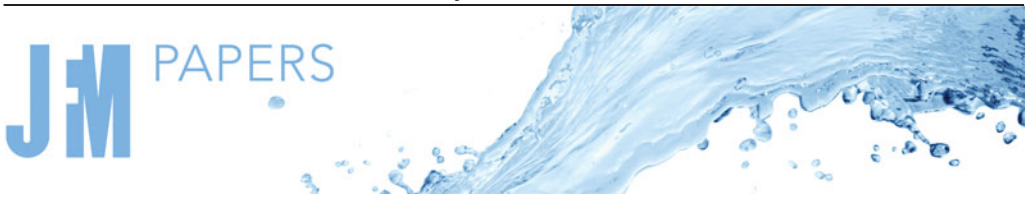
J. Fluid Mech. (2021), *vol*. 920, A27, doi:10.1017/jfm.2021.444



Coherent structure identification in turbulent channel flow using latent Dirichlet allocation

Mohamed Frihat[1], Bérengère Podvin[1],†, Lionel Mathelin[1], Yann Fraigneau[1] and François Yvon[1]

[1]Université Paris-Saclay, CNRS, Laboratoire Interdisciplinaire des Sciences du Numérique, 91400 Orsay, France

(Received 21 October 2020; revised 15 May 2021; accepted 17 May 2021)

Identification of coherent structures is an essential step to describe and model turbulence generation mechanisms in wall-bounded flows. To this end, we present a clustering method based on latent Dirichlet allocation (LDA), a generative probabilistic model for collection of discrete data. The method is applied to a set of snapshots featuring the Reynolds stress (Q_- events) for a turbulent channel flow at a moderate Reynolds number $R_\tau = 590$. Both two- and three-dimensional analysis show that LDA provides a robust and compact flow description in terms of a combination of motifs, which are latent variables inferred from the set of snapshots. We find that the characteristics of the motifs scale with the wall distance, in agreement with the wall-attached eddy hypothesis of Townsend (*Physics of Fluids*, 1961, pp. 97–120). Latent Dirichlet allocation motifs can be used to reconstruct fields with an efficiency that can be compared with the proper orthogonal decomposition (POD). Moreover, the LDA model makes it possible to generate a collection of synthetic fields that captures the intermittent characteristics of the original dataset more clearly than its POD-generated counterpart. These findings highlight the potential of LDA for turbulent flow analysis, efficient reconstruction of actual fields and production of a new field with suitable statistics.

Key words: boundary layer structure, turbulent boundary layers

1. Introduction

The introduction of coherent structures (Townsend 1947; Kline *et al.* 1967) represented a major paradigm shift for turbulence theory and has had a significant impact in various related fields, ranging from geophysical flows to industrial applications. Coherent structure identification has become a key step towards modelling and controlling wall-bounded turbulent flows. However, a recurrent stumbling block is the absence of a precise definition

† Email address for correspondence: podvin@limsi.fr

of structures, as is apparent from several comprehensive reviews (Cantwell 1981; Robinson 1991; Jimenez 2013; Dennis 2015).

Studies originating in the 1960s (Kline *et al.* 1967; Kim, Kline & Reynolds 1971) have established that most of the turbulence in the near-wall region occurred in a highly intermittent manner in both space and time, during what was originally termed 'bursting events'. Quadrant analysis of the Reynolds stress in the plane of streamwise and wall-normal fluctuation (u', v') was introduced by Wallace, Eckelmann & Brodkey (1972) and Willmarth & Lu (1972) to characterize these events. Bursting events were found to be associated with low-speed streaks being lifted away from the wall, as well as with sweeping motions of high-speed fluid towards the wall which, respectively, correspond to quadrant II ($u' < 0, v' > 0$) and quadrant IV ($u' > 0, v' < 0$) events. The two quadrants corresponding to $-u'v' > 0$ can be termed Q_- events and represent the major contribution to the Reynolds stress (Lozano-Duran, Flores & Jimenez 2012).

An interpretation of these bursts is that they are the signature of coherent structures or eddies advected by the mean field. Determining the characteristics of these structures has been the object of considerable effort (Jimenez 2018).

A central element of wall turbulence theory is the attached eddy model, reviewed in detail by Marusic & Monty (2019). The model is based on the idea that turbulence arises as a field of randomly distributed eddies, identified as organized flow patterns which extend to the wall, in the sense that their characteristics are influenced by the wall. Further assumptions require that the entire geometry of the eddies scales with the wall distance, with a constant characteristic velocity scale. The model was extended by Perry & Chong (1982), who introduced the idea of a hierarchy of discrete scales, with an inverse-scale probability distribution. Woodcock & Marusic (2015) showed that this inverse probability distribution was in fact a direct consequence of the self-similarity of the eddies. Further extensions of the model for the logarithmic layer include a wider variety of structures, such as wall-detached ones (Perry & Marusic 1995; Chandran, Monty & Marusic 2020; Hu, Yang & Zheng 2020).

Detection of self-similarity in boundary layers has been the focus of several experimental studies, such as Baars, Hutchins & Marusic (2017), who used spectral coherence analysis to provide evidence of self-similar structures in the streamwise velocity fluctuations of pipe flow. Numerical simulation has proved a powerful tool to explore three-dimensional (3-D) flow fields using a clustering approach. Examples include the work of Alamo *et al.* (2006), who showed that the logarithmic region of turbulent channel was organized in self-similar vortex clusters, and Lozano-Duran *et al.* (2012) developed a 3-D extension of quadrant analysis to detect self-similarity in numerical data at various Reynolds numbers. More recently, wall-attached structures were identified in the streamwise fluctuations of a turbulent boundary layer (Hwang & Sung 2018) as well as in pipe flow (Hwang & Sung 2019). The structures were shown to scale with the wall distance while their population density scales inversely with the distance to the wall. Cheng *et al.* (2020) detected the signature of wall-attached eddies in the streamwise and spanwise velocity fluctuations in turbulent channel flow simulations at low Reynolds numbers. Evidence of self-similarity has been found as well in the context of resolvent analysis (Sharma & McKeon 2013). It has also emerged from proper orthogonal decomposition (POD) results, such as channel flow simulations at low Reynolds numbers (Podvin *et al.* 2010; Podvin & Fraigneau 2017) or pipe flow experiments (Hellström, Marusic & Smits 2016).

The increase of available data, whether through numerical simulation or experiment, has strengthened the need for new identification methods, such as those provided by machine

learning (see Brunton, Noack & Koumoutsakos (2020) for a review). The challenge is to extract structural information about the data without pre-existing knowledge, which defines an unsupervised learning problem. Solutions to this problem should be robust, easy to implement and scalable. One example of an unsupervised learning method that meets these criteria is POD (Lumley 1967), a now classical approach to decompose turbulent fields. Proper orthogonal decomposition is a statistical technique which provides an objective representation of the data as a linear combination of spatial eigenfunctions, which can be hierarchized with respect to a given norm. Although the reconstruction is optimal with respect to this norm (Holmes, Lumley & Berkooz 1996), a potential limitation of the decomposition is that the physical interpretation of the eigenfunctions is not clear. In particular, in the case of homogeneous statistics, the eigenfunctions are spatial Fourier modes over the full domain (see Holmes *et al.* (1996) for a proof), even though instantaneous patterns are strongly localized in space. The connection between POD spatial eigenfunctions with observed coherent structures is therefore not necessarily straightforward. Moreover, the amplitudes of the spatial eigenfunctions are generally strongly interdependent, even though they are by construction uncorrelated. This makes it difficult to give a physical meaning to individual amplitudes, especially in the absence of a probabilistic framework in which to interpret them.

In this paper we consider such a framework to explore an alternative unsupervised learning approach called latent Dirichlet allocation (LDA), which can be derived from POD (Hofmann 1999). Latent Dirichlet allocation is a generative probabilistic model that mimics the characteristics of a collection of data and can be used to create new data. It is based on a soft clustering approach, which was first developed for text mining applications (Blei, Ng & Jordan 2003), but has been extended to other fields in recent years (Aubert *et al.* 2013). By soft clustering, we mean that each data point can belong to more than one cluster. The goal of LDA (Blei *et al.* 2003) is to find short descriptions of the members of a collection that enable efficient processing of large collections while preserving the essential statistical relationships that are useful for basic tasks such as classification, novelty detection, summarization, as well as similarity and relevance. Latent Dirichlet allocation is a three-level hierarchical Bayesian model, in which each member of a collection is modelled as a finite mixture over an underlying set of topics or motifs.

In the field of natural language processing, the dataset to which LDA is applied consists of a set of documents, each of which is considered as a 'bag of words', that is an unordered set of words taken from a finite vocabulary. A particular word may appear several times in the document, or not appear at all. The number of occurrences of each vocabulary word in a document can be seen as an entry of a sparse matrix where the lines correspond to the vocabulary words and the columns to the documents. Based on this typically sparse word count matrix, the classification method returns a set of N_T topics, where the topics are latent variables inferred from the word counts in the documents and the number of topics N_T is a user-defined parameter (Blei *et al.* 2003).

Unlike 'hard' clustering, such as the k-means approach (MacQueen 1967), where each document is assigned to a specific topic, LDA represents each document as a mixture of topics, where the coefficients of the mixture represent the probability of the topic in the document. Each document is characterized by a subset of topics, and each topic can be characterized by a relatively small set of vocabulary words which will appear frequently in the documents.

An interesting application of the LDA method was carried out for a dataset containing images by Griffiths & Steyvers (2004). The dataset considered was a collection of greyscale images where each image consists of an array of pixels, each of which is

associated with a grey level. In this framework, each image is the equivalent of a document, each pixel represents an individual vocabulary word, and the grey-level intensity measured at each pixel is taken as the analogue of the word count matrix entry (the lines of the matrix now represent the pixels, while the columns represent the snapshots). The sum of the intensities over the pixels, which will be called throughout the paper the total intensity, is the analogue of the total number of words observed in the document. Given a set of original patterns constituting the topics or motifs, a collection of synthetic images was generated from random mixtures of the patterns. It was shown that LDA was able to recover the underlying or hidden patterns from the observations of the generated images.

An important point is that identification is carried out without any assumption about the data organization, unlike other analysis methods which are based on an initial selection of suitable objects, usually through the application of a threshold to the data.

Following Griffiths & Steyvers (2004), the idea of the paper is to look for objective evidence of organization in turbulent flow snapshots by identifying LDA topics or motifs. The relevant grey-level intensity is based on the value of Q_- (unlike in the work of Griffiths & Steyvers (2004), it corresponds to a physical field). We thus propose the following analogy: each scalar field observed in a collection of snapshots results from a mixture of N_T spatial topics that will be referred to as motifs in the remainder of the paper. This can be compared with the standard view that each realization of a turbulent flow is constituted of a random superposition of discrete eddies, characterized by a hierarchy of scales.

The paper is organized as follows. We show in § 2 how the POD method of snapshots, which is equivalent to latent semantic allocation (LSA), can be generalized to a probabilistic framework (probabilistic latent semantic allocation (PLSA)) which is then further extended into LDA in § 3. Application to the extraction of motifs for a turbulent channel flow is introduced in § 4 and results are discussed in § 5. The potential of the approach for flow reconstruction and flow generation is considered in § 6 before § 7 closes the paper.

2. A probabilistic extension of POD

To suitably introduce and contextualize the LDA, several established approaches to represent data are first briefly discussed.

2.1. *POD*

2.1.1. *General formulation*

Proper orthogonal decomposition is arguably the most popular tool for representation and analysis of turbulent flow fields (Lumley 1967).

The POD method allows us to derive an orthogonal basis for the (sub)space of the fluctuations of a multidimensional quantity f of finite variance. One can show that a basis for the space of fluctuations, defined as $f'(t) := f(t) - \langle f \rangle$, with $\langle \cdot \rangle$ the statistical mean, is given by the set of elements $\{\boldsymbol{\phi}_n\}_n$, eigenvectors of the following eigenvalue problem (Holmes *et al.* 1996):

$$C\boldsymbol{\phi}_n = \lambda_n \boldsymbol{\phi}_n, \tag{2.1}$$

with λ_n the eigenvalue and $C \in \mathbb{R}^{N_x \times N_x}$ the empirical two-point covariance matrix,

$$C = \frac{1}{N_s} \sum_{i=1}^{N_s} f'(t_i) f'(t_i), \tag{2.2}$$

with $\{t_i\}_i$ the time instants for which the field f is available. Some conditions on the temporal sampling scheme apply for the empirical covariance $\widehat{C}$ to be an accurate approximation of C (Holmes *et al.* 1996). Proper orthogonal decomposition modes are identified as the eigenvectors $\boldsymbol{\phi}_n$.

2.1.2. *Method of snapshots*

The above method is a quite natural implementation of the underlying Hilbert–Schmidt decomposition theory. However, the algorithmic complexity associated with the eigenvalue problem (2.1) scales as $O(N_s {N_x}^2)$, where the number of field instances N_s was assumed to be lower than the size N_x of the discrete field, $N_s \leq N_x$. For large field vectors (large N_x), the computational and memory cost is hence high. For this widely encountered situation, a possible workaround was suggested in Sirovich (1987) and consists of solving the following eigenvalue problem:

$$\tilde{C}\,\boldsymbol{a}_n = \lambda_n \boldsymbol{a}_n, \quad \boldsymbol{a}_n \in \mathbb{R}^{N_s}, \tag{2.3}$$

with

$$\tilde{C}_{i,i'} \propto \left\langle f'(t_i), f'(t_{i'}) \right\rangle_{\Omega}, \quad \forall\, i, i' \in [1, N_s] \subset \mathbb{N} \tag{2.4}$$

and $\langle \cdot, \cdot \rangle_{\Omega}$ the Euclidean inner product. Since the correlation matrix $\tilde{C}$ is Hermitian, its eigenvalues are real and non-negative, $\lambda_n \geq 0$, $\forall\, n$, and its eigenvectors $\{\boldsymbol{a}_n\}_n$ are orthogonal and can be made orthonormal in an Euclidean sense, $\boldsymbol{a}_n^{\mathrm{T}}\, \boldsymbol{a}_{n'} \propto \delta_{n,n'}$, with δ the Kronecker delta. The spatial POD modes are finally retrieved via projection as follows:

$$\boldsymbol{\phi}_n = \lambda_n^{-1/2}\, F'\, \boldsymbol{a}_n, \quad \forall\, n, \tag{2.5}$$

where the ith column of the matrix F' is the snapshot f_i'.

