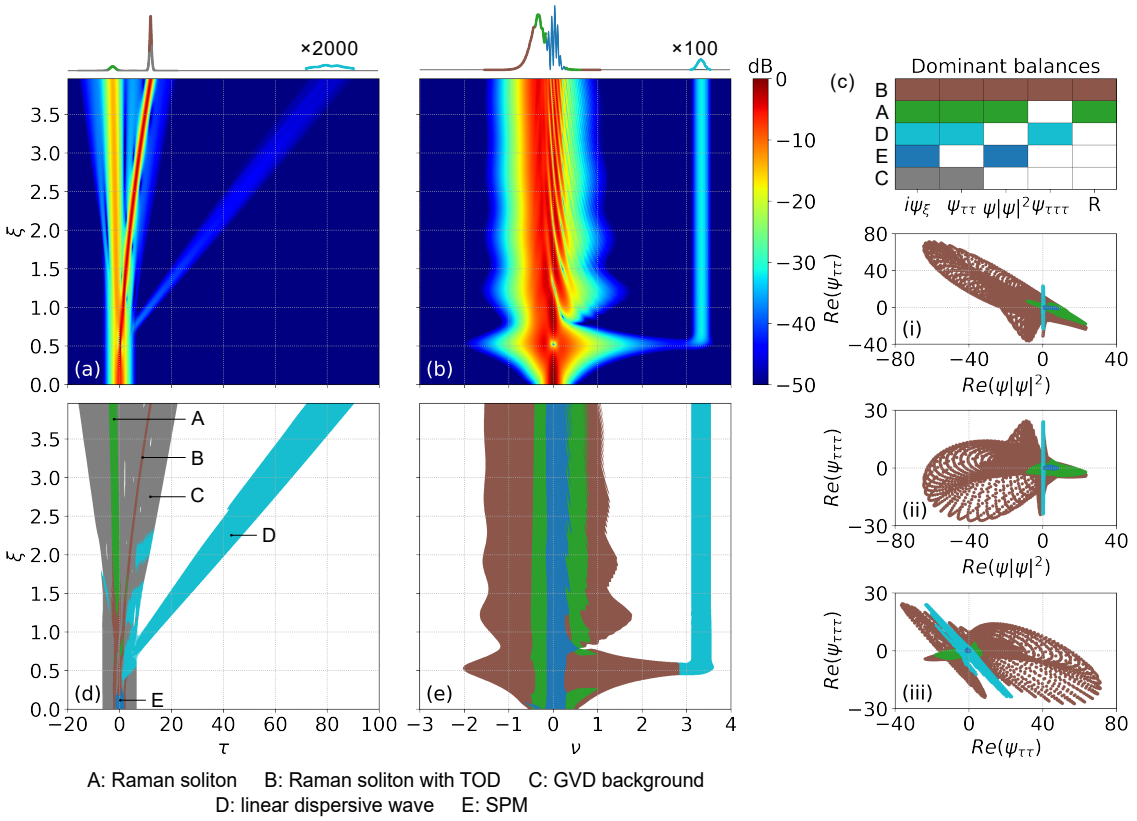


Figure 4.11: **(a)** spatio-temporal and **(b)** spatio-spectral intensity maps of the soliton fission dynamics in the presence of TOD and Raman effects (in both cases the maps are normalised to the maximum intensity and plotted on the same logarithmic scale). The insets at the top show the corresponding output intensity profiles on a linear scale where the colors represent the dominant balance regions (see the inset frame in the top right corner of the figure). In **(c)** three projections of the equation space are shown in subfigures **(i)**, **(ii)** and **(iii)**. **(d)** and **(e)** show the identified dominant balance models mapped onto the original domains for comparison with the intensity evolution maps. The same color code is used for the dominant balance models identified in the spatio-spectral domain, while meaning the dominant balances between the Fourier transformed temporal terms shown in the legend. (d) labels particular temporal features - A: Raman soliton; B: Raman soliton with TOD (all terms); C: dispersive (GVD) background; D: linear dispersive wave; E: initial temporal pulse compression.

The above conditions are known to result in rich soliton fission dynamics [54]. Indeed, 5 dominant balance models (out of 15 candidate dominant balance models) are identified by the algorithm in this case [see Fig. 4.11(c), (d) and (e)]. The initial stage of the propagation that corresponds to the temporal pulse compression [see label E in Fig. 4.11(d)] is again associated with the nonlinear dominant balance (blue region). We can directly see how the soliton fission process, occurring at $\xi \sim 0.7$, is governed by the full propagation model with 5 active terms (brown region).

The soliton fission is followed by the ejection of the dispersive wave which is associated with the $\{i\psi_\xi, \psi_{\tau\tau}, \psi_{\tau\tau\tau}\}$ dominant balance [see label D in Fig. 4.11(d)] in both temporal and spectral domains (light blue regions). This is essentially the same mechanism of the dispersive wave generation that is observed in the case of the soliton fission induced in

the presence of isolated TOD effect considered earlier. In addition to that, we can notice two light blue regions located at $\xi \sim 1.2$ and $\xi \sim 2$ in the temporal domain. Interestingly, the positions of these spots along with the one at $\xi \sim 0.7$ correspond to the propagation distances at which the phase matching condition for the dispersive wave generation is satisfied [57, 161].

At the later stage of the propagation we can observe the evolution of the two ejected solitons in the temporal domain [see labels A and B in Fig. 4.11(d)]. It is peculiar that the dominant balance algorithm suggests that these temporal structures are governed by slightly different physical effects. The higher amplitude soliton is associated with the full propagation model (brown), while the lower-amplitude soliton (green) practically does not see the effect of TOD. From the physical point of view, this explains the acceleration of the higher amplitude soliton arising due to the action of TOD. Narrow temporal duration and high peak power of this temporal structure leads to the dominance of the full propagation model (brown) in the spectral wings [see Fig. 4.11(e)]. The fact that different propagation models are proposed here for the two ejected solitons clearly shows the ability of the dominant balance search algorithm to provide the complementary insight into the observed physics.

Finally, the inset plots on the top of the Figs. 4.11(a) and (b) show the output intensity profiles on the linear scale. Note that due to the low relative amplitude of the dispersive waves, their profiles are scaled by the factors shown. This demonstrates the ability of the dominant balance search algorithm to capture the dynamical patterns at different scales. Interestingly, the line profile of the output temporal field shows that the wings of the accelerating soliton contain a strong dispersive contribution [see label C in Fig. 4.11(d)], and indeed this aspect of the dominant balance analysis may shed new light on the well-known ability of solitons to couple through dispersive interactions.

General conclusion and perspectives

Data-driven methods and algorithms are an exciting area of recent development in machine learning (ML) with great potential to fundamentally modify how research in physics is conducted. In this thesis, we have shown the feasibility of data-driven methods to extract interpretable physical models from data describing the complex nonlinear dynamics in optical fiber propagation.

