A probability theoretic approach to drifting data in continuous time domains

Fabian Hinder, André Artelt, and Barbara Hammer

Cognitive Interaction Technology (CITEC)
Bielefeld University
Inspiration 1, D-33619 Bielefeld, Germany
{fhinder,aartelt,bhammer}@techfak.uni-bielefeld.de

December 4, 2019

Abstract

The notion of drift refers to the phenomenon that the distribution, which is underlying the observed data, changes over time. Albeit many attempts were made to deal with drift, formal notions of drift are application-dependent and formulated in various degrees of abstraction and mathematical coherence. In this contribution, we provide a probability theoretical framework, that allows a formalization of drift in continuous time, which subsumes popular notions of drift. In particular, it sheds some light on common practice such as change-point detection or machine learning methodologies in the presence of drift. It gives rise to a new characterization of drift in terms of stochastic dependency between data and time. This particularly intuitive formalization enables us to design a new, efficient drift detection method. Further, it induces a technology, to decompose observed data into a drifting and a non-drifting part.

Keywords: Online learning, learning theory, stochastic processes, learning with drift, continuous time models, drift decomposition

1 INTRODUCTION

One fundamental assumption in classical machine learning is the fact that observed data are i.i.d. according to some unknown underlying probability measure P_X, i.e. the data generating process is stationary. Yet, this assumption is often violated as soon as machine learning faces real world problems: models are subject to seasonal changes, changed demands of individual costumers, ageing of sensors, etc. In such settings, life-long model adaptation rather than classical batch learning is required for optimum performance. Since drift, i.e. the fact that data is no longer identically distributed, is a major issue in many real-world applications of machine learning, many attempts were made to deal with this setting (Ditzler et al., 2015).

Depending on the domain of data and application, the presence of drift is modelled in different ways. As an example, covariate shift refers to the situation of training and test set having different marginal distributions (Gretton et al., 2009). Learning for data streams extends this setting to an unlimited (but usually countable) stream of observed data, mostly in supervised learning scenarios (Gama et al., 2014). Here one distinguishes virtual and real drift, i.e. non-stationarity of the marginal distribution only or also the posterior. Learning technologies for such situations often rely on windowing techniques, and adapt the model based on the characteristics of the data in an observed time window. Active methods explicitly detect drift, usually referring to drift of the classification error, and trigger model adaptation this way, while passive methods continuously adjust the model (Ditzler et al., 2015).

Data streams also occur naturally whenever times series are dealt with, such as time series prediction. Unlike streaming data as considered by Ditzler et al. (2015) or Gama et al. (2014), time series modeling relies on the assumption of a direct functional relation of subsequent observations. One distinguishes stationary and non-stationary time series, and one particularly interesting challenge is change point detection, i.e. time points where abrupt variations are observed (Aminikhanghahi and Cook, 2017; Alippi et al., 2017)

Interestingly, the overwhelming majority of such drift learning approaches deals with discrete time processes rather than continuous time (Roveri, 2019). Further, the majority refers to supervised learning scenarios with an emphasis on minimization of a cost function such as the interleaved train-test error. Only first approaches consider the particularly relevant question how to substantiate such models by methods for understanding drift (Webb et al., 2017).

The purpose of this contribution is to provide a proper probabilistic definition of drift for streaming data in continuous time, which subsumes common definitions of drift in the literature. Unlike Goldenberg and Webb (2019), we are not interested how to identify and measure different types of drift (such as real drift, virtual drift, reoccurring drift, etc.); rather, we are interested in a unifying probabilistic model of drift in continuous time processes, which also justifies common practice to deal with drift, such as identifying change points or learning from time windows.

Now, we will introduce a measure-theoretic setting to define drift in continuous time first, and we introduce different notions of drift from the literature and show their equivalence. Then, we establish a new characterization by relating drift to an independence criterion of time and data, giving rise to particularly efficient drift detection models as well as an elegant way to disentangle drifting and non-drifting parts in observed data. We demonstrate these methods in experiments.

2 A THEORY OF DRIFT

In the following we will define the notion of a drift process. Afterwards we will give several definitions of drift, that have been proposed in different fields, and investigate their relationships. In particular we introduce a new definition of drift for continuous time processes in Section 2.6. Due to space restrictions, all proofs (and some explanations of well known definitions) are contained in the auxiliary material (identifiable by numeration starting with an "A").