The algorithmic complexity is now $O({N_s}^3)$ and scales much better than the standard POD approach ($O(N_s {N_x}^2)$) in the usual situation where $N_s \ll N_x$. In this work, we rely on this so-called method of snapshots to implement POD.

Formally the decomposition of the snapshot matrix F' is equivalent to a singular value decomposition (SVD)

$$F' = \Phi\, \Sigma A^{\mathrm{T}}, \tag{2.6}$$

where Φ is the matrix constituted by the n columns $\boldsymbol{\phi}_n$.

Here, A is the matrix containing the n columns $\boldsymbol{a}_n$ and Σ is a diagonal matrix whose entries are $\lambda_n^{-1/2}$. The snapshot matrix can thus be decomposed into a snapshot-mode matrix A and into a cell-mode matrix Φ. The spatial modes or structures can be seen as latent variables allowing optimal reconstruction of the data in the L_2 norm or an equivalent. The decomposition can be truncated to retain only the N_T largest values corresponding to the N_T first columns of each matrix.

2.2. *PLSA*

In all that follows we will consider a collection of N_s scalar fields $\{f_i\}_{i=1,\ldots,N_s}$. Each field is of dimension N_x and consists of either positive or zero integer values on each grid cell. The key point is to interpret the value of the field f_i measured on a grid cell l of the snapshot i as the activity level of the grid cell i, or in other words the number of times (occurrences) the grid cell i can be considered as active. The representation of a snapshot is therefore equivalent to a list of cell indices i, each of which appears f_i times.

This representation assumes that each snapshot f_i consists of a mixture of structures z_n. Probabilistic latent semantic allocation adds a probabilistic flavour as follows:

(i) given a snapshot f_i, the structure z_n is present in that snapshot with probability $p(z_n|f_i)$;
(ii) given a structure z_n, the grid cell $\boldsymbol{x}_l$ is reported as active (i.e. appears in the list) with probability $p(\boldsymbol{x}_l|z)$.

Formally, the joint probability of seeing a given snapshot f_i and finding a grid cell $\boldsymbol{x}_l$ in the list of cells representing the snapshot is

$$p(f_i, \boldsymbol{x}_l) = p(f_i) \sum_n p(z_n|f_i) p(\boldsymbol{x}_l|z_n), \tag{2.7}$$

where $p(f_i)$, $p(z_n|f_i)$ and $p(\boldsymbol{x}_l|f_i)$ are the parameters of the model – $p(f_i)$ is the probability of obtaining such a snapshot f_i and is constant in our case, $p(f_i) = 1/N_s$; $p(z_n|f_i)$ and $p(\boldsymbol{x}_l|z_n)$ can be inferred using the expectation–maximization (EM) algorithm of Dempster, Laird & Rubin (1977).

Using Bayes' theorem, $p(f_i, \boldsymbol{x}_l)$ can be equivalently written as

$$p(f_i, \boldsymbol{x}_l) = \sum_n p(z_n) p(\boldsymbol{x}_l|z_n) p(f_i|z_n). \tag{2.8}$$

This alternative formulation shows a direct link between PLSA model and POD model (as mentioned above, POD is called latent semantic allocation, or LSA in text mining). The formulations in (2.6) and (2.8) are similar:

(i) the value of f at cell n for the snapshot i represents the number of times cell n will appear in the list of cells corresponding to snapshot i, and is therefore proportional to the probability of detecting cell n in snapshot i;
(ii) the structure probability $p(z_n)$ is the equivalent of the diagonal matrix entry Σ_n;
(iii) the probability of the snapshot f_i given the structure z_n, $p(f_i|z_n)$ corresponds to the snapshot-mode matrix entry $A_{i,n}$;
(iv) the probability to detect the cell $\boldsymbol{x}_l$ as active given the structure z_n, $p(\boldsymbol{x}_l|z_n)$, corresponds to the matrix entry $\Phi_{l,n}$.

3. LDA

3.1. *The probabilistic framework of LDA*

Latent Dirichlet allocation extends PLSA to address its limitations. Its specificity is as follows.

(i) The introduction of a probabilistic model for the collection of snapshots: each snapshot is now characterized by a distribution over the structures which will be now called motifs.
(ii) The use of Dirichlet distributions to model both motif–cell and snapshot–motif distributions.
The Dirichlet distribution is a multivariate probability distribution over the space of multinomial distributions. It is parameterized by a vector of positive-valued

parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)$ as follows:

$$p\,(x_1, \ldots, x_N; \alpha_1, \ldots, \alpha_N) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{n=1}^{N} x_n^{\alpha_n - 1}, \tag{3.1}$$

where B is a normalizing constant, which can be expressed in terms of the Gamma function Γ,

$$B(\boldsymbol{\alpha}) = \frac{\prod_{n=1}^{N} \Gamma(\alpha_n)}{\Gamma(\sum_{n=1}^{N} \alpha_n)}. \tag{3.2}$$

The support of the Dirichlet distribution is the set of N-dimensional discrete distributions, which constitutes the $N-1$ simplex. Introduction of the Dirichlet distribution allows us to specify the prior belief about the snapshots. The Bayesian learning problem is now to estimate $p(z_n, f_i)$ and $p(\boldsymbol{x}_l, z_n)$ from F given our prior belief $\boldsymbol{\alpha}$, and it can be shown that Dirichlet distributions offer a tractable, well-posed solution to this problem (Blei *et al.* 2003).

3.2. *Principles of LDA*

Latent Dirichlet allocation is therefore based on the following representation.

(i) Each motif z_n is associated with a multinomial distribution $\boldsymbol{\varphi}_n$ over the grid cells ($p(\boldsymbol{x}_l|z_n) = \varphi_{l,n}$). This distribution is modelled with a Dirichlet prior parameterized with an N_x-dimensional vector $\boldsymbol{\beta}$. The components β_l of $\boldsymbol{\beta}$ control the sparsity of the distribution: values of β_l larger than unity correspond to evenly dense distributions, while values lower than unity correspond to sparse distributions. In all that follows, we will assume a non-informative prior, meaning that $\boldsymbol{\beta} = \beta \mathbf{1}_{N_x}$.

(ii) Each snapshot f_i, is associated with a distribution of motifs $\boldsymbol{\theta}_i$ such that $\theta_{n,i} = p(z_n|f_i)$. The probabilities of each motif sum to unity in each snapshot. This distribution is modelled with an N_T-dimensional Dirichlet distribution of parameter $\boldsymbol{\alpha}$. The magnitude of $\boldsymbol{\alpha}$ characterizes the sparsity of the distribution (low values of α_n correspond to snapshots with relatively few motifs). The same assumption of a non-informative prior leads us to assume $\boldsymbol{\alpha} = \alpha \mathbf{1}_{N_T}$.

3.3. *Snapshot generation using LDA*

The generative process performed by LDA with N_T motifs is the following.

(i) For each motif z_n, a cell–motif distribution $\boldsymbol{\varphi}_n$ is drawn from the Dirichlet distribution of parameter β.

(ii) For each snapshot f_i:
   (a) a snapshot–motif distribution $\boldsymbol{\theta}_i$ is drawn;
   (b) each intensity unit $1 \leq j \leq N_i$ where N_i is the total intensity with $N_i = \sum_l f_{l,i}$ is then distributed among the different cells as follows. Firstly, a motif z_n is first selected from $\boldsymbol{\theta}_i$ (motif z_n occurs with probability $\theta_{n,i}$ in the snapshot). Secondly, for this motif, a cell l is chosen among the cells using $\varphi_{l,n}$ and the intensity associated with cell l is incremented by one.

The generative process can be summarized as follows.

Algorithm 1: LDA Generative Model.

```
for each of the N_T motifs n do
    sample φ_n ~ Dir(β)
end
for each of the N_s snapshots i do
    sample θ_i ~ Dir(α)
    for each of the N_i intensity units do
        1. sample a motif z_n from θ_{n,i}
        2. for this motif sample a cell l from φ_{l,n}
    end
end
```

The snapshot–motif distribution $\boldsymbol{\theta}_i$ and the cell–motif distribution $\boldsymbol{\varphi}_n$ are determined from the observed f_i. They are, respectively, N_T- and N_x-dimensional categorical distributions. Finding the distributions $\boldsymbol{\theta}_i$ and $\boldsymbol{\varphi}_n$ that are most compatible with the observations is an inference problem that can be solved by either a variational formulation (Blei *et al.* 2003) or a Markov chain Monte Carlo (known as MCMC) algorithm such as a Gibbs sampler (Griffiths & Steyvers 2002). In the variational approach, the objective function to minimize is the Kullback–Leibler divergence. The solution *a priori* depends on the number of motifs and on the values of the Dirichlet parameters α and β.

From a computational perspective, both Markov chain Monte Carlo and variational methods require the performance of several passes over each snapshot and cell to estimate the motif distribution, yielding a computational complexity of the order of $(N_x N_s N_T)$. In practice, stochastic approximations are used to speed up estimation, as is done in the implementation used for our experiments. It allows us to process snapshots one at a time or in small batches, to perform stochastic gradient steps, and achieves convergence much faster. Latent Dirichlet allocation has been shown to scale to millions of snapshots ($N_s > 10^6$), with tens of thousands of cells, and hundreds to thousands of motifs (Hofmann *et al.* 2013).

3.4. *A simple example*

We now give a short illustration of LDA for a one-dimensional example. We consider five basis functions represented in figure 1(*a*). These basis functions can be interpreted as motifs $\boldsymbol{\varphi}_n$. Using the generating algorithm described in the previous section, we created 5000 snapshots as random mixtures of the motifs. Each snapshot was obtained from 3000 individual draws (the number of draws was kept constant for the sake of simplicity). For each draw, we first sampled a topic n from a Dirichlet distribution of parameter 0.2, and we then determine a cell by sampling the corresponding distribution $\boldsymbol{\varphi}_n$. Examples of snapshots can be seen in figure 2. Application of LDA to the collection of snapshots for a number of five motifs yields the motifs represented in figure 1(*b*), which can be seen to be a relatively good approximation of the five original ones. It can be seen that the motifs are very different from POD eigenmodes, represented in figure 3.

3.5. *Remarks*

We conclude this section with two remarks.

(i) Latent Dirichlet allocation can generalize to new snapshots more easily than PLSA, due to the snapshot–motif distribution formalism. In PLSA, the snapshot

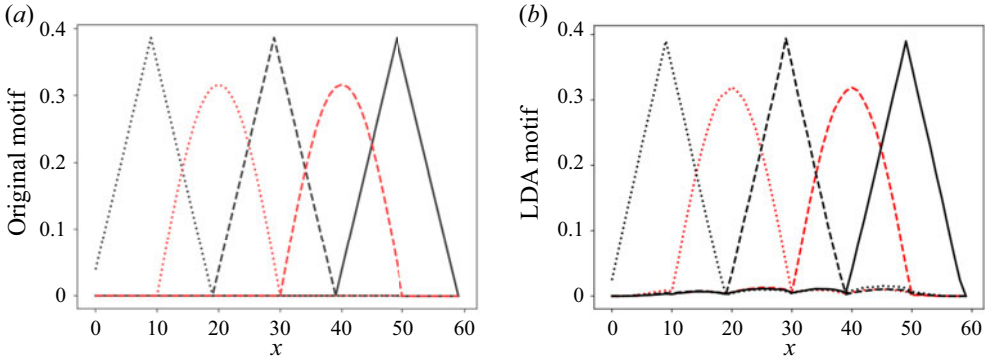


Figure 1. (*a*) Original motifs; (*b*) motifs recovered through application of LDA.

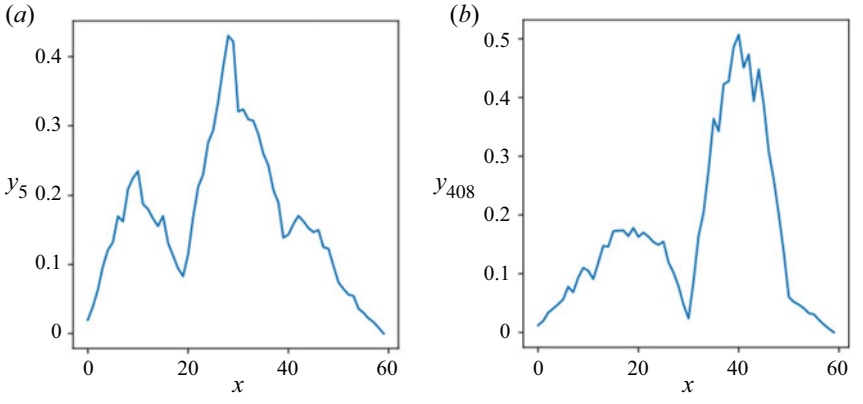


Figure 2. Selected one-dimensional snapshots.

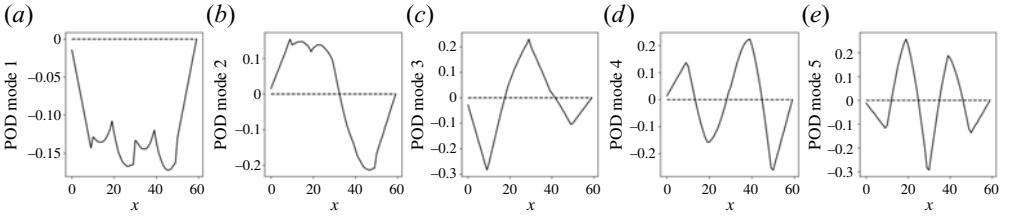


Figure 3. Proper orthogonal decomposition modes extracted from the collection.

probability is a fixed point in the dataset, which cannot be estimated directly if it is missing.

In LDA, the dataset serves as training data for the Dirichlet distribution of snapshot–motif distributions. If a snapshot is missing, it can easily be sampled from the Dirichlet distribution instead.

(ii) An alternative viewpoint can also be adopted in interpreting the LDA in the form of a regularized matrix factorization (MF) method. This is further discussed in Appendix A.

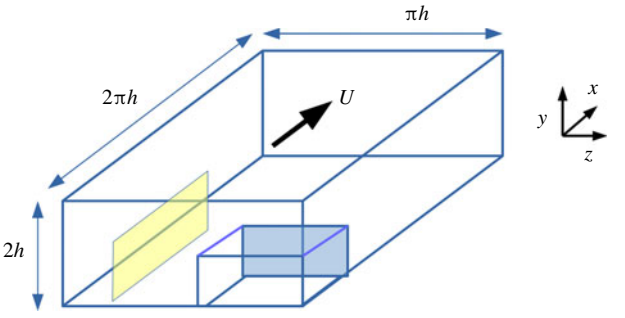


Figure 4. Numerical domain D. The shaded surfaces correspond to the two types of planes used in the analysis. The volume considered for 3-D analysis is indicated in bold lines.