Firstly, we have demonstrated the possibility of data-driven discovery of governing differential equations from noisy dynamical data in nonlinear fiber optics. Specifically, the results presented in Chapter 2 show that the sparse identification of nonlinear dynamics (SINDy) algorithm can successfully retrieve the differential equation model describing ideal four wave mixing (FWM) dynamics from data even at moderate noise levels, which are likely to be obtainable in experiments. These results are especially valuable since FWM is a central process not only in nonlinear optics, but also in many other domains of physics.

To successfully retrieve the physically-consistent FWM model from noisy dynamical data, we proposed an approach to interpret the SINDy results by analyzing an ensemble of input data series corresponding to different initial conditions. This approach involves examining the histogram of the number of terms associated with multiple returned differential equations in order to select the parsimonious model describing the dynamics. We have also demonstrated how this allows us to directly estimate the uncertainty limits of the returned SINDy model by calculating the mean and variance in the associated term coefficients. We expect that this approach will prove very useful in future attempts to extract governing differential equation models from experimental measurements.

In addition to works focused on improving the robustness of the SINDy algorithm for the case of noisy data, an important direction is the development of algorithms to find coordinates where the dynamics have a sparse representation [102]. In particular, it can be seen from the results of Chapter 2 that the SINDy algorithm is notably less efficient in finding the dynamics of coupled amplitude equations compared to the Hamiltonian model, which is more sparse in the equation space. For many dynamical systems in physics, the coordinate systems in which the dynamics have a sparse representation may not be known in advance. On this basis, we expect that the development of coordinate search algorithms for sparse representations of the dynamics may also prove to be an important step towards extracting models from experimental data.

Apart from finding differential equation models from experimental data, the SINDy algorithm can be an excellent tool for finding reduced models in nonlinear optics from data obtained by directly solving the Maxwell's equations. For example, we anticipate the application of the SINDy algorithm to finite-difference time-domain simulation data to help interpret complex dynamics. The retrieved reduced models can then be used for optimization and control.

Next, we have demonstrated how the data-driven dominant balance approach can complement conventional theoretical analysis by providing insights into the locally dominant physics in optical fiber propagation. The results presented in Chapters 3 and 4 clearly show the power of this approach to isolate local spatio-temporal and spatio-spectral regions exhibiting nonlinear and dispersion-dominated dynamical regimes in a completely unsupervised manner. In particular, we have shown how the combination of unsupervised Gaussian mixture modeling (GMM) clustering along with a specific model selection algorithm can help identify the dominant balance models in various propagation scenarios, including noise-driven modulation instability, optical wave breaking, Riemann wave shocks, perturbed fundamental soliton dynamics and soliton fission in the presence of third-order dispersion and Raman scattering.

A natural direction for further development of the data-driven dominant balance approach is to optimise both the clustering step and the model selection algorithm. The unsupervised GMM clustering of equation space dynamics can be potentially improved by embedding a Bayesian information criterion. In particular, this strategy can be used to set the initial number of GMM clusters to best partition the equation space. It is also of interest to propose any additional physical criteria to improve the efficiency of the dominant model selection algorithm.

The data-driven dominant balance approach can potentially be applied to various complex ultrafast optical systems where the spatio-temporal and spatio-spectral dynamics exhibit different behaviour depending on control parameters. Specifically, we anticipate that this approach can be used to interpret various dynamical regimes arising in fiber laser systems. The ability of the data-driven dominant balance technique to separate different dynamical regimes opens the way to a fully automated method to gain an "intuitive" understanding of the underlying physics. We also expect that data-driven dominant balance can be used in order to simplify numerical integration schemes. In particular, one can consider implementing adaptive step strategies based on the returned local dominant balance models.

As a final remark, we would like to emphasize that data-driven methods are not intended to replace existing conventional approaches to interpreting complex nonlinear dynamics and making discoveries in nonlinear fiber optics. Rather, these novel ML approaches should be viewed as complementary tools to help interpret observed dynamics.

Bibliography

[1] T. S. Kuhn, *The Structure of Scientific Revolutions*. University of Chicago press, 1962.

[2] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, pp. 255–260, 2015.

[3] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2008.

[4] T. Hey, S. Tansley, and K. Tolle, eds., *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, Washington: Microsoft Research, 2009.

[5] B. A. Richards, T. P. Lillicrap, P. Beaudoin, Y. Bengio, R. Bogacz, A. Christensen, C. Clopath, R. P. Costa, A. de Berker, S. Ganguli, C. J. Gillon, D. Hafner, A. Kepecs, N. Kriegeskorte, P. Latham, G. W. Lindsay, K. D. Miller, R. Naud, C. C. Pack, P. Poirazi, P. Roelfsema, J. Sacramento, A. Saxe, B. Scellier, A. C. Schapiro, W. Senn, G. Wayne, D. Yamins, F. Zenke, J. Zylberberg, D. Therien, and K. P. Kording, "A deep learning framework for neuroscience," *Nature Neuroscience*, vol. 22, no. 11, pp. 1761–1770, 2019.

[6] N. M. Ball and R. J. Brunner, "Data mining and machine learning in astronomy," *International Journal of Modern Physics D*, vol. 19, no. 07, pp. 1049–1106, 2010.

[7] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, "Machine learning and the physical sciences," *Reviews of Modern Physics*, vol. 91, no. 4, p. 045002, 2019.

[8] S. L. Brunton and J. N. Kutz, *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control*. Cambridge University Press, 2019.

[9] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, and L. Yang, "Physics-informed machine learning," *Nature Reviews Physics*, vol. 3, no. 6, pp. 422–440, 2021.

[10] E. Lewis, C. Sheridan, M. O'Farrell, D. King, C. Flanagan, W. Lyons, and C. Fitzpatrick, "Principal component analysis and artificial neural network based approach to analysing optical fibre sensors signals," *Sensors and Actuators A: Physical*, vol. 136, no. 1, pp. 28–38, 2007.

[11] W. Ma, Z. Liu, Z. A. Kudyshev, A. Boltasseva, W. Cai, and Y. Liu, "Deep learning for the design of photonic structures," *Nature Photonics*, vol. 15, no. 2, pp. 77–90, 2020.

[12] G. Wetzstein, A. Ozcan, S. Gigan, S. Fan, D. Englund, M. Soljačić, C. Denz, D. A. B. Miller, and D. Psaltis, "Inference in artificial intelligence with deep optics and photonics," *Nature*, vol. 588, no. 7836, pp. 39–47, 2020.

[13] P. Freire, E. Manuylovich, J. E. Prilepsky, and S. K. Turitsyn, "Artificial neural networks for photonic applications—from algorithms to implementation: tutorial," *Advances in Optics and Photonics*, vol. 15, no. 3, p. 739, 2023.

[14] G. Genty, L. Salmela, J. M. Dudley, D. Brunner, A. Kokhanovskiy, S. Kobtsev, and S. K. Turitsyn, "Machine learning and applications in ultrafast photonics," *Nature Photonics*, vol. 15, no. 2, pp. 91–101, 2020.