2.1 Definition of a drift process

In the usual, time invariant setup of machine learning one considers a generative process P_X, i.e. a probability measure, on a measurable space $(\mathfrak{X}, \mathcal{A})$. In this context one views the realizations of a $\mathfrak{X}$-valued, P_X distributed random variable X as samples. Depending on the objective, learning algorithms try to infer the data distribution based on these samples or, in the supervised setting, a posterior distribution. We will not distinguish these settings and only consider distributions in general, this way subsuming the notion of both, real drift and virtual drift.

Many processes in real-world applications are not time independent, so it is reasonable to incorporate time into our considerations. One prominent way to do so is to consider an index set $\mathfrak{T}$, representing time, and a collection of probability measures p_t on $\mathfrak{X}$ indexed over $\mathfrak{T}$ (Gama et al., 2014).

In the following we investigate the relationship of those p_t, with drift referring to a property of the relationship of several p_t at different time points t. A first, and mathematically equivalent, step to do so is by considering $p : \mathfrak{T} \to \mathbf{P}(\mathfrak{X})$, $t \mapsto p_t$ as a map rather than a conglomerate, here $\mathbf{P}(\mathfrak{X})$ denotes the set of all probability measures on $\mathfrak{X}$. We will sometimes refer to this as a *non-probabilistic drift process.*

For continuous time, we need more structure; hence we view p_t in a measure theoretic setup, which yields:

Definition 1. Let $(\mathfrak{T}, \mathcal{B})$ and $(\mathfrak{X}, \mathcal{A})$ be two measurable spaces. A *drift process* (p_t, P_T) is a Markov kernel[1] p_t from $\mathfrak{T}$ to $\mathfrak{X}$ and a probability measure P_T on $\mathfrak{T}$.

When $(\mathfrak{X}, \mathcal{A})$ and $(\mathfrak{T}, \mathcal{B})$ or P_T are clear, we sometimes just write (P_T, p_t) resp. p_t for simplicity. Notice that this is a very minor restriction regarding p_t as compared to a non-probabilistic drift process, since we basically only state that we want $t \mapsto p_t(D)$ to be a measurable map for all $D \in \mathcal{A}$ [2].

Note that this notion of drift processes is actually extremely natural:

Remark 1. By Fubini's theorem every drift process (p_t, P_T) induces a probability measure $p_t \otimes P_T$ [3] on $\mathfrak{X} \times \mathfrak{T}$.
Conversely under some mild assumptions (e.g. $\mathfrak{T}$ and $\mathfrak{X}$ are polish spaces (Parthasarathy, 1967)), every probability measure P on $\mathfrak{X} \times \mathfrak{T}$ gives rise to a drift process, i.e. we may find a Markov kernel p_t with $p_t \otimes P_T = P$, where P_T is the marginalization of P onto $\mathfrak{T}$.
In particular if P_T has no null sets, i.e. $P_T(\{t\}) > 0$ for all $t \in \mathfrak{T}$, then we have $p_t = P(\cdot \mid \mathfrak{X} \times \{t\})$ the conditional expectation given $t \in \mathfrak{T}$.

We will now define drift: A very common notion specifies drift as the fact that distributions vary over time (Gama et al., 2014), i.e. there exist $t, s \in \mathfrak{T}$ such that $p_t \neq p_s$. Conversely a process has no drift iff $p_t = p_s$ for all $t, s \in \mathfrak{T}$. In the following definition, this notion is adapted to the measure theoretic setup.

Definition 2. Let (p_t, P_T) be a drift process. We say that p_t has *no drift* or does not drift if $p_t = p_s$ holds P_T-a.s., i.e. $(P_T \times P_T)(\{(s,t) \in \mathfrak{T} \times \mathfrak{T} \mid p_t = p_s\}) = 1$[4]. We say that p_t has *drift* or is drifting if it is not the case that it does not drift.

Here we allow differences of distributions in null sets, i.e. we allow that $p_t \neq p_s$ if we do not expect to ever observe a sample at t resp. s, which makes this difference irrelevant for applications. In particular if there exists a measure P_T on $\mathfrak{T}$ that has no null set then both notions coincide (see Lemma A.1).