4. Application of LDA to turbulent flows

4.1. *Numerical configuration*

The idea of this paper is to apply this methodology to snapshots of turbulent flows in order to determine latent motifs from observations of Q_- events. We will consider the configuration of turbulent channel flow at a moderate Reynolds number of $R_\tau = u_\tau h/\nu = 590$ (Moser, Kim & Mansour 1999; Muralidhar *et al.* 2019), where R_τ is the Reynolds number based on the fluid viscosity ν, channel half-height h and friction velocity u_τ. Wall units based on the friction velocity and fluid viscosity will be denoted with a subscript $_+$. The streamwise, wall-normal and spanwise directions will be referred to as x, y and z, respectively. The horizontal dimensions of the numerical domain are $(2\pi, \pi)h$. Periodic boundary conditions are used in the horizontal directions. The resolution of $(256)^3$ points is based on a uniform spacing in the horizontal directions, which corresponds to streamwise and spanwise spacings of $\Delta_{x+} = 7.24$ and $\Delta_{z+} = 3.62$, and a hyperbolic tangent stretching function for the vertical direction. More details can be found in Podvin & Fraigneau (2017). The configuration is shown in figure 4. More details about the numerical simulation can be found in Muralidhar *et al.* (2019).

4.2. *LDA inputs*

In this section, we introduce the different parameters of the study. The python library scikit-learn (Pedregosa *et al.* 2011) was used to implement LDA. The sensitivity of the results to these parameters will be examined in a subsequent section.

We first focus on two-dimensional (2-D) vertical subsections of the domain, then present 3-D results. The vertical extent of the domain of investigation was the half-channel height. Since this is an exploration into a new technique, a limited range of scales was considered in the horizontal dimensions: the spanwise dimension of the domain was limited to 450 wall units. The streamwise extent of the domain was in the range of 450–900 wall units. The number of realizations considered for 2-D analysis was $N_s = 800$, with a time separation of 60 wall time units. The number of snapshots was increased to 2400 for 3-D analysis.

The scalar field f of interest corresponds to Q_- events. It is defined as the positive part of the product $-u'v'$, where fluctuations are defined with respect to an average taken over all snapshots and horizontal planes. The LDA procedure requires that the input field consists of integer values: it was therefore rescaled and digitized and the scalar field f was

defined as

$$f = [A\tau_-], \tag{4.1}$$

where $\tau_- = \max(-u'v', 0)$ and $[\cdot]$ represents the integer part. The rescaling factor A was chosen in order to yield a sufficiently large, yet still tractable, total intensity. In practice we used $A = 40$, which led to a total intensity $\sum_i \sum_l f_{l,i}$ of approximately 10^8 for plane sections and corresponds to 1516 discrete intensity levels. The effect of the rescaling factor will be examined in a subsequent section.

Latent Dirichlet allocation is characterized by a user-defined number of motifs N_T, a parameter α which characterizes the sparsity of prior Dirichlet snapshot–motif distribution, and a parameter β which characterizes the sparsity of the prior Dirichlet motif–cell distribution. Results were obtained assuming uniform priors for α and β with a default value of $1/N_T$. The sensitivity of the results to the priors will be evaluated in § 5.2.

4.3. *LDA outputs*

For a collection of N_s snapshots and a user-defined number of motifs N_T, LDA returns N_T motif–cell distributions $\boldsymbol{\varphi}_n$ and N_s snapshot–motif distributions $\boldsymbol{\theta}_i$. Each motif is defined by a probability distribution $\boldsymbol{\varphi}_n$ associated with each grid cell. It is therefore analogous to a structure or a portion of structure since it contains spatial information – note, however, that its definition is different from standard approaches. The motif–snapshot distribution $\boldsymbol{\theta}_i$ characterizes the prevalence of a given motif in the snapshot.

As will be made clear below, the motifs most often consist of single connected regions, although occasionally a couple of distinct regions were identified. In most cases, the motifs can thus be characterized by a characteristic location $\boldsymbol{x}^c$ and a characteristic dimension in each direction $L_j, j \in \{x, y, z\}$.

To determine these characteristics, we first define for each motif a mask associated with a domain D. The origin of the domain was defined as the position, $\boldsymbol{x}_m$ corresponding to its maximum probability $p_m = \boldsymbol{\varphi}_n(\boldsymbol{x}_m)$. The dimensions of the domain in each direction (for instance L_x) were defined as the segment extending from the domain origin over which the probability remained larger than 1 % of its maximum value p_m. The position and characteristic dimension of a motif for instance in the x-direction are then defined as

$$x^c = \frac{\int_D x\boldsymbol{\varphi}_n \mathrm{d}D}{\int_D \boldsymbol{\varphi}_n \mathrm{d}D}, \tag{4.2}$$

$$L_x^2 = 4\frac{\int_D (x - x^c)^2 \boldsymbol{\varphi}_n \mathrm{d}D}{\int_D \boldsymbol{\varphi}_n \mathrm{d}D}. \tag{4.3}$$

Analogous definitions can be given for y^c and z^c.

5. Results

5.1. *Vertical planes*

In order to investigate in detail the vertical organization of the flow, LDA was first applied to vertical sections of the flow. Both cross-flow (y, z) and longitudinal (x, y) sections were considered. Due to the horizontal homogeneity of the flow, we do not expect significant changes in the cell–motif and the motif-document distributions when the sections are translated in the horizontal direction.

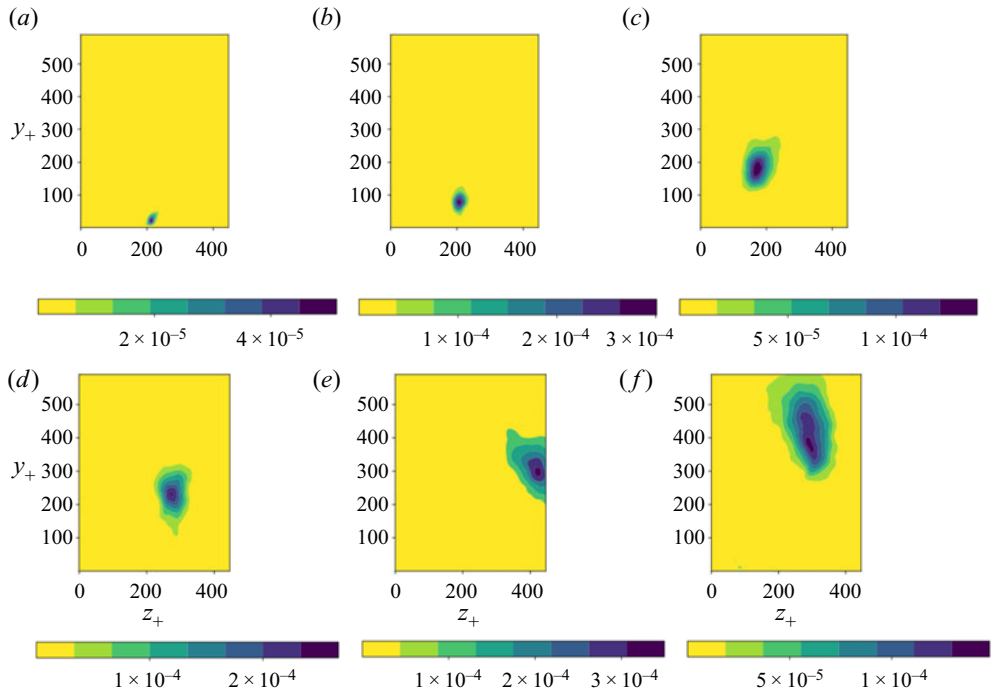


Figure 5. Probability contours of selected motifs in a cross-flow plane for a number of motifs $N_T = 96$. The probability integrates to unity over the domain.

5.1.1. *Cross-flow planes*

The dimensions of the cross-sections were $d_{z+} = 450$ and $d_{y+} = 590$. Figure 5 shows selected motifs for a total number of motifs $N_T = 96$ on a vertical plane at $x = 0$. What is represented is the spatial distribution of the motif among the cells. If a given motif appears in a snapshot, all the cells associated with a high cell–motif probability are likely to contribute to the Reynolds stress measured in the snapshot. The motifs consist of isolated regions, the dimensions of which increase with the wall distance. This is confirmed by figure 6, which represents characteristic sizes of LDA motifs of a succession of four vertical planes separated by a distance of 100 wall units (+). The characteristic dimensions are computed from (4.2)) and ((4.3).

We point out that observing motifs which are detached from the wall does not invalidate the presence of wall-attached structures, as they would be consistent with a cross-section of a wall-attached structure elongated in the streamwise direction.

Results for several motif numbers (three different motif numbers $N_T = 48, 96, 144$ are shown in figure 6). Characteristic sizes were compared with a linear fit and a $y^{1/2}$ fit, which is the scaling of the Taylor microscale (Jimenez 2013). Close to the wall, the linear scaling seems to hold for the vertical dimension but is not as clear for the spanwise dimension. Farther away, it was found that both spanwise and vertical dimensions increase linearly with the wall distance in the region $y_+ > 100$. Again, this is in agreement with the Townsend (1961) hypothesis of a hierarchy of structures of increasing dimensions, which was also confirmed numerically by Flores & Jimenez (2010). We note that the size of the motifs in the upper boundary layer may be somewhat underestimated, as Reynolds stress events at these heights are likely to cross over into the other half-channel, as was observed by Lozano-Duran *et al.* (2012). The aspect ratio L_z/L_y is constant with the wall

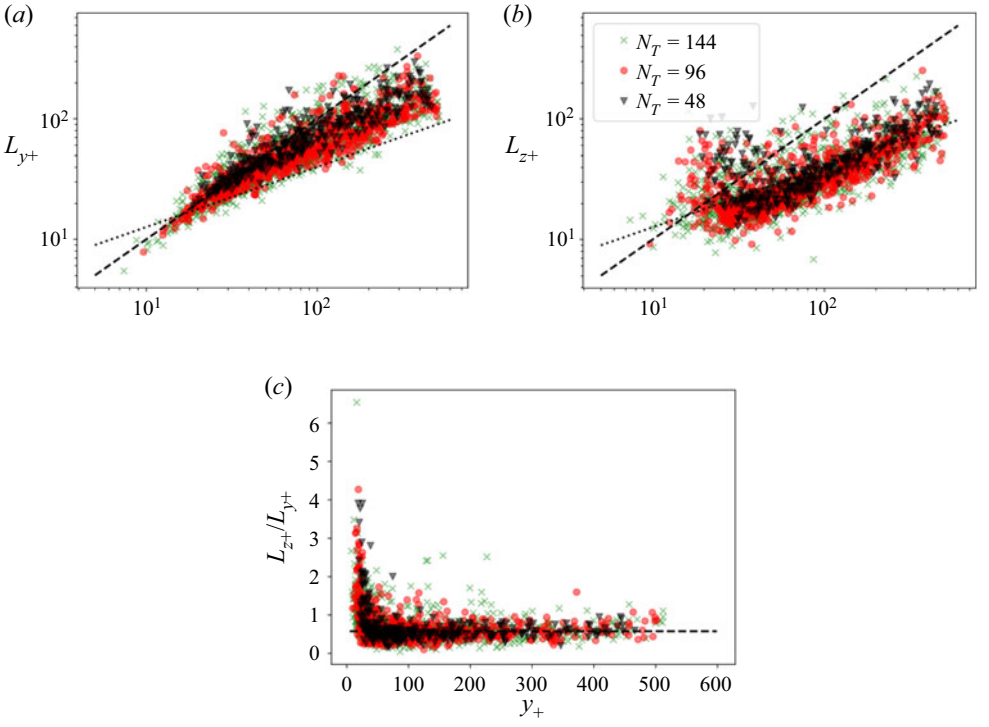


Figure 6. (*a*,*b*) Cross-plane motif characteristic sizes: (*a*) vertical dimension L_y; (*b*) spanwise dimension L_z. The dashed line corresponds to the fit y_+ and the dotted line to a $4y_+^{1/2}$ fit. (*c*) Aspect ratio L_y/L_z. The horizontal line corresponds to the mean value 0.6. Each dot corresponds to a motif.

distance above $y_+ > 100$, with a typical value of 0.6 ± 0.2. We note that Lozano-Duran *et al.* (2012) found with an different definition that Q_- events were characterized by nearly equal spanwise and vertical sizes $\Delta z \sim \Delta y$, while Alamo *et al.* (2006) found a scaling of $\Delta z \sim 1.5\Delta y$ for vortex clusters. The discrepancy can be due to the fact that motifs are typically tilted at a slight angle to the coordinate axes, and that Lozano-Duran *et al.* (2012) use circumscribing boxes, while we use (4.3) to compute standard deviations. If the motif was a rectangle of horizontal and vertical dimensions a and b, tilted by an angle θ with respect to the horizontal axis, the vertical to horizontal aspect ratio of the enclosing box would be $(a\tan\theta + b)/(a + b\tan\theta)$, while our definition would give $[(a\tan\theta)^2 + b^2)/(a^2 + (b\tan\theta)^2)]^{1/2}$.

It would be interesting to determine whether the linear scaling corresponds to structures that are attached or detached from the wall. Chandran *et al.* (2020) found that a major contribution to the cross-stream velocity spectra was made by wall-detached eddies, which scale with the wall distance but are not attached to the wall. Meanwhile, Lozano-Duran *et al.* (2012) showed that 60 % of the Reynolds stress was carried out by wall-attached structures. Although it is tempting to identify the isolated regions corresponding to LDA motifs with detached structures, it is not possible to give a definitive answer at this point since the analysis is 2-D, and, furthermore, limited to a relatively small section. We will come back to this point in § 5.3.

Figure 7 shows the distribution of the vertical location $p(y_m)$ of the motif maximum probability. Comparison of two different plane locations x confirms that results do not depend on the location of the plane, which reflects the statistical homogeneity of the flow

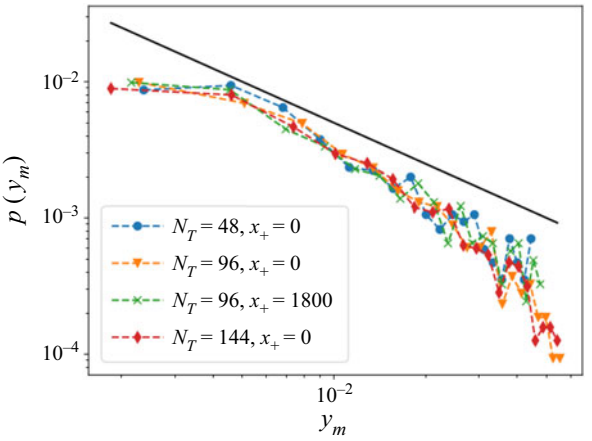


Figure 7. Distribution of the motif maximum location y^c. The line corresponds to a y^{-1} fit.