[15] S. L. Brunton and J. N. Kutz, "Promising directions of machine learning for partial differential equations," *Nature Computational Science*, 2024 (https://doi.org/10.1038/s43588-024-00643-2).

[16] S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," *Proceedings of the National Academy of Sciences*, vol. 113, no. 15, pp. 3932–3937, 2016.

[17] J. L. Callaham, J. V. Koch, B. W. Brunton, J. N. Kutz, and S. L. Brunton, "Learning dominant physical processes with data-driven balance models," *Nature Communications*, vol. 12, no. 1, p. 1016, 2021.

[18] B. E. Kaiser, J. A. Saenz, M. Sonnewald, and D. Livescu, "Automated identification of dominant physical processes," *Engineering Applications of Artificial Intelligence*, vol. 116, p. 105496, 2022.

[19] M. Born, E. Wolf, A. B. Bhatia, P. C. Clemmow, D. Gabor, A. R. Stokes, A. M. Taylor, P. A. Wayman, and W. L. Wilcock, *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light*. Cambridge University Press, 1999.

[20] J. D. Jackson, *Classical Electrodynamics*. Wiley, 2003.

[21] N. Akhmediev and A. Ankiewicz, *Solitons*. Springer US, 1997.

[22] G. P. Agrawal, *Nonlinear Fiber Optics*. Elsevier Science & Technology, 2019.

[23] L. Allen and J. H. Eberly, *Optical resonance and two-level atoms*. Wiley, 1975.

[24] L. Landau and E. Lifshitz, *Electrodynamics of Continuous Media*. Elsevier, 1984.

[25] A. V. Ermolaev and T. A. Vartanyan, "Rigorous theory of thin-vapor-layer linear optical properties: The case of specular reflection of atoms colliding with the walls," *Physical Review A*, vol. 101, no. 5, p. 053850, 2020.

[26] A. V. Ermolaev and T. A. Vartanyan, "Theory of thin-vapor-layer linear-optical properties: The case of quenching of atomic polarization upon collisions of atoms with dielectric walls," *Physical Review A*, vol. 105, no. 1, p. 013518, 2022.

[27] L. Tsang, J. A. Kong, and K. Ding, *Scattering of Electromagnetic Waves: Theories and Applications*. Wiley, 2000.

[28] R. Deiterding, R. Glowinski, H. Oliver, and S. Poole, "A reliable split-step fourier method for the propagation equation of ultra-fast pulses in single-mode optical fibers," *Journal of Lightwave Technology*, vol. 31, no. 12, pp. 2008–2017, 2013.

[29] V. E. Zakharov, "Stability of periodic waves of finite amplitude on the surface of a deep fluid," *Journal of Applied Mechanics and Technical Physics*, vol. 9, no. 2, pp. 190–194, 1972.

[30] R. K. Dodd, J. C. Eilbeck, J. D. Gibbon, and H. C. Morris, *Solitons and nonlinear wave equations*. Academic Press, 1982.

[31] Y. Kivshar and G. Agrawal, *Optical Solitons: From Fibers to Photonic Crystals*. Academic Press, 2003.

[32] T. B. Benjamin and J. E. Feir, "The disintegration of wave trains on deep water. Part I. Theory.," *Journal of Fluid Mechanics*, vol. 27, pp. 417–430, 1967.

[33] V. I. Bespalov and V. I. Talanov, "Filamentary Structure of Light Beams in Nonlinear Liquids," *JETP Letters*, vol. 3, pp. 307–310, 1966.

[34] D. H. Peregrine, "Water waves, nonlinear Schrödinger equations and their solutions," *The Journal of the Australian Mathematical Society. Series B. Applied Mathematics*, vol. 25, no. 01, pp. 16–43, 1983.

[35] E. Kuznetsov, "Solitons in a parametrically unstable plasma," *Soviet Physics Doklady*, vol. 22, pp. 507—-508, 1977.

[36] N. N. Akhmediev and V. I. Korneev, "Modulation instability and periodic solutions of the nonlinear Schrödinger equation," *Theoretical and Mathematical Physics*, vol. 69, pp. 1089–1093, 1986.

[37] E. Brainis, D. Amans, and S. Massar, "Scalar and vector modulation instabilities induced by vacuum fluctuations in fibers: Numerical study," *Physical Review A*, vol. 71, no. 2, p. 023808, 2005.

[38] P. Walczak, S. Randoux, and P. Suret, "Optical rogue waves in integrable turbulence," *Physical Review Letters*, vol. 114, no. 14, p. 143903, 2015.

[39] S. Randoux, P. Walczak, M. Onorato, and P. Suret, "Nonlinear random optical waves: Integrable turbulence, rogue waves and intermittency," *Physica D: Nonlinear Phenomena*, vol. 333, pp. 323–335, 2016.

[40] A. Gelash, D. Agafontsev, V. Zakharov, G. El, S. Randoux, and P. Suret, "Bound state soliton gas dynamics underlying the spontaneous modulational instability," *Physical Review Letters*, vol. 123, no. 23, p. 234102, 2019.

[41] A. E. Kraych, D. Agafontsev, S. Randoux, and P. Suret, "Statistical properties of the nonlinear stage of modulation instability in fiber optics," *Physical Review Letters*, vol. 123, no. 9, p. 093902, 2019.

[42] A. M. Perego, F. Bessin, and A. Mussot, "Complexity of modulation instability," *Physical Review Research*, vol. 4, no. 2, p. l022057, 2022.

[43] J. M. Dudley, G. Genty, F. Dias, B. Kibler, and N. Akhmediev, "Modulation instability, akhmediev breathers and continuous wave supercontinuum generation," *Optics Express*, vol. 17, no. 24, pp. 21497–21508, 2009.

[44] S. Toenger, T. Godin, C. Billet, F. Dias, M. Erkintalo, G. Genty, and J. M. Dudley, "Emergent rogue wave structures and statistics in spontaneous modulation instability," *Scientific Reports*, vol. 5, p. 10380, 2015.

[45] A. A. Gelash and V. E. Zakharov, "Superregular solitonic solutions: a novel scenario for the nonlinear stage of modulation instability," *Nonlinearity*, vol. 27, no. 4, pp. R1–38, 2014.

[46] A. Gelash, D. Agafontsev, P. Suret, and S. Randoux, "Solitonic model of the condensate," *Physical Review E*, vol. 104, no. 4, p. 044213, 2021.

[47] P. Suret, S. Randoux, A. Gelash, D. Agafontsev, B. Doyon, and G. El, "Soliton gas: Theory, numerics, and experiments," *Physical Review E*, vol. 109, no. 6, p. 061001, 2024.

[48] T. Congy, G. El, G. Roberti, A. Tovbis, S. Randoux, and P. Suret, "Statistics of extreme events in integrable turbulence," *Physical Review Letters*, vol. 132, no. 20, p. 207201, 2024.