Though our definition is already weaker than the one given in Gama et al. (2014), it is still too strict for applications. This is mainly caused by the fact that the decomposition described in Remark 1 is not unique. Therefore we need a notion of drift where we no longer distinguish drift processes that do not differ in this sense. This leads to the following definition:

Definition 3. Let (p_t, P_T) be a drift process. We say that p_t has *no proper drift* iff there exists a drift process (p'_t, P'_T), such that $p_t \otimes P_T = p'_t \otimes P'_T$, that does not drift. We say that p_t has *proper drift* iff it is not the case that it has no proper drift.

Remark 2. Proper drift implies drift but the converse does not hold in general. However under some assumptions, e.g. $\mathfrak{T}$ is (at most) countable or $\mathfrak{X} = \mathbb{R}^d$ (or more general if $\mathcal{A}$ is generated by a countable set, stable under finite intersections[5]; see Lemma A.2), drift and proper drift are equivalent.

2.2 Road map

The results we are going to show, i.e. the fact that this notion subsumes several popular definitions from the literature, are summarized in Figure 1.

(1) and (2) do not hold in general. (1) holds for example if X_t has $\mathbb{P}$-a.s. continuous paths.[6] (2) holds if $\mathcal{A}$ has a intersection stable, countable generating set. If $\mathfrak{T}$ is (at most) countable, then (1) and (2) hold. In this case every probability measure P on $\mathfrak{X} \times \mathfrak{T}$ gives rise to a drift process.

2.3 Drift as change of distribution

In this subsection, we discuss that Definition 2 can be simplified to the fact that the probability distribution is constant, i.e. p_t does not depend on time (up to a null set). This is the standard setting of classical (drift free) machine learning.

[1] See Definition A.1
[2] See Remark A.1 for more details
[3] See Remark A.4 for definition and more details
[4] See Definition A.4 for details
[5] See Definition A.2 for details
[6] See Definition A.5

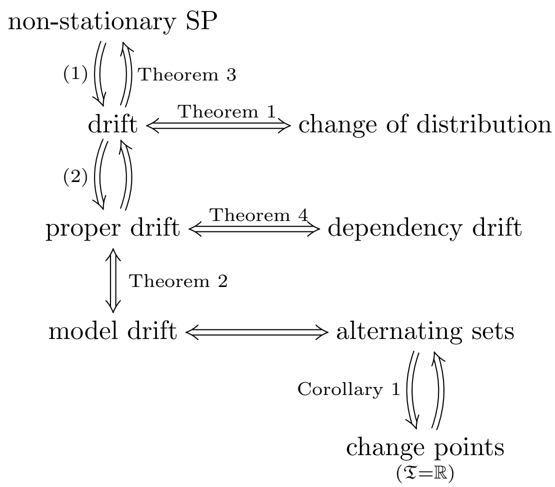


Figure 1: Equivalent Notions Of Drift

Definition 4. Let (p_t, P_T) be a drift process. We say that p_t is *constant* if there exists a probability measure $P_X \in \mathbf{P}(\mathfrak{X})$ such that $p_t = P_X$ for P_T-a.s. all $t \in \mathfrak{T}$. We say that p_t has a *change of distribution* or is changing iff it is not the case that p_t is constant.

It is clear that P_X is uniquely determined by this property. Furthermore we can show that P_X given p_t is characterized by the expectation with respect to P_T (Corollary A.1). Indeed, we can even find a $t_0 \in \mathfrak{T}$ such that $P_X = p_{t_0}$, i.e. P_X actually appears at some (actually nearly every) point in time (Lemma A.3).

This enables us to characterize the relation between drifting processes and change of distribution:[7]

Theorem 1. *Let (p_t, P_T) be a drift process. Then p_t is constant if and only if p_t has no drift.*

Though it does not hold that drift processes with no proper drift are constant, it can be proven, that a drift process has no proper drift if an only if $p_t \otimes P_T = P_X \times P_T$[8], for some probability measure P_X (Lemma A.4).

[7] Notice that we have to take care of the null sets here.