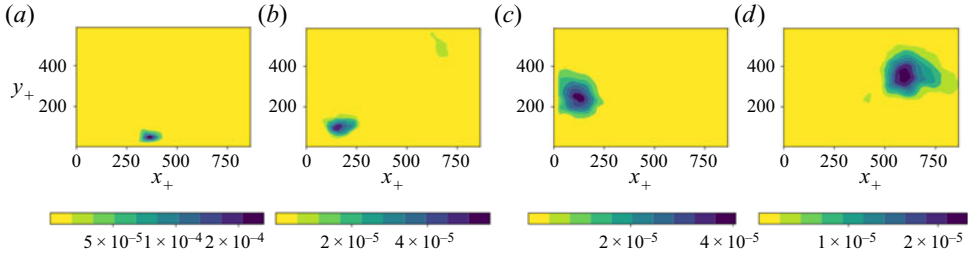


Figure 8. Probability contours of selected motifs for a longitudinal plane for a number of motifs $N_T = 96$. The probability integrates to unity over the domain.

in the horizontal direction. The probability decreases roughly as the inverse of the wall distance on all planes. This is in agreement with Townsend's self-similarity hypothesis that the number of structures decreases with the wall distance as $1/y$ (Townsend 1961; Woodcock & Marusic 2015). A slight discrepancy with a purely inverse proportional fit is observed in the results, but it is not clear whether this is due to the relatively moderate Reynolds number and the limited size of the logarithmic region.

5.1.2. *Longitudinal planes*

We now examine results for the longitudinal sections (x, y). The streamwise and vertical dimensions of the sections are, respectively, $d_{x+} = 900$ and $d_{y+} = 590$ wall units, although some tests were also carried out for a streamwise extent of 450 units. Figure 8 presents selected motifs for the longitudinal planes for $N_T = 96$. As in the cross-flow plane, for $y_+ > 100$, the dimensions of the motifs increase with the wall distance, which is confirmed by figure 9. The characteristic dimensions seem essentially independent of the total number of motifs (see also next section). There is a wide disparity in streamwise characteristic dimensions near the wall. The streamwise dimension appears to be nearly constant there. The motif aspect ratio is highest near the wall and decreases sharply in the region $0 < y_+ < 50$. The vertical dimension increases linearly with the wall distance in the region $y_+ > 100$, as well as the streamwise dimension, with an aspect ratio of L_x/L_y of the order of unity (with a mean value of 1.1 and a standard deviation of 0.4).

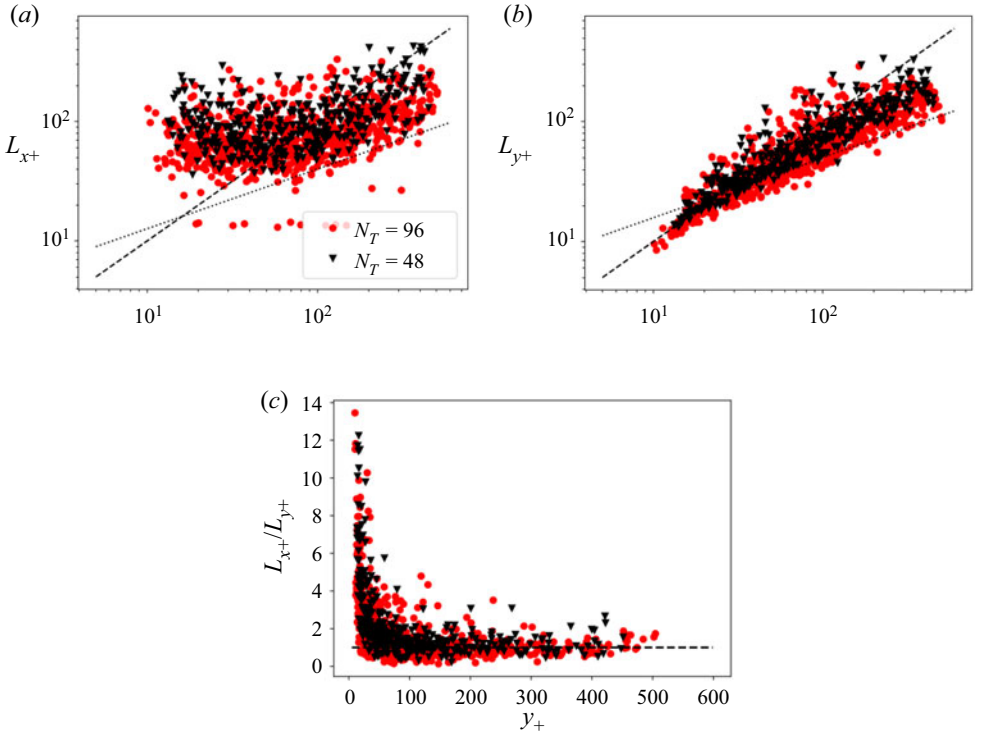


Figure 9. (*a*,*b*) Longitudinal motif characteristic dimensions: (*a*) streamwise dimension L_x; (*b*) vertical dimension L_y. The dashed line corresponds to $y+$ and the dotted line to $4y_+^{1/2}$. (*c*) Aspect Ratio L_x/L_y. The dashed line corresponds to the mean value of 1.1. Each dot corresponds to a motif.

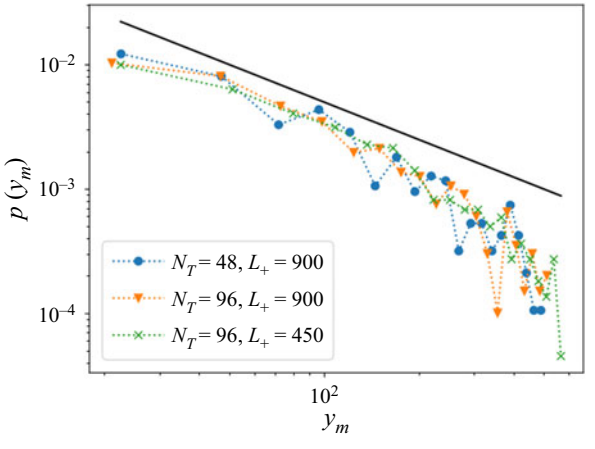


Figure 10. Distribution of the motif maximum location y^c – the line corresponds to a y^{-1} fit.

Figure 10 shows the distribution of the motif maximum probability location for two different sets $N_T = 48, 96$, and for two domain lengths. The shape of the distribution does not appear to change, and again fits well with the distribution $p \simeq 1/y$.

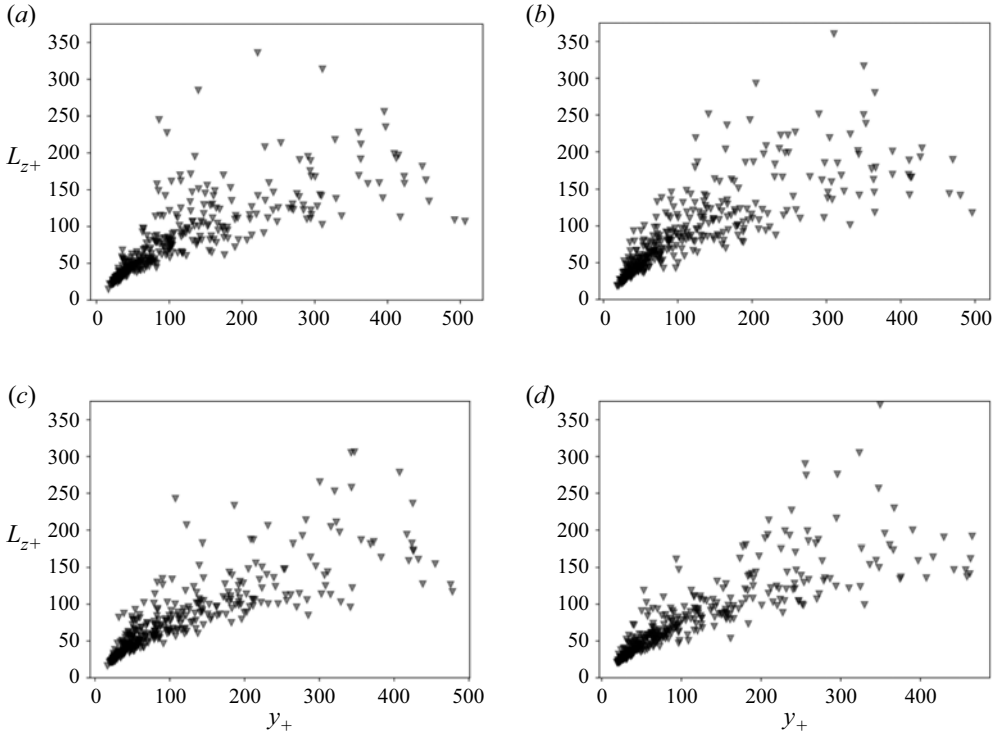


Figure 11. Motif characteristic vertical dimension for $N_T = 48$. (*a*,*b*) Influence of dataset size, N_s: (*a*) N_s= 800; (*b*) $N_s = 400$. (*c*,*d*) Effect of rescaling factor: (*c*) $A = 60$; (*d*) $A = 20$.

5.2. *Sensitivity of the results*

In this section we examine if and how the characteristics of the motifs depend on the various parameters of LDA. We point out that the probabilistic framework of the model makes exact comparison difficult, since there is no convergence in the L_2 sense, and the Kullback–Leibler divergence, which measures the difference between two distributions is not a true metric tensor (see appendices).

The criteria we chose to assess the robustness of the results were the characteristic size of the topics and the distribution of their locations. We first examine the influence of various LDA parameters on the results obtained for cross-flow sections for a constant number of topics $N_T = 48$. The reference case corresponded to an amplitude $A = 40$, prior values of $\alpha = \beta = 1/N_T$ and a total number of snapshots $N_s = 800$.

Figure 11(*a*,*b*) shows that the characteristic dimension is not modified when the number of snapshots was reduced by 50 %, indicating that the procedure has converged. Figure 11(*c*,*d*) shows the characteristic vertical dimension L_y of the structures when the rescaling parameter A was varied. Similar results (not shown) were found for L_z. Although some fluctuations were observed in the individual characteristic dimensions, no significant statistical change was observed. Figure 12 shows the characteristic dimensions of the structures for different prior choices for α and β, which govern the sparsity of the representation. No significant statistical trend was modified when α and β were made to vary within $1/10$ and up to 10 times their default values of $1/N_T$. Figure 13 shows that the distribution of the maximum location of the motifs follows the same inverse law and does not depend on the choice of parameters chosen for LDA.

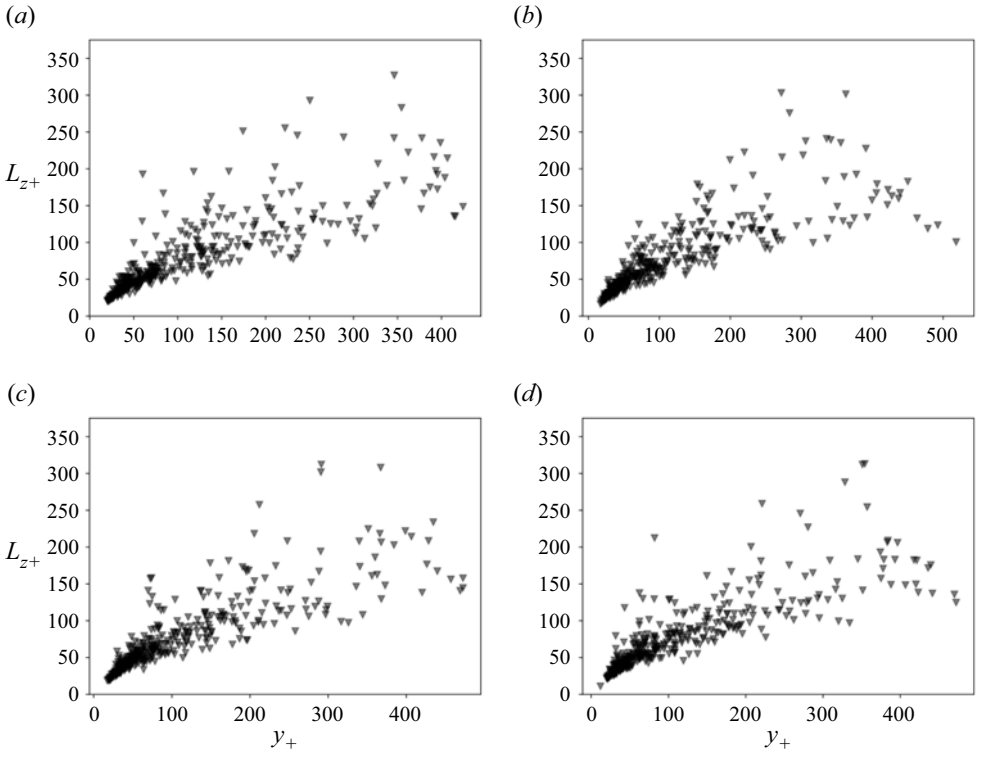


Figure 12. Characteristic vertical motif length for different LDA priors, $N_T = 48$: (*a*) $\alpha = 0.1/N_T$, $\beta = 1/N_T$; (*b*) $\alpha = 10/N_T$, $\beta = 1/N_T$; (*c*) $\alpha = 1/N_T$, $\beta = 0.1/N_T$; (*d*) $\alpha = 1/N_T$, $\beta = 10/N_T$.

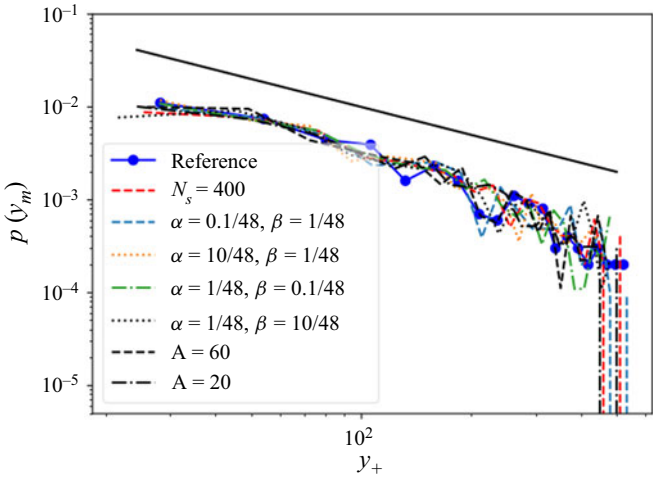


Figure 13. Distribution p of motif/cell distribution maximum y_m for different parameters. The solid line corresponds to y^{-1}.