[49] M. H. Frosz, "Validation of input-noise model for simulations of supercontinuum generation and rogue waves," *Optics Express*, vol. 18, no. 14, p. 14778, 2010.

[50] C. Sulem and P.-L. Sulem, *The Nonlinear Schrödinger Equation: Self-Focusing and Wave Collapse*. Springer, 2007.

[51] B. Kibler, J. Fatome, C. Finot, G. Millot, F. Dias, G. Genty, N. Akhmediev, and J. M. Dudley, "The Peregrine soliton in nonlinear fibre optics," *Nature Physics*, vol. 6, no. 10, pp. 790–795, 2010.

[52] B. Kibler, J. Fatome, C. Finot, G. Millot, G. Genty, B. Wetzel, N. Akhmediev, F. Dias, and J. M. Dudley, "Observation of Kuznetsov-Ma soliton dynamics in optical fibre," *Scientific Reports*, vol. 2, no. 1, p. 463, 2012.

[53] B. Frisquet, B. Kibler, and G. Millot, "Collision of akhmediev breathers in nonlinear fiber optics," *Physical Review X*, vol. 3, no. 4, p. 041032, 2013.

[54] J. M. Dudley, G. Genty, and S. Coen, "Supercontinuum generation in photonic crystal fiber," *Reviews of Modern Physics*, vol. 78, pp. 1135–1184, 2006.

[55] J. M. Dudley and J. R. Taylor, *Supercontinuum Generation in Optical Fibers*. Cambridge University Press, 2010.

[56] N. Akhmediev and M. Karlsson, "Cherenkov radiation emitted by solitons in optical fibers," *Physical Review A*, vol. 51, no. 3, pp. 2602–2607, 1995.

[57] I. Cristiani, R. Tediosi, L. Tartara, and V. Degiorgio, "Dispersive wave generation by solitons in microstructured optical fibers," *Optics Express*, vol. 12, pp. 124–135, 2004.

[58] K. Blow and D. Wood, "Theoretical description of transient stimulated raman scattering in optical fibers," *IEEE Journal of Quantum Electronics*, vol. 25, no. 12, pp. 2665–2673, 1989.

[**59**] R. H. Stolen, W. J. Tomlinson, H. A. Haus, and J. P. Gordon, "Raman response function of silica-core fibers," *Journal of the Optical Society of America B*, vol. 6, no. 6, p. 1159, 1989.

[**60**] N. Akhmediev, W. Królikowski, and A. Lowery, "Influence of the raman-effect on solitons in optical fibers," *Optics Communications*, vol. 131, no. 4–6, pp. 260–266, 1996.

[**61**] J. P. Gordon, "Theory of the soliton self-frequency shift," *Optics Letters*, vol. 11, no. 10, p. 662, 1986.

[**62**] E. M. Dianov, A. M. Prokhorov, and V. N. Serkin, "Dynamics of ultrashort-pulse generation by raman fiber lasers: cascade self-mode locking, optical pulsons, and solitons," *Optics Letters*, vol. 11, no. 3, p. 168, 1986.

[**63**] Y. Kodama and A. Hasegawa, "Nonlinear pulse propagation in a monomode dielectric guide," *IEEE Journal of Quantum Electronics*, vol. 23, no. 5, pp. 510–524, 1987.

[**64**] G. Van der Sande, D. Brunner, and M. C. Soriano, "Advances in photonic reservoir computing," *Nanophotonics*, vol. 6, no. 3, pp. 561–576, 2017.

[**65**] P. Grelu and N. Akhmediev, "Dissipative solitons for mode-locked lasers," *Nature Photonics*, vol. 6, no. 2, pp. 84–92, 2012.

[**66**] J. N. Kutz and S. L. Brunton, "Intelligent systems for stabilizing mode-locked lasers and frequency combs: Machine learning and equation-free control paradigms for self-tuning optics," *Nanophotonics*, vol. 4, no. 4, pp. 459–471, 2015.

[**67**] X. Fu, S. L. Brunton, and J. Nathan Kutz, "Classification of birefringence in mode-locked fiber lasers using machine learning and sparse representation," *Optics Express*, vol. 22, no. 7, p. 8585, 2014.

[**68**] J. N. Kutz, "Mode-locked soliton lasers," *SIAM Review*, vol. 48, no. 4, pp. 629–678, 2006.

[**69**] D. J. Richardson, J. Nilsson, and W. A. Clarkson, "High power fiber lasers: current status and future perspectives [invited]," *Journal of the Optical Society of America B*, vol. 27, no. 11, p. B63, 2010.

[**70**] R. I. Woodward and E. J. R. Kelleher, "Towards 'smart lasers': self-optimisation of an ultrafast pulse source using a genetic algorithm," *Scientific Reports*, vol. 6, no. 37616, 2016.

[**71**] U. Andral, R. S. Fodil, F. Amrani, F. Billard, E. Hertz, and P. Grelu, "Fiber laser mode locked through an evolutionary algorithm," *Optica*, vol. 2, no. 4, pp. 275–278, 2015.

[**72**] J. Girardot, A. Coillet, M. Nafa, F. Billard, E. Hertz, and P. Grelu, "On-demand generation of soliton molecules through evolutionary algorithm optimization," *Optics Letters*, vol. 47, no. 1, pp. 134–137, 2021.

[**73**] X. Fu and J. N. Kutz, "High-energy mode-locked fiber lasers using multiple transmission filters and a genetic algorithm," *Optics Express*, vol. 21, no. 5, p. 6526, 2013.

[**74**] M. Närhi, L. Salmela, J. Toivonen, C. Billet, J. M. Dudley, and G. Genty, "Machine learning analysis of extreme events in optical fibre modulation instability," *Nature Communications*, vol. 9, no. 1, p. 4923, 2018.

[**75**] L. Salmela, N. Tsipinakis, A. Foi, C. Billet, J. M. Dudley, and G. Genty, "Predicting ultrafast nonlinear dynamics in fibre optics with a recurrent neural network," *Nature Machine Intelligence*, vol. 3, no. 4, pp. 344–354, 2021.

[**76**] M. Mabed, L. Salmela, A. Ermolaev, C. Finot, G. Genty, and J. M. Dudley, "Neural network analysis of unstable temporal intensity peaks in continuous wave modulation instability," *Optics Communications*, vol. 541, p. 129570, 2023.

[**77**] L. Salmela, M. Hary, M. Mabed, A. Foi, J. M. Dudley, and G. Genty, "Feed-forward neural network as nonlinear dynamics integrator for supercontinuum generation," *Optics Letters*, vol. 47, no. 4, pp. 802–805, 2022.

[**78**] M. Schmidt and H. Lipson, "Distilling free-form natural laws from experimental data," *Science*, vol. 324, no. 5923, pp. 81–85, 2009.

[**79**] J. N. Kutz, S. L. Brunton, B. W. Brunton, and J. L. Proctor, *Dynamic Mode Decomposition: Data-Driven Modeling of Complex Systems.* Society for Industrial and Applied Mathematics, 2016.

[**80**] M. Raissi, P. Perdikaris, and G. Karniadakis, "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations," *Journal of Computational Physics*, vol. 378, pp. 686–707, 2019.

[**81**] Y.-H. He, "Ai-driven research in pure mathematics and theoretical physics," *Nature Reviews Physics*, vol. 6, pp. 546–553, Aug. 2024.

[**82**] S. L. Brunton and J. N. Kutz, *Data-Driven Science and Engineering.* Cambridge University Press, 2022.

[**83**] T. Baumeister, S. L. Brunton, and J. Nathan Kutz, "Deep learning and model predictive control for self-tuning mode-locked lasers," *Journal of the Optical Society of America B*, vol. 35, no. 3, p. 617, 2018.

[**84**] S. Coulibaly, F. Bessin, M. G. Clerc, and A. Mussot, "Precursors-driven machine learning prediction of chaotic extreme pulses in kerr resonators," *Chaos, Solitons &; Fractals*, vol. 160, p. 112199, 2022.

[**85**] V. Pammi, M. Clerc, S. Coulibaly, and S. Barbay, "Extreme events prediction from nonlocal partial information in a spatiotemporally chaotic microcavity laser," *Physical Review Letters*, vol. 130, no. 22, p. 223801, 2023.

[**86**] J. Riishede and O. Sigmund, "Inverse design of dispersion compensating optical fiber using topology optimization," *Journal of the Optical Society of America B*, vol. 25, no. 1, p. 88, 2007.

[**87**] S. Molesky, Z. Lin, A. Y. Piggott, W. Jin, J. Vucković, and A. W. Rodriguez, "Inverse design in nanophotonics," *Nature Photonics*, vol. 12, no. 11, pp. 659–670, 2018.

[**88**] Z. He, J. Du, X. Chen, W. Shen, Y. Huang, C. Wang, K. Xu, and Z. He, "Machine learning aided inverse design for few-mode fiber weak-coupling optimization," *Optics Express*, vol. 28, no. 15, p. 21668, 2020.

[**89**] X. Hu and A. Schulzgen, "Design of negative curvature hollow core fiber based on reinforcement learning," *Journal of Lightwave Technology*, vol. 38, no. 7, pp. 1959–1965, 2020.

[**90**] F. Meng, J. Ding, Y. Zhao, H. Liu, W. Su, L. Yang, G. Tao, A. Pryamikov, X. Wang, H. Mu, Y. Niu, J. He, X. Zhang, S. Lou, X. Sheng, and S. Liang, "Artificial intelligence designer for optical fibers: Inverse design of a hollow-core anti-resonant fiber based on a tandem neural network," *Results in Physics*, vol. 46, p. 106310, 2023.

[**91**] A. Mendible, S. L. Brunton, A. Y. Aravkin, W. Lowrie, and J. N. Kutz, "Dimensionality reduction and reduced-order modeling for traveling wave physics," *Theoretical and Computational Fluid Dynamics*, vol. 34, no. 4, pp. 385–400, 2020.

[**92**] A. Sheveleva, S. Hamdi, A. Coillet, C. Finot, and P. Colman, "Analysis of the dispersive fourier transform dataset using dynamic mode decomposition: evidence of multiple vibrational modes and their interplay in a three-soliton molecule," *Optics Letters*, vol. 48, no. 11, p. 3015, 2023.

[**93**] M. Sonnewald, C. Wunsch, and P. Heimbach, "Unsupervised learning reveals geography of global ocean dynamical regions," *Earth and Space Science*, vol. 6, no. 5, pp. 784–794, 2019.

[**94**] A. V. Ermolaev, M. Mabed, C. Finot, G. Genty, and J. M. Dudley, "Analysis of interaction dynamics and rogue wave localization in modulation instability using data-driven dominant balance," *Scientific Reports*, vol. 13, p. 10462, 2023.

[**95**] A. V. Ermolaev, C. Finot, G. Genty, and J. M. Dudley, "Automating physical intuition in nonlinear fiber optics with unsupervised dominant balance search," 2024.

[**96**] Rudy, Samuel H. and Brunton, Steven L. and Proctor, Joshua L. and Kutz, J. Nathan, "Data-driven discovery of partial differential equations," *Science Advances*, vol. 3, no. 4, p. e1602614, 2017.

[**97**] J. Berg and K. Nyström, "Data-driven discovery of PDEs in complex datasets," *Journal of Computational Physics*, vol. 384, pp. 239–252, 2019.

[**98**] A. V. Ermolaev, A. Sheveleva, G. Genty, C. Finot, and J. M. Dudley, "Data-driven model discovery of ideal four-wave mixing in nonlinear fibre optics," *Scientific Reports*, vol. 12, no. 1, p. 12711, 2022.

[**99**] T. Bikmukhametov and J. Jäschke, "Combining machine learning and process engineering physics towards enhanced accuracy and explainability of data-driven models," *Computers & Chemical Engineering*, vol. 138, p. 106834, 2020.

[**100**] E. P. Alves and F. Fiuza, "Data-driven discovery of reduced plasma physics models from fully-kinetic simulations," *arXiv:2011.01927 [physics.plasm-ph]*, 2020.

[**101**] M. Sorokina, S. Sygletos, and S. Turitsyn, "Sparse identification for nonlinear optical communication systems: SINO method," *Optics Express*, vol. 24, no. 26, pp. 30433–30443, 2016.

[**102**] K. Champion, B. Lusch, J. N. Kutz, and S. L. Brunton, "Data-driven discovery of coordinates and governing equations," *Proceedings of the National Academy of Sciences*, vol. 116, no. 45, pp. 22445–22451, 2019.

[**103**] B. de Silva, K. Champion, M. Quade, J.-C. Loiseau, J. Kutz, and S. Brunton, "PySINDy: A python package for the sparse identification of nonlinear dynamical systems from data," *Journal of Open Source Software*, vol. 5, no. 49, p. 2104, 2020.

[**104**] S. L. Brunton and J. N. Kutz, "Machine learning for partial differential equations," 2023.

[**105**] Y. Chen and A. W. Snyder, "Four-photon parametric mixing in optical fibers: effect of pump depletion," *Optics Letters*, vol. 14, no. 1, p. 87, 1989.

[**106**] T. Christopoulos, O. Tsilipakos, G. Sinatkas, and E. E. Kriezis, "Degenerate four-wave mixing in nonlinear resonators comprising two-dimensional materials: A coupled-mode theory approach," *Physical Review B*, vol. 98, no. 23, p. 235421, 2018.

[**107**] S. K. Turitsyn, A. E. Bednyakova, M. P. Fedoruk, S. B. Papernyi, and W. R. L. Clements, "Inverse four-wave mixing and self-parametric amplification in optical fibre," *Nature Photonics*, vol. 9, no. 9, pp. 608–614, 2015.