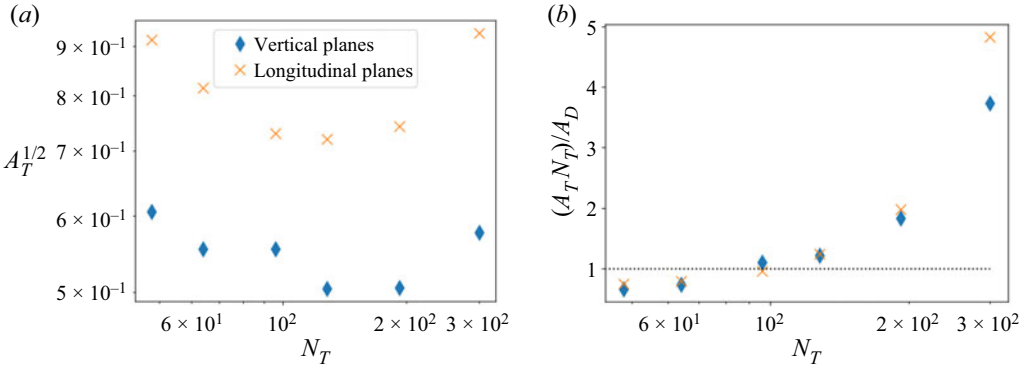


Figure 14. (*a*) Motif characteristic dimension L_T for different datasets as a function of the number of motifs; (*b*) relative fraction of the area captured by the sum of the topics $N_T A_T / A_D$.

We now study if and how the characteristics of the motifs are modified when the total number of topics N_T varies. Both types of vertical planes are considered. We have seen in the previous sections that the motif dimensions appear essentially independent of the number of motifs considered. To quantify this more precisely, we first define a characteristic motif size L_T as $L_T = \sqrt{\langle A_T \rangle}$ where A_T is the area corresponding to the ellipsoid with the same characteristic dimensions as the motif and $\langle \cdot \rangle$ represents the average over the motifs. Figure 14 summarizes how the motif size evolves with the number of motifs for both vertical and longitudinal planes. In all cases, it was found that the characteristic size varies slowly around a minimal value (figure 14*a*), and that the characteristic area of the motif was a minimum when the sum of the motif characteristic areas $N_T A_T$ was comparable with the total domain area A_D (figure 14*b*).

5.3. *3-D analysis*

Latent Dirichlet allocation was then applied to a volumetric section of the flow of size $450 \times 590 \times 450$ wall units. We note that the volume of study is limited: in particular the streamwise direction is short compared with the typical extent of a coherent structure carrying the Reynolds stress, which typically spans a few hundred wall units (Zhou *et al.* 1999). As a result, it is difficult to capture full-length structures. We note that the structures carrying the streamwise velocity fluctuations are even longer since they can extend over a few thousand wall units (Robinson 1991; Jimenez 2013).

Figure 15 shows isocontours of the probability for different topics. One can see that some consist of streaks or 'legs' with a strong streamwise coherence near the wall (15*a–c*) while others consist of lumps farther away from the wall (15*d–f*), which may be reminiscent of 'heads' and are not always connected to the wall. It is unclear whether the lack of connection between 'legs' and 'heads' is an artefact of the limited domain size or whether this constitutes evidence of wall-detached coherent structures. An investigation over a larger domain will be necessary in order to settle this question.

The characteristic dimensions of the motifs are reported in figure 16. Since the number of samples is small, we checked that the trends reported below held for 2-D motifs identified on wall-parallel planes at different heights. Further, L_y appears to grow linearly over both regions. Two different regions can be identified based on the evolution of the horizontal dimensions. The critical height separating the two regions is of the order of $y_+ \sim 100$, in agreement with experimental observations (Stanislas, Perret &

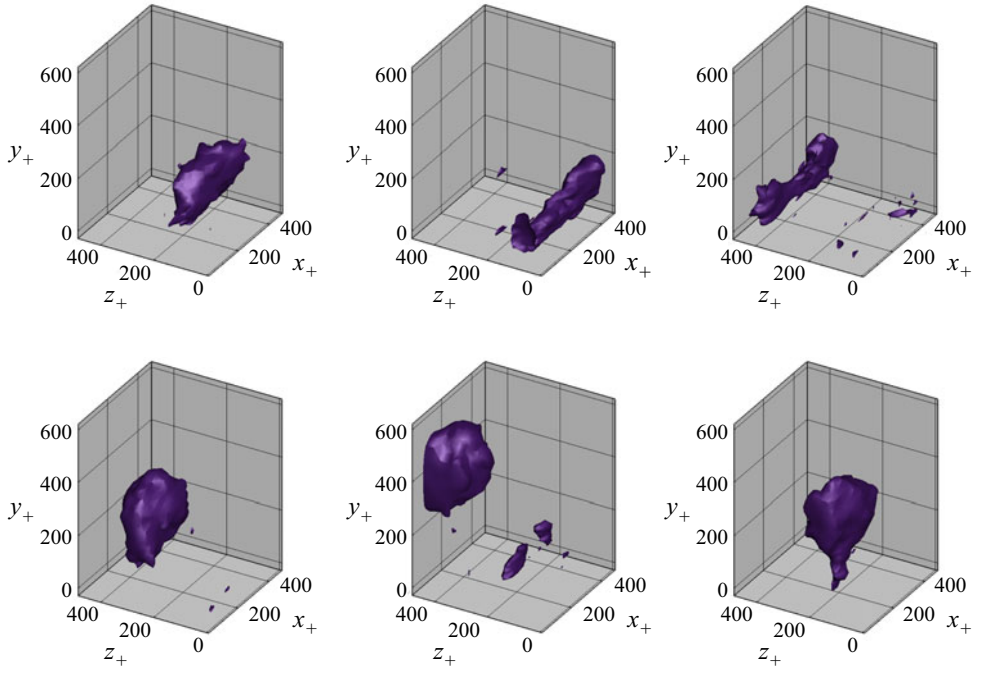


Figure 15. Examples of 3-D topics represented by their isoprobability contour $p = 16\ 10^{-5}$. The topics indexed with N were selected out of a total number of topics $N_T = 48$. The probability for each topic sums to unity over the domain. Here (*a*) $N = 33$; (*b*) $N = 38$; (*c*) $N = 42$; (*d*) $N = 46$; (*e*) $N = 7$; (*f*) $N = 16$.

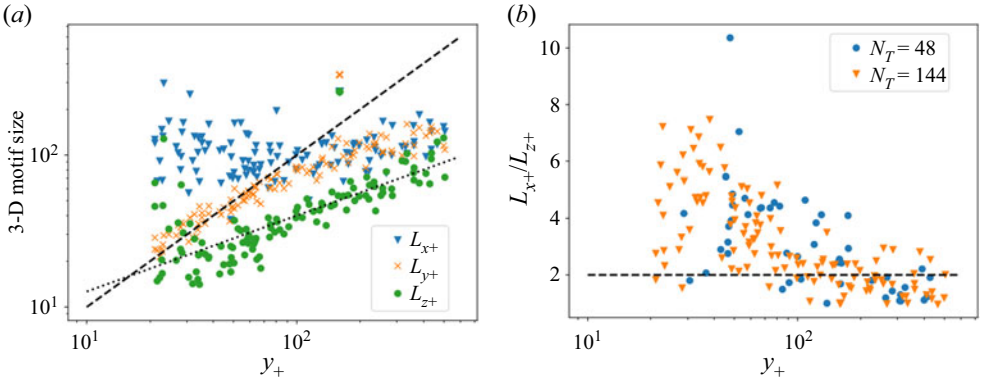


Figure 16. (*a*) Characteristic dimensions of the 3-D motifs, $N_T = 144$; the dashed line corresponds to the fit y_+ and the dotted line to a $4y_+^{1/2}$ fit. (*b*) Evolution of ratio L_x/L_z with height for $N_T = 144$ and $N_T = 48$. The dashed line corresponds to a mean value of 2.

Foucaut 2008). For $y_+ \lessapprox 100$ the region is characterized by a wide distribution of L_x and L_z, with a ratio of L_x/L_z that varies between 2 and 8. For $y_+ \gtrapprox 100$ both L_x and L_z appear to increase nearly linearly with a ratio that decreases slowly to a value of the order of 2 (with a standard deviation of 1). This ratio is consistent with results from analysis of POD eigenfunctions in Podvin *et al.* (2010), as well as from vortex cluster analysis from Alamo *et al.* (2006). Three-dimensional motif characteristic sizes are consistent with those obtained for vertical planes, which shows that information about the 3-D organization of

the flow can be obtained from analysis performed on 2-D sections. This is of particular interest as it suggests that the LDA method could be usefully applied to particle image velocimetry experimental data limited to a plane. We also note that the data do not have to be resolved in time.

6. Field reconstruction and generation

6.1. *Reconstruction*

We now examine how the flow can be reconstructed using LDA. In all that follows, without loss of generality, we will focus on one of the cross-flow planes examined in § 5, specifically the cross-section at $x = 0$ of dimensions $d_{y+} = 590$ and $d_{z+} = 450$. As described in the algorithm presented in § 3, both the motif–snapshot and the cell–motif distributions can be sampled for the total intensity $N_i = \sum_l f_{l,i}$ in the ith snapshot. This total intensity is defined as the rescaled integral value of the Reynolds stress (digitized and restricted to Q_- events) over the plane. Since results were found to be essentially independent of the rescaling, we can make the simplifying assumption that N_i is large enough so that the distribution φ_n is well approximated by the samples. For a given total intensity N_i, a reconstruction of the ith snapshot can then be obtained at each grid cell $\boldsymbol{x}_l$ from

$$\tau^{R\text{-}LDA}(\boldsymbol{x}, t_i) = \frac{1}{A} f_i(\boldsymbol{x}) = \frac{N_i}{A} \sum_{n=1}^{N_T} \theta_{n,i} \varphi_n(\boldsymbol{x}), \tag{6.1}$$

where

(i) $\varphi_n(\boldsymbol{x})$ is the motif–cell distribution;
(ii) the snapshot–motif distribution $\theta_{n,i}$ represents the likelihood of motif $\boldsymbol{z}_n$ in the ith snapshot.

It seems natural to compare this reconstruction with the POD representation of the field which has a similar expression,

$$\tau^{R\text{-}POD}(\boldsymbol{x}, t_i) = \sum_{n=0}^{N_{POD}-1} a_{n,i} \phi_n(\boldsymbol{x}), \tag{6.2}$$

where

(i) $\phi_n(\boldsymbol{x})$ are the POD eigenfunctions extracted from the autocorrelation tensor $C_{i,i'}$ obtained from the N_s snapshots;
(ii) $a_{n,i}$ corresponds to the amplitude of the nth POD mode in the ith snapshot.

The first six fluctuating POD modes are represented in figure 17. We note that the zeroth POD mode represents the temporal average of the field. As expected, the fluctuating POD modes consist of Fourier modes in that spanwise direction (due to homogeneity of the statistics), and their intensity reaches a maximum at around $y_+ \simeq 25$.

If the number of POD modes is equal to the number of motifs N_T, by construction, POD will provide a better representation of the statistics at least up to second order (Holmes *et al.* 1996). We note that, in terms of computational requirement, POD is less expensive than LDA, as it requires solving an SVD problem versus implementing an iterative EM algorithm (Dempster *et al.* 1977). However, the performance of the EM algorithm can be improved, in particular with online updates (Hofmann 1999).

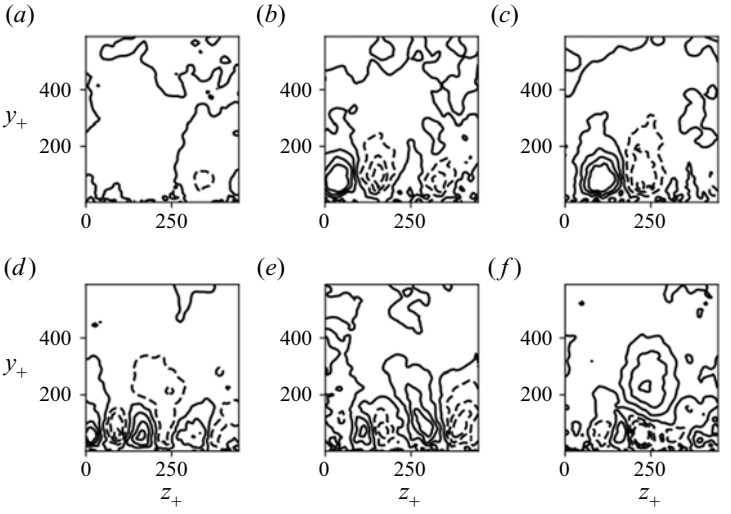


Figure 17. Contour plot of the first six fluctuating normalized POD spatial modes; contour values go from −0.03 to 0.03. Negative values are indicated by dashed lines. (*a*) $n = 1$, (*b*) $n = 2$, (*c*) $n = 3$, (*d*) $n = 4$, (*e*) $n = 5$, (*f*) $n = 6$.

In terms of storage, a reconstructed snapshot requires N_{POD} modes for POD and N_T topics for LDA. However, storage reduction could be obtained in the case of LDA by filtering out the motifs with a low probability $\theta_{n,i}$, i.e. lower than a threshold κ. We note that, in this case, it is necessary to store the indices n of the motifs as well as the value of $\theta_{n,i}$, so that if n modes (respectively, topics) are kept, storage will consist of $2n$ variables per snapshot. We see that storage reduction can be achieved if the fraction of retained modes $\eta = n/N_T$ is sufficiently small. The LDA storage data length per snapshot $2\eta N_T$ should then be compared with the POD data length N_{POD}.

For $N_T = 96$, choosing a threshold of $\kappa = 0.015$ resulted in less than 8 % difference between the filtered and unfiltered LDA reconstructions (the L_2 norm was used). The average value for η was 0.2, which means that the number of POD modes that would represent a storage equivalent to that of LDA with $N_T = 96$ is $N_{POD} \simeq 2\eta N_T \simeq 40$. We note that the total storage cost should further take into account the size of the LDA basis $\{\boldsymbol{z}_n\}_n$, which will be larger than the POD basis $\{\boldsymbol{\phi}_n\}_n$ since they are, respectively, equivalent to N_T and N_{POD} fields. However, efficient storage of the LDA basis can be achieved by making use of the limited spatial support of $\boldsymbol{z}_n$, in particular for motifs located close to the wall.