[**108**] S. Trillo and S. Wabnitz, "Dynamics of the nonlinear modulational instability in optical fibers," *Optics Letters*, vol. 16, no. 13, pp. 986–988, 1991.

[**109**] G. Cappellini and S. Trillo, "Third-order three-wave mixing in single-mode fibers: exact solutions and spatial instability effects," *J. Opt. Soc. Am. B*, vol. 8, pp. 824–838, 1991.

[**110**] D. G. Steel and J. F. Lam, "Degenerate four-wave mixing in plasmas," *Optics Letters*, vol. 4, no. 11, p. 363, 1979.

[**111**] L. Deng, E. W. Hagley, J. Wen, M. Trippenbach, Y. Band, P. S. Julienne, J. E. Simsarian, K. Helmerson, S. L. Rolston, and W. D. Phillips, "Four-wave mixing with matter waves," *Nature*, vol. 398, no. 6724, pp. 218–220, 1999.

[**112**] Y. Jiang, Y. Mei, Y. Zuo, Y. Zhai, J. Li, J. Wen, and S. Du, "Anti-parity-time symmetric optical four-wave mixing in cold atoms," *Physical Review Letters*, vol. 123, no. 19, p. 193604, 2019.

[**113**] U. Fasel, Kutz, J. Nathan, B. W. Brunton, and S. L. Brunton, "Ensemble-SINDy: Robust sparse model discovery in the low-data, high-noise limit, with active learning and control," *arXiv:2111.10992 [math.NA]*, 2021.

[**114**] M. Erkintalo, K. Hammani, B. Kibler, C. Finot, N. Akhmediev, J. M. Dudley, and G. Genty, "Higher-order modulation instability in nonlinear fiber optics," *Physical Review Letters*, vol. 107, no. 25, p. 253901, 2011.

[**115**] M. E. Marhic, *Fiber Optical Parametric Amplifiers, Oscillators and Related Devices*. Cambridge, 2016.

[**116**] X. Hu, W. Chen, Y. Lu, Z. Yu, M. Chen, and Z. Meng, "Distributed measurement of fermi–pasta–ulam recurrence in optical fibers," *IEEE Photonics Technology Letters*, vol. 30, no. 1, pp. 47–50, 2018.

[**117**] A. Mussot, C. Naveau, M. Conforti, A. Kudlinski, F. Copie, P. Szriftgiser, and S. Trillo, "Fibre multi-wave mixing combs reveal the broken symmetry of fermi–pasta–ulam recurrence," *Nature Photonics*, vol. 12, no. 5, pp. 303–308, 2018.

[**118**] G. Vanderhaegen, P. Szriftgiser, A. Kudlinski, M. Conforti, S. Trillo, M. Droques, and A. Mussot, "Observation of four fermi-pasta-ulam-tsingou recurrences in an ultra-low-loss optical fiber," *Optics Express*, vol. 28, no. 12, p. 17773, 2020.

[**119**] A. Sheveleva, U. Andral, B. Kibler, P. Colman, J. M. Dudley, and C. Finot, "Idealized four-wave mixing dynamics in a nonlinear schrödinger equation fiber system," *Optica*, vol. 9, no. 6, p. 656, 2022.

[**120**] G. Vanderhaegen, *Intensity and phase measurements along optical fibers to reveal the complex nonlinear dynamic of Fermi-Pasta-Ulam-Tsingou recurrences*. PhD thesis, Université de Lille, 2022.

[**121**] A. Sheveleva, *Experimental characterization of fundamental four-wave mixing: discovering the idealized nonlinear dynamics*. PhD thesis, Université Bourgogne Franche-Comté, 2023.

[**122**] C. Yao and E. M. Bollt, "Modeling and nonlinear parameter estimation with kronecker product representation for coupled oscillators and spatiotemporal systems," *Physica D: Nonlinear Phenomena*, vol. 227, no. 1, pp. 78–99, 2007.

[**123**] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[**124**] D. E. Shea, S. L. Brunton, and J. N. Kutz, "SINDy-BVP: Sparse identification of nonlinear dynamics for boundary value problems," *Physical Review Research*, vol. 3, no. 2, p. 023255, 2021.

[**125**] M. Jacobs, B. W. Brunton, S. L. Brunton, J. N. Kutz, and R. V. Raut, "Hypersindy: Deep generative modeling of nonlinear stochastic governing equations," 2023.

[**126**] N. Zolman, U. Fasel, J. N. Kutz, and S. L. Brunton, "Sindy-rl: Interpretable and efficient model-based reinforcement learning," 2024.

[**127**] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[**128**] K. Kaheman, J. N. Kutz, and S. L. Brunton, "SINDy-PI: a robust algorithm for parallel implicit sparse identification of nonlinear dynamics," *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 476, no. 2242, p. 20200279, 2020.

[**129**] C. B. Delahunt and J. N. Kutz, "A toolkit for data-driven discovery of governing equations in high-noise regimes," *IEEE Access*, vol. 10, pp. 31210–31234, 2022.

[**130**] B. M. de Silva, D. M. Higdon, S. L. Brunton, and J. N. Kutz, "Discovery of physics from data: Universal laws and discrepancies," *Frontiers in Artificial Intelligence*, vol. 3, p. 25, 2020.

[**131**] J. D. Anderson, "Ludwig Prandtl's boundary layer," *Physics Today*, vol. 58, no. 12, pp. 42–48, 2005.

[**132**] G. D. Portwood, S. M. de Bruyn Kops, J. R. Taylor, H. Salehipour, and C. P. Caulfield, "Robust identification of dynamically distinct regions in stratified turbulence," *Journal of Fluid Mechanics*, vol. 807, 2016.

[**133**] G. I. Barenblatt, *Scaling, self-similarity, and intermediate asymptotics*. Cambridge University Press, 1996.

[**134**] R. B. White, *Asymptotic Analysis of Differential Equations*. Imperial college press, 2005.

[**135**] L. Poropat, D. Jones, S. D. A. Thomas, and C. Heuzé, "Unsupervised classification of the northwestern european seas based on satellite altimetry data," *Ocean Science*, vol. 20, no. 1, pp. 201–215, 2024.

[**136**] D. Sturm, S. Maddu, and I. F. Sbalzarini, "Learning locally dominant force balances in active particle systems," 2023 (https://doi.org/10.48550/arXiv.2307.14970).

[**137**] V. E. Zakharov and L. A. Ostrovsky, "Modulation instability: The beginning," *Physica D: Nonlinear Phenomena*, vol. 238, pp. 540–548, 2009.

[**138**] D. S. Agafontsev and A. A. Gelash, "Rogue Waves With Rational Profiles in Unstable Condensate and Its Solitonic Model," *Frontiers in Physics*, vol. 9, p. 610896, 2021.

[**139**] A. Gelash, G. Xu, and B. Kibler, "Management of breather interactions," *Physical Review Research*, vol. 4, no. 3, p. 033197, 2022.

[**140**] J. M. Dudley, F. Dias, M. Erkintalo, and G. Genty, "Instabilities, breathers and rogue waves in optics," *Nature Photonics*, vol. 8, no. 10, pp. 755–764, 2014.

[**141**] B. Kibler, A. Chabchoub, A. Gelash, N. Akhmediev, and V. E. Zakharov, "Superregular Breathers in Optics and Hydrodynamics: Omnipresent Modulation Instability beyond Simple Periodicity," *Physical Review X*, vol. 5, no. 4, p. 041026, 2015.

[**142**] G. Xu, A. Gelash, A. Chabchoub, V. Zakharov, and B. Kibler, "Ghost interaction of breathers," *Frontiers in Physics*, vol. 8, p. 608894, 2020.

[**143**] A. Chabchoub, N. P. Hoffmann, and N. Akhmediev, "Rogue wave observation in a water wave tank," *Physical Review Letters*, vol. 106, no. 20, p. 204502, 2011.

[**144**] A. Chabchoub and N. Akhmediev, "Observation of rogue wave triplets in water waves," *Physics Letters A*, vol. 377, no. 38, pp. 2590–2593, 2013.

[**145**] V. E. Zakharov and A. A. Gelash, "Nonlinear stage of modulation instability," *Physical Review Letters*, vol. 111, no. 5, p. 054101, 2013.

[**146**] P. Suret, R. E. Koussaifi, A. Tikan, C. Evain, S. Randoux, C. Szwaj, and S. Bielawski, "Single-shot observation of optical rogue waves in integrable turbulence using time microscopy," *Nature Communications*, vol. 7, no. 1, p. 13136, 2016.

[**147**] S. Chen, L. Bu, C. Pan, C. Hou, F. Baronio, P. Grelu, and N. Akhmediev, "Modulation instability—rogue wave correspondence hidden in integrable systems," *Communications Physics*, vol. 5, no. 1, p. 297, 2022.

[**148**] M. Baer, "findiff software package," 2018. https://github.com/maroba/findiff.

[**149**] J. L. Callaham, J. V. Koch, B. W. Brunton, J. N. Kutz, and S. L. Brunton, "Learning Dominant Physical Processes With Data-driven Balance Models. Methods and Codes. Release accompanying publication.," *https://doi.org/10.5281/zenodo.4428904*, 2021.

[**150**] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[**151**] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *Journal of Computational and Graphical Statistics*, vol. 15, pp. 265–286, 2006.

[**152**] N. Akhmediev, A. Ankiewicz, and J. M. Soto-Crespo, "Rogue waves and rational solutions of the nonlinear Schrödinger equation," *Physical Review E*, vol. 80, no. 2, p. 026601, 2009.

[**153**] C. Sulem and P. Sulem, *The Nonlinear Schrödinger Equation. Self-Focusing and Wave Collapse*. Springer, 1999.

[**154**] L. Zou, X. Luo, D. Zeng, L. Ling, and L.-C. Zhao, "Measuring the rogue wave pattern triggered from Gaussian perturbations by deep learning," *Physical Review E*, vol. 105, no. 5, p. 054202, 2022.

[**155**] B. Wetzel, D. Bongiovanni, M. Kues, Y. Hu, Z. Chen, S. Trillo, J. M. Dudley, S. Wabnitz, and R. Morandotti, "Experimental generation of riemann waves in optics: A route to shock wave control," *Physical Review Letters*, vol. 117, no. 7, p. 073902, 2016.

[**156**] P. L. François, "Nonlinear propagation of ultrashort pulses in optical fibers: total field formulation in the frequency domain," *Journal of the Optical Society of America B*, vol. 8, pp. 276–293, 1991.

[**157**] G. B. Whitham, *Linear and Nonlinear Waves*. Wiley, 1999.

[**158**] C. Finot, B. Kibler, L. Provost, and S. Wabnitz, "Beneficial impact of wave-breaking for coherent continuum formation in normally dispersive nonlinear fibers," *Journal of the Optical Society of America B*, vol. 25, pp. 1938–1948, 2008.

[**159**] D. Castelló-Lurbe, P. Andrés, and E. Silvestre, "Dispersion-to-spectrum mapping in nonlinear fibers based on optical wave-breaking," *Optics Express*, vol. 21, pp. 28550–28558, 2013.

[**160**] A. M. Heidt, J. S. Feehan, J. H. V. Price, and T. Feurer, "Limits of coherent supercontinuum generation in normal dispersion fibers," *Journal of the Optical Society of America B*, vol. 34, pp. 764–775, 2017.

[**161**] T. X. Tran and F. Biancalana, "Dynamics and control of the early stage of supercontinuum generation in submicron-core optical fibers," *Phys. Rev. A*, vol. 79, p. 065802, 2009.

A

List of acronyms

FWM four wave mixing

NLSE nonlinear Schrödinger equation

GVD group velocity dispersion

TOD third order dispersion

SPM self-phase modulation

MI modulation instability

AB Akhmediev breather

PS Peregrine soliton

KM Kuznetsov-Ma soliton

SFB solitons on finite background

SC supercontinuum

ML machine learning

SINDy sparse identification of nonlinear dynamics

LHS left-hand side

RHS right-hand side

LASSO least absolute shrinkage and selection operator

STLS sequential thresholded least squares

GMM Gaussian mixture modeling

SPCA sparse principal component analysis

B

Dominant balance analysis of optical wave breaking dynamics: Python code

To better illustrate the data-driven dominant balance search algorithm used in Chapter 4, we provide an example of the Python code used for the optical wave breaking example presented in Section 4.3.2. The code is available from the repository:

https://github.com/AndreiErmolaev/Dominant-balance-analysis-of-optical-wave-breaking

C

List of publications

Publications in peer-review journals

1. **Andrei V. Ermolaev**, Christophe Finot, Goery Genty, John M. Dudley, "Automating physical intuition in nonlinear fiber optics with unsupervised dominant balance search," *Optics Letters*, 49(15), pp. 4202-4205 (2024).

2. **Andrei V. Ermolaev**, Mehdi Mabed, Lauri Salmela, Goëry Genty, Christophe Finot, John M. Dudley, "Analysis of interaction dynamics and rogue wave localization in modulation instability using data-driven dominant balance," *Scientific Reports*, 13(1), 10462, 2023.

3. Mehdi Mabed, Lauri Salmela, **Andrei V. Ermolaev**, Christophe Finot, Goëry Genty, John M. Dudley, "Neural network analysis of unstable temporal intensity peaks in continuous wave modulation instability," *Optics Communications*, 541, 129570, 2023.