In the remainder of this section we will compare a filtered LDA reconstruction of 96 motifs (where values of $\theta_{n,i}$ lower than $\kappa = 0.015$ are excluded from the reconstruction), with a POD representation of $N_{POD} = 48$ modes, which captures approximately 75 % of the total energy. Figure 18 compares an instantaneous field with its LDA reconstruction and its POD reconstruction. A more general assessment is provided by figure 19, which shows the correlation coefficient between each snapshot and its reconstruction based on POD as well as that based on LDA. Although POD appears to be slightly superior, as can be expected from its optimality property, the correlation coefficients are very close with respective average values of 0.75 for LDA and 0.77 for POD.

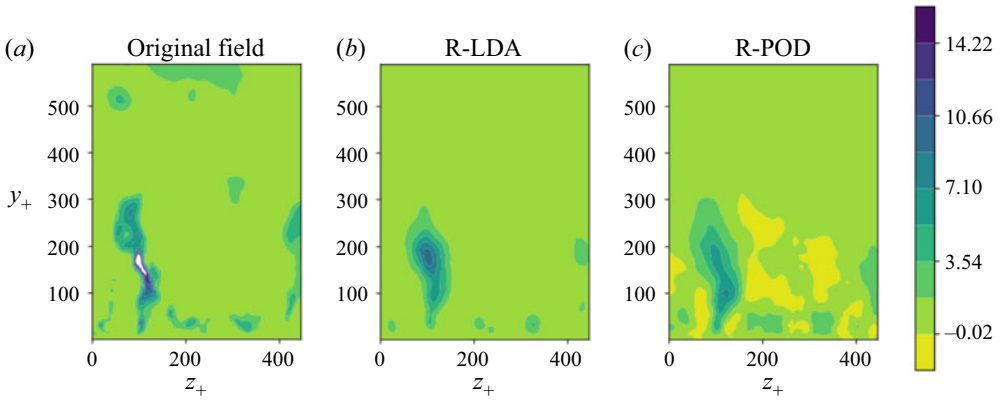


Figure 18. Instantaneous Reynolds stress field (limited to Q_- events) (*a*) True field; (*b*) LDA-reconstructed field using 96 modes; (*c*) POD-reconstructed field using 48 modes.

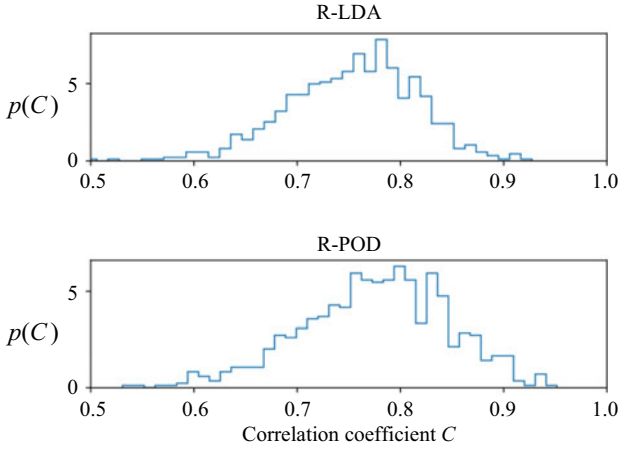


Figure 19. Distribution of the correlation coefficient between each original snapshot and its reconstruction based on (*a*) LDA or (*b*) POD.

6.2. *Generation*

Latent Dirichlet allocation is a generative model, so it is straightforward to generate synthetic snapshots by sampling from distributions θ and φ for a total intensity $N_i = \sum_l f_{l,i}$, which is modelled as a Poisson process with the same mean and standard deviation as the original database.

In contrast, POD is not a generative model *per se*. We will use a simplified version of the probabilistic extension of POD (PPCA) derived by Tipping & Bishop (1999), which is presented in Appendix B, where we will make the additional assumption that no noise is present in the model, POD-based synthetic fields will be reconstructed from deterministic spatial POD modes $\boldsymbol{\phi}_n$ and random POD amplitudes $\boldsymbol{a}_n$ which are assumed to be Gaussian variables. Examination of figure 20, which represents the distribution of the first fluctuating POD coefficients $n \geq 1$, suggests that it is quite acceptable as a first approximation to assume Gaussian distributions for the amplitudes $\boldsymbol{a}_n$ – alternatively, the amplitudes could be sampled from the empirical distributions. The amplitude of the zeroth

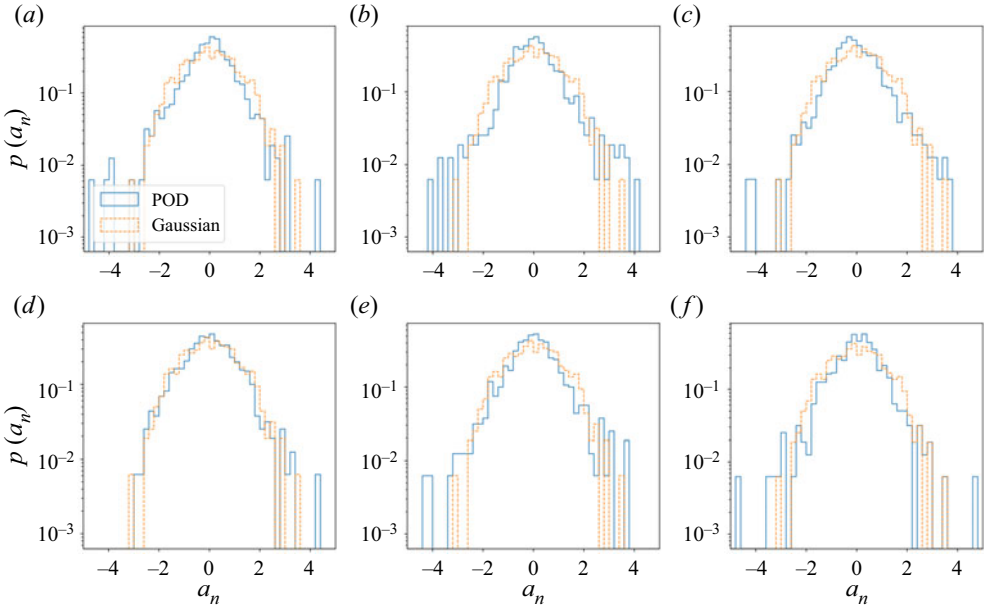


Figure 20. Histograms of the normalized amplitudes of the first six fluctuating POD modes and comparison with a sampled Gaussian distribution. (*a*) $n = 1$, (*b*) $n = 2$, (*c*) $n = 3$, (*d*) $n = 4$, (*e*) $n = 5$, (*f*) $n = 6$.

mode, which corresponds to the average of the field over the snapshots, will be assumed to be constant for all snapshots.

We can therefore compare the databases reconstructed from and generated with LDA with those obtained from POD. The generated databases consist of N_s snapshots corresponding to arbitrary instants $\tilde{t}_i$. Overall, the statistics of five different databases can be compared:

(i) the true database $\tau_-(y, z, t_i)$ corresponding to the actual values of the Q_- events;

(ii) the POD-reconstructed (R-POD) or POD-projected database,

$$\tau_-^{R\text{-}POD}(y, z, t_i) = \sum_{n=0}^{N_{POD}-1} a_{i,n}\phi_n(y, z), \tag{6.3}$$

where ϕ_n are the POD eigenfunctions and $a_{i,n}$ are the amplitudes of the nth POD mode in the ith snapshot;

(iii) the POD-generated (G-POD) database,

$$\tau_-^{G\text{-}POD}(y, z, \tilde{t}_i) = \sum_{n=0}^{N_{POD}-1} \tilde{a}_{i,n}\phi_n(y, z), \tag{6.4}$$

where $\tilde{a}_{i,0} = \langle a_{i,0}\rangle$, with $\langle\cdot\rangle$ the average over all snapshots and $\tilde{a}_{i,n}$, $n \geq 1$, centred Gaussian random variables with variance $\langle\tilde{a}_{i,n}^2\rangle$;

(iv) the LDA-reconstructed database (R-LDA),

$$\tau_-^{R\text{-}LDA}(y, z, t_i) = \frac{N_i}{A}\sum_{n=1}^{N_T} \theta_{n,i}\varphi_n(y, z), \tag{6.5}$$

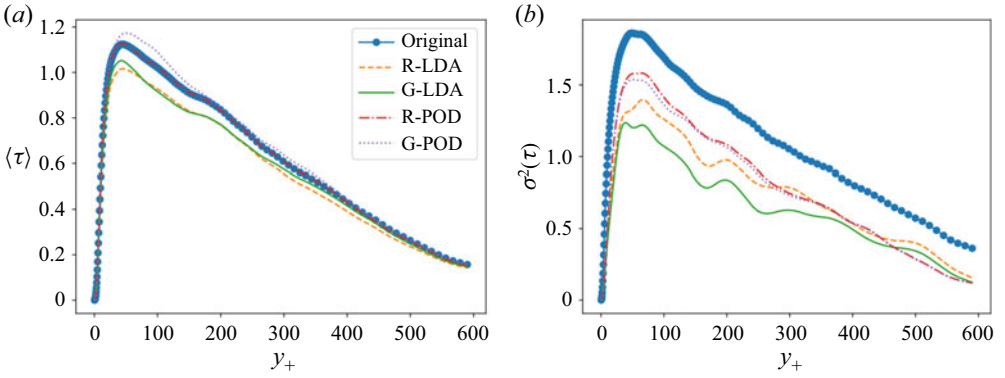


Figure 21. Statistics of the different databases averaged over the spanwise direction and the number of snapshots: (*a*) mean value; (*b*) standard deviation.

where N_i is the total intensity measured in the ith snapshot, $\theta_{n,i}$ is the distribution of motif n on the ith snapshot and $\varphi_n(y, z)$ is the identified distribution of the cell at (y, z) on motif n.

(v) the LDA-generated database (G-LDA),

$$\tau_-^{G\text{-}LDA}(y, z, \tilde{t}_i) = \frac{\tilde{N}_i}{A} \sum_{n=1}^{N_T} \tilde{\theta}_{n,i} \varphi_n(y, z), \tag{6.6}$$

where $\tilde{N}_i$ is the total intensity, which is sampled from a Poisson process, $\varphi_n(y, z)$ is the identified distribution of the cell at (y, z) on motif n and $\tilde{\theta}_{n,i}$ is sampled for each n from the empirical distribution $\theta_{n,i}$ over the snapshots.

Figure 21 shows the statistics of the different databases as a function of the wall distance. Averages are taken over all snapshots and in the streamwise direction. The mean value of the Reynolds stresses is correctly recovered by all methods. The second-order statistics are slightly better recovered by the POD-reconstructed and POD-generated snapshot sets, but both LDA approaches also capture a significant portion of the variance. The POD databases capture 75 % of the total variance, while the reconstructed and generated LDA databases capture 68 % and 60 % of the variance, respectively. Again, the superiority of POD over LDA is expected since it is optimal in the L2 sense, however, it should be noted that the gain is relatively small. Figure 22 shows the vertical spatial autocorrelation of τ_- defined as $R(y, y') = \langle \tau_-(x, y, z, t)\tau_(x, y', z, t)\rangle$ (where $\langle\cdot\rangle$ represents an average taken in time and in the spanwise position). We can see that the generated LDA autocorrelation is very similar to its reconstructed POD counterpart, which shows that the LDA synthetic fields capture as much as the spatial structure as the POD reconstructed ones. We note that the autocorrelation at large separations is well reproduced by all datasets.

Figure 23 shows histograms of the fields at different heights. We note that unlike the LDA approach, which is a non-negative decomposition (since it is based on probabilities), some negative values are observed for the POD approach, even though the original field values considered are always positive. We can see that at different wall distances the POD-reconstructed database reproduces well the distribution of the original database, but the POD-generated database does not. This failure is due to the fact that although POD amplitudes are uncorrelated by construction, they are not independent. We note that the same failure was observed when sampling the POD coefficients from their data-observed

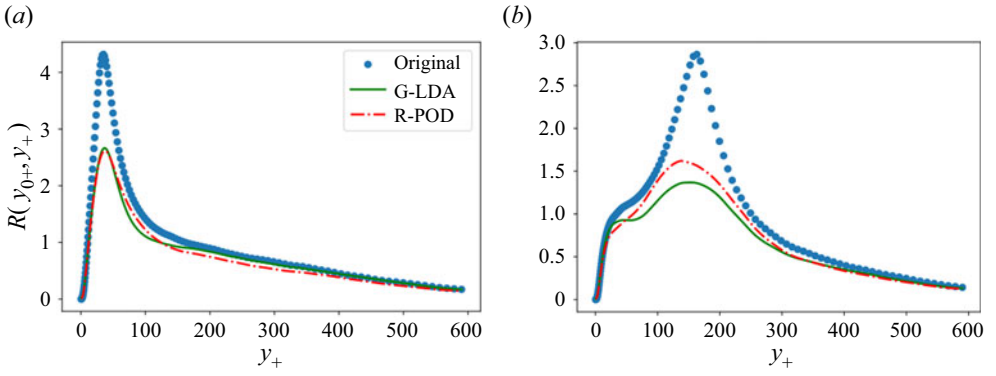


Figure 22. Spatial autocorrelation of the Reynolds stress (limited to Q_- events) in the vertical direction at different heights. The average is taken over snapshots and in the spanwise direction. (*a*) $y_{0+} = 33$, (*b*) $y_{0+} = 163$.

distributions $a_{i,n}$ instead of Gaussian processes. In contrast, both reconstructed and generated LDA methods yield very similar distributions, which reproduce the main features of the original Reynolds stress values, such as the intermittency (sharp peak at zero) and the asymptotic decay for positive values. The LDA is therefore able to reproduce strongly non-Gaussian statistics, which is not possible with its POD equivalent.

7. Conclusion

This paper presents exploratory work about the application of LDA to the identification of coherent structures in turbulent flows. In the probabilistic framework of LDA, latent factors or motifs are inferred from a collection of snapshots. Each snapshot is characterized by a motif distribution, and each motif itself is distributed over space. Implementation was carried out for a scalar field representing Reynolds stress Q_- events. Evidence of self-similarity was found in the motifs: the spanwise and vertical dimensions of the motifs increase linearly with the wall distance in the logarithmic region, and the number of structures evolves inversely with the wall distance. This is in agreement with the eddy attached model hypotheses. The characteristics of the motifs were established to be robust with respect to the LDA parameters.