4. **Andrei V. Ermolaev**, Anastasiia Sheveleva, Lauri Salmela, Goëry Genty, Christophe Finot, John M. Dudley, "Data-driven model discovery of ideal four-wave mixing in nonlinear fiber optics," *Scientific Reports*, 12(1), 12711, 2022.

International conferences[1]

1. **Andrei V. Ermolaev***, Christophe Finot, Goëry Genty, John M. Dudley, "Identifying dominant physical interactions in nonlinear fibre optics using machine learning," SPIE Photonics Europe, Strasbourg, France, 7 - 11 April 2024.

2. **Andrei V. Ermolaev***, Mehdi Mabed, Christophe Finot, Goëry Genty, John M. Dudley, "Analysing interaction and localization dynamics in modulation instability via data-driven dominant balance," The European Optical Society Annual Meeting (EOSAM), Dijon, France, 11 - 15 September 2023.

3. **Andrei V. Ermolaev***, Mehdi Mabed, Lauri Salmela, Goëry Genty, Christophe Finot, John M. Dudley, "Identifying extreme localization and rogue waves in fibre optics using data-driven dominant balance," Conference on Lasers and Electro-Optics/Europe – European Quantum Electronics Virtual Conferences (CLEO®/Europe-EQEC 2023), Munich, Germany, 26-30 June 2023.

[1] In the publication list below, * indicates the presenting author.

4. Mehdi Mabed*, Lauri Salmela, **Andrei V. Ermolaev**, Christophe Finot, Goëry Genty, John M Dudley, "Machine Learning analysis of temporal instability peaks under Continuous Wave excitation in optical fiber Modulation Instability," Conference on Lasers and Electro-Optics/Europe – European Quantum Electronics Virtual Conferences (CLEO®/Europe-EQEC 2023), Munich, Germany, 26-30 June 2023.

5. Goëry Genty*, Lauri Salmela, Mathilde Hary, Mehdi Mabed, **Andrei V. Ermolaev**, John M. Dudley, "Applications of artificial intelligence to nonlinear fiber-optics," AI and Optical Data Sciences IV, SPIE Photonics West 2023, San Francisco, 28 January-2 February 2023.

6. Anastasiia Sheveleva*, **Andrei V. Ermolaev**, Pierre Colman, John M. Dudley, Christophe Finot, "Experimental Investigation of Phase-Space Portraits of Ideal Four-Wave Mixing," 23rd International Conference on Transparent Optical Networks (ICTON), 2-6 July 2023.

7. John M. Dudley*, Mehdi Mabed, **Andrei V. Ermolaev**, Lauri Salmela, Fanchao Meng, Christophe Finot, Goëry Genty, "Applications of Machine Learning to Ultrafast Nonlinear Dynamics," International Conference on Photonics, IEEE Photonics Society Malaysia Chapter, 8-10 August 2022.

8. John M. Dudley*, M. Mabed, **Andrei V. Ermolaev**, Coraline Lapre, Fanchao Meng, Anastasiia Sheveleva, Christophe Finot, Goëry Genty, "Recent advances understanding rogue waves and analogies in nonlinear fibre optics," International Workshop on Rogue Waves, University of Vermont, Vermont, USA, 17-18 December 2022.

9. Sonia Boscolo*, Anastasiia Sheveleva, **Andrei V. Ermolaev**, John M Dudley, Christophe Finot, "Machine learning for ultrafast nonlinear photonics," Advanced Fiber Laser Conference (AFL2022), Changsha, China, 11-13 November 2022.

National conferences

1. **Andrei V. Ermolaev***, Mehdi Mabed, Lauri Salmela, Goëry Genty, Christophe Finot, John M. Dudley, "Identifying extreme localization and rogue waves in fibre optics modulation instability using data-driven dominant balance," International Workshop on Optical Wave & Waveguide Theory and Numerical Modeling (OWTNM), Marseille, France, 4-5 May 2023.

2. Mehdi Mabed*, **Andrei V. Ermolaev**, Lauri Salmela, Goëry Genty, John M. Dudley, "Machine Learning analysis of Continuous Wave fiber Modulation Instability," International Workshop on Optical Wave & Waveguide Theory and Numerical Modeling (OWTNM), Marseille, France, 4-5 May 2023.

3. **Andrei V. Ermolaev***, Anastasiia Sheveleva, Lauri Salmela, Goëry Genty, Christophe Finot, John M. Dudley, "Data-Driven Discovery in Nonlinear Fiber Optics Using Sparse Regression," Journées du GDR Elios, Marseille, France, 17-18 October 2022.

4. **Andrei V. Ermolaev***, Anastasiia Sheveleva, Lauri Salmela, Goëry Genty, Christophe Finot, John M. Dudley, "Data-driven Discovery of the Ideal Four Wave Mixing Model in Nonlinear Fiber Optics," Congrès Général Optique de la Société Française d'Optique, Nice, France, 4-8 July 2022.

Document generated with LaTeX and:
the LaTeX style for PhD Thesis created by S. Galland — http://www.multiagent.fr/ThesisStyle
the tex-upmethodology package suite — http://www.arakhne.org/tex-upmethodology/

école doctorale **sciences pour l'ingénieur et microtechniques**

Title: Data-driven methods for analysing nonlinear propagation in optical fibres

Keywords: Nonlinear fibre optics, Machine Learning, Data-driven discovery

Abstract:

This thesis aims to apply machine learning methods specifically tailored to the analysis and interpretation of optical pulses as they propagate in a single pass through an optical fiber, and under a variety of conditions. In particular, we will focus on data-driven model discovery approaches that involve the use of machine learning to analyze physical system data with the aim of discovering interpretable and generalizable models and developing methods that can substantially accomplish and/or complement conventional theoretical analysis. To this end, both supervised and unsupervised learning methods will be used to deepen understanding of ultrafast nonlinear phenomena in fiber optics systems.

Titre : Méthodes basées sur les données pour l'analyse de la propagation non linéairedans les fibres optiques

Mots-clés : Optique non-linéaire fibrée, apprentissage automatique, découverte basée sur les données

Résumé :

Cette thèse vise à appliquer et à développer les nouvelles méthodes d'apprentissage automatique spécialement adaptées à l'analyse et à l'interprétation des impulsions lors de leur propagation en un seul passage dans une fibre optique, et sous diverses conditions. En particulier, nous nous concentrerons sur les approches de découverte de modèles guidées par les données qui impliquent l'utilisation de l'apprentissage automatique pour analyser les données du système physique dans le but de découvrir les modèles interprétables et généralisables et de développer les méthodes qui peuvent substantiellement accomplir et/ou compléter l'analyse théorique conventionnelle. À cette fin, les méthodes d'apprentissage supervisé et non supervisé seront utilisées pour approfondir la compréhension des phénomènes non linéaires ultrarapides dans les systèmes de fibres optiques.

UBFC UNIVERSITÉ BOURGOGNE FRANCHE-COMTÉ

Université Bourgogne Franche-Comté
32, avenue de l'Observatoire
25000 Besançon, France