Latent Dirichlet allocation yields a sparse, efficient reconstruction of the snapshots that compares reasonably well with POD representation. Adding in the fact that the motifs have a local spatial support, even when statistics are homogeneous, could make the LDA representation of interest for estimation and control purposes. Further, a strong benefit of LDA is its inherent generative property, which makes it possible to generate a set of synthetic snapshots which is statistically similar to the original one. In particular, strongly non-Gaussian features are captured and can be reproduced by the LDA representation.

The first results obtained with the LDA method open up exciting prospects for data analysis and modelling of turbulent flows. The present study was carried out at a moderate Reynolds number for domains which are relatively small compared with the typical extent of coherent structures. The observations will therefore need to be confirmed in larger domains at higher Reynolds numbers in future work.

In addition, while the investigation was limited to a positive scalar field in the present implementation, it would be useful to extend the capabilities of LDA to fully real, as well as multidimensional fields. One possibility to handle real fields would be to code the field

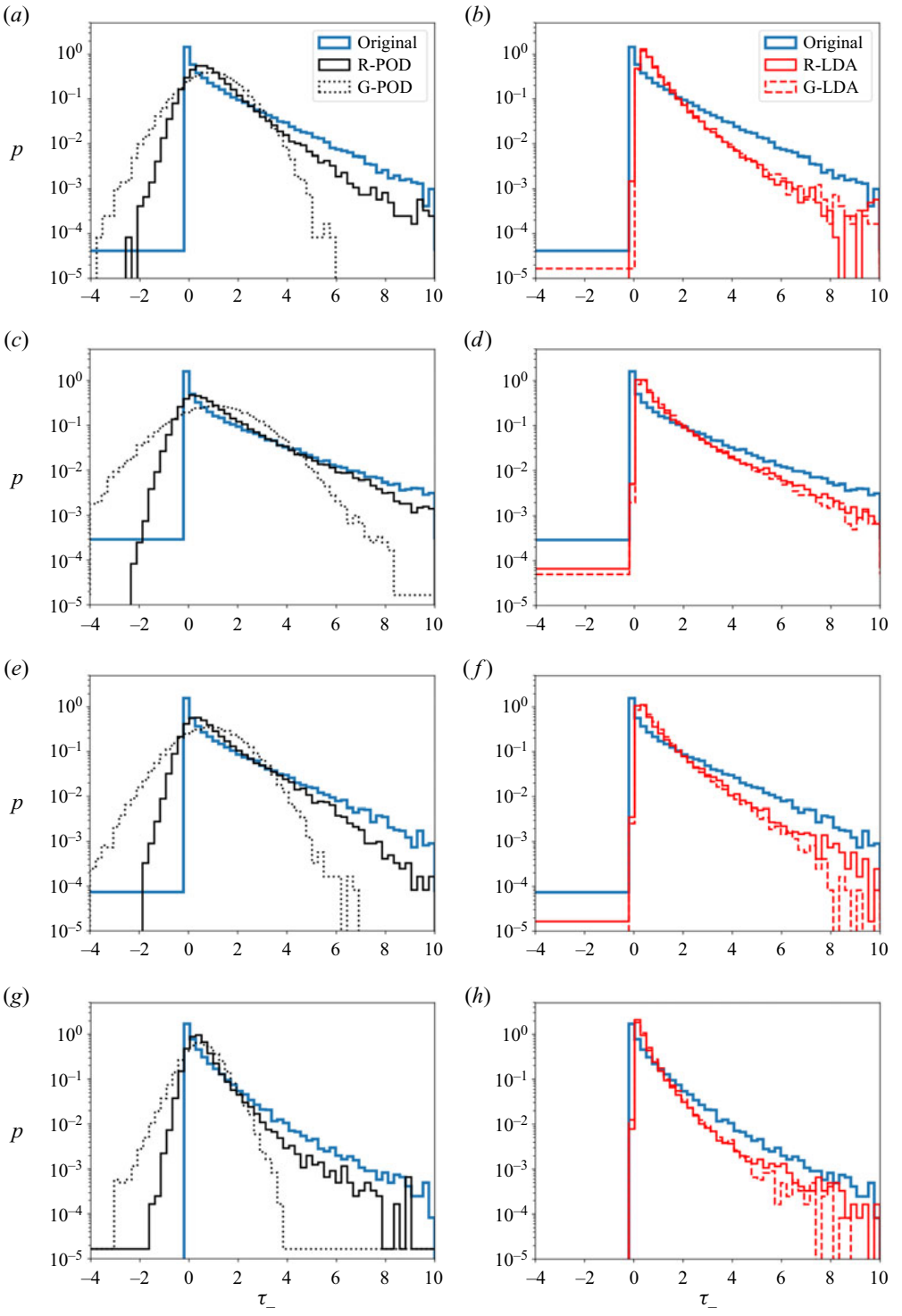


Figure 23. Histograms of the Reynolds stress (limited to Q_- events) corresponding to the different databases at different heights. (a,b) $y_+ = 19$, (c,d) $y_+ = 61$, (e,f) $y_+ = 157$, (g,h) $y_+ = 343$.

sign into the cell index, which would be equivalent to considering two different, positive scalar fields on two identical grids.

Moreover, the LDA technique appears well suited to describe intermittent phenomena, since it decomposes the domain of investigation into a set of localized regions (motifs), each of which is characterized by a specific, snapshot-dependent intensity. It would, therefore, be interesting to apply it to strongly inhomogeneous flow regions such as the turbulent/non-turbulent interface (Philip *et al.* 2014) or other types of intermittency (Johnson & Meneveau 2017).

Acknowledgements. The authors are grateful to the anonymous Referees for their helpful comments on the first version of the manuscript.

Funding. This work was supported by the Center of Data Science from the Paris-Saclay University. Computations were carried out at IDRIS-GENCI (project 02262).

Declaration of interests. The authors report no conflict of interest.

Author ORCIDs.
Bérengère Podvin https://orcid.org/0000-0001-7003-719X.

Appendix A. LDA as a factorization method

To further shed light on the interpretation of LDA, we now adopt a different viewpoint and briefly explore the connections between the decomposition methods discussed above in the framework of MF. Specifically, we now explain how model decomposition methods, such as POD, k-means and LDA, can be interpreted in terms of MF.

A.1. *MF*

Letting $F \in \mathbb{R}^{N_x \times N_s}$ be a data matrix to be approximated, MF consists in the following decomposition:

$$F = XY, \tag{A1}$$

with $X \in \mathbb{R}^{N_x \times N_T}$ and $Y \in \mathbb{R}^{N_T \times N_s}$ two real-valued matrices. Compression is achieved whenever $N_T < \min(N_x, N_s)$, which is considered hereafter. Matrix factorization can be formulated as an optimization problem,

$$(X, Y) \in \underset{\widetilde{X} \in \mathcal{S}_X, \widetilde{Y} \in \mathcal{S}_Y}{\arg\min} \left\| F - \widetilde{X}\widetilde{Y} \right\|^2 + \mathcal{R}\left(\widetilde{X}, \widetilde{Y}\right), \tag{A2}$$

with $\|\cdot\|$ a given norm, $\mathcal{S}_X$ and $\mathcal{S}_Y$ admissibility sets for X and Y, respectively, and $\mathcal{R}$ a regularization term.

A.2. *POD–MF equivalence*

Let the SVD of the $N_x \times N_s$ real-valued data matrix F be

$$F = \Psi \Sigma B^{\mathrm{T}}, \tag{A3}$$

with Ψ and B two orthonormal matrices and Σ being diagonal. The Eckart–Young theorem makes precise in which sense this decomposition is optimal (Eckart &

Young 1936). In particular, it follows that

$$\Psi_{N_T}, \left(\Sigma B^{\mathrm{T}}\right)_{N_T} \in \underset{\widetilde{\Psi}^{\mathrm{T}}\widetilde{\Psi}=I_{N_T}}{\arg\min} \left\| F - \widetilde{\Psi}\left(\widetilde{\Sigma B^{\mathrm{T}}}\right)\right\|_F, \quad \forall N_T \leq \min\left(N_x, N_s\right), \tag{A4}$$

where $(\Sigma B^{\mathrm{T}})_{N_T} = \Sigma_{N_T} B_{N_T}^{\mathrm{T}}$ and with Ψ_{N_T} and B_{N_T} the restriction of Ψ and B to their columns associated with the dominant N_T singular values $\mathrm{diag}(\Sigma_{N_T})$.

From (A3), it follows that

$$FF^{\mathrm{T}}\Psi = \Psi_{N_T}\Sigma_{N_T}B_{N_T}^{\mathrm{T}}B_{N_T}\Sigma_{N_T}^{\mathrm{T}}\Psi_{N_T}^{\mathrm{T}}\Psi_{N_T} = \Psi_{N_T}\Sigma_{N_T}^2 = C_{N_T}\Psi_{N_T}. \tag{A5}$$

Referring to (2.1) and (2.2), the diagonal matrix $\Sigma_{N_T}^2$ and Ψ_{N_T} then directly identify with the N_T dominant eigenvalues Λ and POD modes Φ, respectively. Denoting the Moore–Penrose pseudo-inverse with a $+$ superscript, the POD projection coefficients are

$$A = \Phi^{+}F = \Phi^{\mathrm{T}}F = \Psi_{N_T}^{\mathrm{T}}F = \Sigma_{N_T}B_{N_T}^{\mathrm{T}}, \tag{A6}$$

so that the POD decomposition is finally seen to satisfy the following MF problem:

$$\Phi, A \in \underset{\Phi^{\mathrm{T}}\Phi=I_{N_T}}{\arg\min} \left\|F - \Phi A\right\|_F, \tag{A7}$$

of the form of (A2) with $\mathcal{R} \equiv 0$ and $\mathcal{S}_X$ such that $X^{\mathrm{T}}X = I_{N_T}$.

A.3. *k-means–MF equivalence*

Clustering is an unsupervised learning technique aiming at identifying groups (clusters) in the data such that data points in the same group have similar features, while data points in different groups have highly dissimilar features.

One of the simplest and popular clustering methods is k-means (MacQueen 1967; Lloyd 1982). The algorithm tries to iteratively partition the dataset into N_T predefined distinct non-overlapping clusters $\{\mathcal{C}_n\}_n$. In its standard deterministic version, each data point belongs to only one cluster. The key idea consists in assigning each data point to the closest centroid (arithmetic mean of all the data points that belong to that cluster). The distance is defined in terms of some chosen norm $\|\cdot\|$. Setting the number of clusters N_T, the algorithm starts with an initial guess for the N_T centroids $\{\boldsymbol{c}_n\}_n$, by randomly selecting N_T data points from the data set without replacement. It then iterates between the data assignment step, assigning each data point f_i to the closest cluster $\mathcal{C}_{n_i^\star}$ and the centroid update step, which computes the centroid of each cluster,

$$n_i^\star \leftarrow \underset{1\leq\widetilde{n}\leq N_T}{\arg\max} \left\|\boldsymbol{c}_{\widetilde{n}} - f_i\right\|^2, \quad \forall\, 1 \leq i \leq N_s, \tag{A8}$$

$$\boldsymbol{c}_n \leftarrow \frac{1}{\mathrm{card}\left[\mathcal{C}_n\right]} \sum_{f_i\in\mathcal{C}_n} f_i, \quad \forall\, 1 \leq n \leq N_T. \tag{A9}$$

Furthermore, k-means is guaranteed to converge to a local optimum but not necessarily to a global optimum. Therefore, we choose to run the algorithm with different initializations of centroids and retain the solution that yielded the lowest loss $\mathcal{L}$,

$$\mathcal{L} = \sum_{n=1}^{N_T} \sum_{f_i\in\mathcal{C}_n} \left\|f_i - \boldsymbol{c}_n\right\|^2. \tag{A10}$$

Solving a clustering problem in the L^2 sense means finding a set of $\{\mathcal{C}_n\}_{n=1}^{N_T}$ disjoint clusters ($\mathcal{C}_n \bigcap \mathcal{C}_{n'} = \{\emptyset\}$, $n \neq n'$), that minimizes the following cost function:

$$\mathcal{L} = \sum_{n=1}^{N_T} \sum_{f_i \in \mathcal{C}_n} \|f_i - \boldsymbol{c}_n\|_2^2 = \sum_{i=1}^{N_s} \|f_i\|_2^2 - \sum_{n=1}^{N_T} \sum_{f_i, f_{i'} \in \mathcal{C}_n} n_n^{-1} f_i^{\mathrm{T}} f_{i'}, \tag{A11}$$

where $\{\boldsymbol{c}_n\}_{n=1}^{N_T}$ are the cluster centroids, $\boldsymbol{c}_n := \sum_{f_i \in \mathcal{C}_n} f_i / n_n$, $n_n := \mathrm{card}\,[\mathcal{C}_n]$.

Let $Y \in [0,1]^{N_s \times N_T}$ be the normalized cluster indicator matrix, $\boldsymbol{y}_n = n_n^{-1/2} \mathbb{1}_{\{f_i \in \mathcal{C}_n\}}$. Disjointedness of clusters implies that columns of Y are orthonormal, $Y^{\mathrm{T}} Y = I_{N_T}$. The clustering problem (A11) may now be reformulated in terms of $Y \geq 0$ as (Ding, He & Simon 2005)

$$\begin{aligned} Y &\in \underset{\widetilde{Y} \geq 0, \widetilde{Y}^{\mathrm{T}} \widetilde{Y} = I_{N_T}}{\arg\min} \ \mathrm{Tr}\left[F^{\mathrm{T}} F\right] - \mathrm{Tr}\left[\widetilde{Y}^{\mathrm{T}} F^{\mathrm{T}} F \widetilde{Y}\right], \\ &\in \underset{\widetilde{Y} \geq 0, \widetilde{Y}^{\mathrm{T}} \widetilde{Y} = I_{N_T}}{\arg\min} \ \left\|F^{\mathrm{T}} F\right\|_F^2 - 2\mathrm{Tr}\left[\widetilde{Y}^{\mathrm{T}} F^{\mathrm{T}} F \widetilde{Y}\right] + \left\|\widetilde{Y}^{\mathrm{T}} \widetilde{Y}\right\|_F^2, \\ &\in \underset{\widetilde{Y} \geq 0, \widetilde{Y}^{\mathrm{T}} \widetilde{Y} = I_{N_T}}{\arg\min} \ \left\|F^{\mathrm{T}} F - \widetilde{Y} \widetilde{Y}^{\mathrm{T}}\right\|_F^2. \end{aligned} \tag{A12}$$

The Euclidean hard-clustering k-means problem hence stems from an orthogonal non-negative matrix factorization form and the clusters are given by $\boldsymbol{c}_n = n_n^{-1/2} F \boldsymbol{y}_n$, $\forall n$.

A.4. *LDA–MF equivalence*

We now focus on LDA and discuss the fact that, similarly to POD and k-means, it can also be interpreted as a MF technique, under certain conditions.

Let us consider the variational LDA flavour, where inferring the LDA parameters from maximizing the posterior distribution p is substituted with an approximated posterior q, easier to sample from. The inference problem then consists in minimizing the approximation error, which is equivalent to maximizing the evidence lower bound (ELBO) $\mathcal{L}$,

$$\mathcal{L} = \mathbb{E}_{\mu_q}\left[p\right] - \mathbb{E}_{\mu_q}\left[q\right]. \tag{A13}$$

Provided suitable approximations in the inference problem are made, and under a symmetric Dirichlet priors hypothesis ($\boldsymbol{\alpha} = \alpha \mathbf{1}$), Faleiros & Lopes (2016) have derived an upper bound for the ELBO associated with variational LDA,

$$\max \mathcal{L} \lessapprox \min \sum_{l}^{N_x} \sum_{i}^{N_s} \left(F_{l,i} \log \frac{F_{l,i}}{(XY)_{l,i}} + \sum_{n}^{N_T} \mathcal{R}(Y_{n,i}, \alpha_n) \right), \tag{A14}$$

where $X \geq 0$ and $Y \geq 0$ are variational parameters to be inferred, normalized as $\sum_l X_{l,n} = \sum_n Y_{n,i} = 1$, and regarded as normalized probability distributions. Here, $\boldsymbol{x}_n$ is related to $\boldsymbol{\beta}$ while $\boldsymbol{y}_i$ is related to the distribution $\boldsymbol{\theta}_i$ of a document f_i. The term $\mathcal{R}(Y_{n,i}, \alpha_n) := (Y_{n,i} - \alpha_n)(\log Y_{n,i} - Y_{n,i}(\log Y_{n,i} - 1))$ corresponds to the prior influence and induces sparsity over the document-topic distribution.

From (A14), it follows that maximizing the ELBO $\mathcal{L}$ under certain approximations takes the form of a non-negative MF problem (known as NMF) of $F \approx XY$ expressed in terms of the Kullback–Leibler divergence,

$$D_I(F\|XY) := \sum_{l,i} \left(F_{l,i} \log \frac{F_{l,i}}{(XY)_{l,i}} - F_{l,i} + (XY)_{l,i} \right), \tag{A15}$$

supplemented with a regularization term.

Details of the derivation are beyond the scope of this paper and one should refer to Faleiros & Lopes (2016) for a more complete discussion.

Appendix B. PPCA/POD

In this section, we give a brief review of PPCA (Tipping & Bishop 1999) which provides a density estimation framework for POD (or PCA/LSA), under hypotheses that are different from those given in § 2.2 for PLSA.

We will assume that the data is zero-centred, without loss of generality. The basic idea of PPCA is to assume a Gaussian probability model for the observed data f'. In that formulation (see § 2.2), the motif–cell matrix $\tilde{\Phi}$ of dimension $N_x \times N_T$ does not have a probabilistic interpretation, but relates each noisy observation to a set of N_T independent normalized Gaussian variables following

$$f' = \tilde{\Phi}\tilde{a} + \epsilon, \tag{B1}$$

where the variables $\tilde{a}$ are defined to be independent and Gaussian with unit variance and ϵ represents noise.

An important assumption to proceed is that the model for the noise should be isotropic,

$$\langle \epsilon\epsilon \rangle = \sigma^2 I, \tag{B2}$$

so that all the dependences between the observations are going to be contained in $\tilde{\Phi}$. One can then show using (B1) that

$$p(f') = \mathcal{N}(0, C), \tag{B3}$$

where $C = \tilde{\Phi}^T\tilde{\Phi} + \sigma^2 I$ is the observation covariance matrix of dimension N_x^2.

The issue is to determine $\tilde{\Phi}$ and $\tilde{\sigma}$, given the observations of f'. Under the assumption of isotropic Gaussian noise, Tipping & Bishop (1999) showed that the maximum likelihood estimators $\hat{\Phi}$ and $\hat{\sigma}^2$ can be obtained from standard POD analysis on the N_s snapshots. They showed that

$$\hat{\Phi} = \Phi(\Lambda_{nmode} - \sigma^2 I_{N_T})^{1/2} R, \tag{B4}$$

where Φ contains the first N_T eigenvectors of the sampled covariance matrix $\tilde{C}$ where $\tilde{C}$ was defined in (2.4) (note that the dimension of $\tilde{C}$ is ${N_s}^2$), Λ_{N_T} is a diagonal matrix containing the N_T first eigenvalues of $\tilde{C}$ and R is an arbitrary rotation matrix.

An estimate for the error variance can then be given by

$$\hat{\sigma}^2 = \frac{1}{N_s - N_T} \sum_{j=N_T+1}^{N_s} \lambda_j, \tag{B5}$$

which represents the variance lost in the project and averaged over the lost dimensions.

References

Alamo, J.C.D., Jimenez, J., Zandonade, P. & Moser, R.D. 2006 Self-similar vortex clusters in the turbulent logarithmic region. *J. Fluid Mech.* **561**, 329–358.

Aubert, A.H., Tavenard, R., Emonet, R., de Lavenne, A., Malinowski, S., Guyet, T., Quiniou, R., Odobez, J.-M., Merot, P. & Gascuel-Odoux, C. 2013 Clustering flood events from water quality time series using latent Dirichlet allocation model. *Water Resour. Res.* **49** (12), 8187–8199.

Baars, W.J., Hutchins, N. & Marusic, I. 2017 Self-similarity of wall-attached turbulence in boundary layers. *J. Fluid Mech.* **823**, R2.

Blei, D., Ng, A. & Jordan, M.I. 2003 Latent Dirichlet allocation. *J. Machine Learning Res.* **3**, 993–1022.

Brunton, S.L., Noack, B.R. & Koumoutsakos, P. 2020 Machine learning for fluid mechanics. *Annu. Rev. Fluid Mech.* **52** (1), 477–508.

Cantwell, B.J. 1981 Organized motion in turbulent flow. *Annu. Rev. Fluid Mech.* **13**, 457–515.

Chandran, D., Monty, J.P. & Marusic, I. 2020 Spectral-scaling-based extension to the attached eddy model of wall turbulence. *Phys. Rev. Fluids* **5**, 104606.

Cheng, C., Li, W., Lozano-Durán, A. & Liu, H. 2020 Uncovering townsend's wall-attached eddies in low-Reynolds-number wall turbulence. *J. Fluid Mech.* **889**, A29.

Dempster, A.P., Laird, N.M. & Rubin, D.B. 1977 Maximum likelihood from incomplete data via the em algorithm. *J. R. Stat. Soc.* B **39** (1), 1–38.

Dennis, J.C.D. 2015 Coherent structures in wall-bounded turbulence. *An. Acad. Bras. Cienc.* **87** (2), 161–193.

Ding, C., He, X. & Simon, H.D. 2005 On the equivalence of nonnegative matrix factorization and spectral clustering. In *SIAM International Conference on Data Mining* (ed. H. Kargupta , J. Srivastava , C. Kamath & A. Goodman). SIAM.

Eckart, C. & Young, G. 1936 The approximation of one matrix by another of lower rank. *Psychometrika* **1** (3), 211–218.

Faleiros, T. & Lopes, A. 2016 On the equivalence between algorithms for non-negative matrix factorization and latent dirichlet allocation. In *ESANN European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning XXIV, Bruges*. Louvain-la-Neuve Ciaco.

Flores, O. & Jimenez, J. 2010 Hierarchy of minimal flow units in the logarithmic layer. *Phys. Fluids* **22**, 071704.

Griffiths, T.L. & Steyvers, M. 2002 A probabilistic approach to semantic representation. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (ed. W.D. Gray & C.D. Schunn). Routledge, Taylor & Francis.

Griffiths, T.L. & Steyvers, M. 2004 Finding scientific topics. *Proc. Natl Acad. Sci.* **101** (suppl 1), 5228–5235.

Hellström, L.H.O., Marusic, I. & Smits, A.J. 2016 Self-similarity of the large-scale motions in turbulent pipe flow. *J. Fluid Mech.* **792**, R1.

Hofmann, M.D., Blei, D.M., Wang, C. & Paisley, J. 2013 Stochastic variational inference. *J. Machine Learning Res.* **14**, 1303–1347.

Hofmann, T. 1999 Probabilistic latent semantic analysis. In *Proceedings of Uncertainty in Artificial Intelligence* (ed. K.B. Laskey & H. Prade). UAI99. Morgan Kaufmann.

Holmes, P., Lumley, J.L. & Berkooz, G. 1996 *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*. Cambridge University Press.

Hu, R., Yang, X.I.A. & Zheng, X. 2020 Wall-attached and wall-detached eddies in wall-bounded turbulent flows. *J. Fluid Mech.* **885**, A30.

Hwang, J. & Sung, H.J. 2018 Wall-attached structures of velocity fluctuations in a turbulent boundary layer. *J. Fluid Mech.* **856**, 958–983.

Hwang, J. & Sung, H.J. 2019 Wall-attached clusters for the logarithmic velocity law in turbulent pipe flow. *Phys. Fluids* **31** (5), 055109.

Jimenez, J. 2013 Near-wall turbulence. *Phys. Fluids* **25** (1), 101302.

Jimenez, J. 2018 Coherent structures in wall-bounded turbulence. *J. Fluid Mech.* **842**, P1.

Johnson, P.L. & Meneveau, C. 2017 Turbulence intermittency in a multiple-time-scale Navier–Stokes-based reduced model. *Phys. Rev. Fluids* **2**, 072601.

Kim, H.T., Kline, S.J. & Reynolds, W.C. 1971 The production of turbulence near a smooth wall in a turbulent boundary layer. *J. Fluid Mech.* **50** (1), 133–160.

Kline, S.J., Reynolds, W.C., Schraub, F.A. & Runstadler, P.W. 1967 The structure of turbulent boundary layers. *J. Fluid Mech.* **30** (4), 741–773.

Lloyd, S.P. 1982 Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **28** (2), 129–137.

Lozano-Duran, A., Flores, O. & Jimenez, J. 2012 The three-dimensional structure of momentum transfer in turbulent channels. *J. Fluid Mech.* **524**, 1–31.

LUMLEY, J.L. 1967 The structure of inhomogeneous turbulent flows. In *Atmospheric Turbulence and Radio Wave Propagation* (ed. A.M. Iaglom & V.I. Tatarski), pp. 221–227. Nauka.

MACQUEEN, J. 1967 Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics* (ed. L.M. Le Cam & J. Neyman), pp. 281–297. University of California Press.

MARUSIC, I. & MONTY, J.P. 2019 Attached eddy model of wall turbulence. *Annu. Rev. Fluid Mech.* **51** (1), 49–74.

MOSER, R., KIM, J. & MANSOUR, N.N. 1999 Direct numerical simulation of turbulent channel flow up to $Re_\tau = 590$. *Phys. Fluids* **11** (4), 943.

MURALIDHAR, S., PODVIN, B., MATHELIN, L. & FRAIGNEAU, Y. 2019 Spatio-temporal proper orthogonal decomposition of turbulent channel flow. *J. Fluid Mech.* **864**, 614–639.

PEDREGOSA, F., *et al.*2011 Scikit-learn: machine learning in Python. *J. Machine Learning Res.* **12**, 2825–2830.

PERRY, A.E. & CHONG, M.S. 1982 On the mechanism of wall turbulence. *J. Fluid Mech.* **119**, 173–217.

PERRY, A.E. & MARUSIC, I. 1995 A wall-wake model for the turbulent structure of boundary layers. Part 1. Extension of the attached eddy hypothesis. *J. Fluid Mech.* **298**, 361–388.

PHILIP, J., MENEVEAU, C., DE SILVA, C.M. & MARUSIC, I. 2014 Multiscale analysis of fluxes at the turbulent/non-turbulent interface in high Reynolds number boundary layers. *Phys. Fluids* **26** (1), 015105.

PODVIN, B. & FRAIGNEAU, Y. 2017 A few thoughts on proper orthogonal decomposition in turbulence. *Phys. Fluids* **29**, 531.

PODVIN, B., FRAIGNEAU, Y., JOUANGUY, J. & LAVAL, J.P. 2010 On self-similarity in the inner wall layer of a turbulent channel flow. *Trans. ASME J. Fluids Engng* **132** (4), 41202.

ROBINSON, S.K. 1991 Coherent motions in the turbulent boundary layer. *Annu. Rev. Fluid Mech.* **23**, 601–639.

SHARMA, A.S. & MCKEON, B.J. 2013 On coherent structure in wall turbulence. *J. Fluid Mech.* **728**, 196–238.

SIROVICH, L. 1987 Turbulence and the dynamics of coherent structures part i: coherent structures. *Q. Appl. Maths* **45** (3), 561–571.

STANISLAS, M., PERRET, L. & FOUCAUT, J.M. 2008 Vortical structures in the turbulent boundary layer: a possible route to a universal representation. *J. Fluid Mech.* **602**, 327–382.

TIPPING, M.E. & BISHOP, C.M. 1999 Probabilistic principal component analysis. *J. R. Stat. Soc.* B **61** (3), 611–622.

TOWNSEND, A.A. 1947 Measurements in the turbulent wake of a cylinder. *Proc. R. Soc. Lond.* A **190**, 551–561.

TOWNSEND, A.A. 1961 Equilibrium layers and wall turbulence. *J. Fluid Mech.* **11** (1), 97–120.

WALLACE, J.M., ECKELMANN, H. & BRODKEY, R.S. 1972 The wall region in turbulent shear flow. *J. Fluid Mech.* **54** (1), 39–48.

WILLMARTH, W.W. & LU, S.S. 1972 Structure of the Reynolds stress near the wall. *J. Fluid Mech.* **55** (1), 65–92.

WOODCOCK, J.D. & MARUSIC, I. 2015 The statistical behavior of attached eddies. *Phys. Fluids* **27**, 015104.

ZHOU, J., ADRIAN, R.J., BALACHANDAR, S. & KENDALL, T.M. 1999 Mechanisms for generating coherent packets of hairpin vortices in channel flow. *J. Fluid Mech.* **387**, 173–217.