[8] See Definition A.4 for details

2.4 Drift as change of model

We will now consider drift in the context of machine learning models; machine learning models in the context of drift often learn a constant model over a time window. It is common practice to detect drift by a change of such model, e.g. a changed error. Here, we are not interested in specific models, rather we consider $\mathfrak{T}$-invariant models, which we will use as prototypical optimum machine learning model:

Definition 5. Let (p_t, P_T) be a drift process. For a P_T-non-null set $A \in \mathcal{B}$ we define the *$\mathfrak{T}$-invariant model* of p_t over A as the marginalization of $(p_t \otimes P_T)(\cdot \mid \mathfrak{X} \times A)$ onto $\mathfrak{X}$ or equivalent

$$p_A := \frac{1}{P_T(A)} \int_A p_t \mathrm{d}P_T.$$

p_A is the optimal, time invariant model in the sense that every static probabilistic model that is capable of universal approximations, trained with data observed during A only, converges to p_A.

Furthermore notice that those models have some Bayesian-like properties: for disjoint non-null sets $A, B \in \mathcal{B}$ it holds

$$p_{A\cup B} = \frac{1}{P_T(A) + P_T(B)} (P_T(A) p_A + P_T(B) p_B).$$

Now we can characterize drift in terms of models derived from the drift process not being constant:

Definition 6. We say that a pair of P_T-non-null sets $A, B \in \mathcal{B}$ are *alternating sets* iff $p_A \neq p_B$. If alternating sets exist, then we say that p_t has *model drift.*

Model drift characterizes the fact that an optimal model for observed streaming data, changes over time. Having in mind that a practical model approximates the behavior of an optimal $\mathfrak{T}$-invariant model, we see that model drift captures common practice: e.g. many drift detection methods refer to a change in model accuracy (Bifet and Gavaldà, 2007; Gama et al., 2004; Baena-García et al., 2006).

We will now investigate the relation of model drift and proper drift:

Theorem 2. *Let (p_t, P_T) be a drift process and let $\mathcal{B}_0 \subset \mathcal{B}$ a generating set (i.e. $\sigma(\mathcal{B}_0) = \mathcal{B}$ [9]), which is stable under finite intersections. Then the following properties are equivalent:*

1. *p_t has proper drift,*
2. *p_t has model drift,*
3. *there exist alternating sets A, A^C, with $A \in \mathcal{B}_0$.*

Besides the observation that model drift is equivalent to proper drift, Theorem 2 has an interesting consequence regarding the structure of alternating sets: alternating sets take the form of complementary subsets of $\mathfrak{T}$. Provided the index set $\mathfrak{T}$ represents time, i.e. is contained in the real numbers, this implies that model drift is the same as the existence of a change-point:

Corollary 1. *Let (p_t, P_T) be a drift process and suppose that $\mathfrak{T} \in \{[a, b], \mathbb{R}_{\geq 0}, \mathbb{R}\}, \mathcal{B} = \mathfrak{B}(\mathfrak{T})$. If p_t has proper drift, then there exists a change-point, i.e. it exists a $t_0 \in \mathbb{R}$ such that $p_{\{t<t_0\}} \neq p_{\{t \geq t_0\}}$.*

This result can be seen as a justification of change-point detection methods in the field of drift detection.

2.5 Drift as non-stationarity of a stochastic process

In the context of time-series, the notion of stationary processes constitutes a prominent concept (Park, 2018). We discuss its relation to drift. In this section let $\mathfrak{T} \in \{\mathbb{N}, \mathbb{Q}, \mathbb{R}\}$, so that we have a natural shift operation $\cdot + \tau$ with $\tau \in \mathfrak{T}$, representing shift in time.

Definition 7. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A stochastic process $X_t : \Omega \times \mathfrak{T} \to \mathfrak{X}$ is *stationary* if

$$\mathbb{P} \circ (X_{t_1}, ..., X_{t_n})^{-1} = \mathbb{P} \circ (X_{t_1+\tau}, ..., X_{t_n+\tau})^{-1}$$

for all $t_1, ..., t_n \in \mathfrak{T}$, $\tau \in \mathfrak{T}$ and $n \in \mathbb{N}$.

Notice that stationary implies having no drift.

Theorem 3. *Let X_t be a stochastic process. For every sequence $t_1, ..., t_n \in \mathfrak{T}$ we obtain a non-probabilistic drift process by setting $p_\tau^{(t_1,...,t_n)} = \mathbb{P} \circ (X_{t_1+\tau}, ..., X_{t_n+\tau})^{-1}$. If X_t is stationary then $(p_\tau^{(t_1,...,t_n)}, P_T)$ has no drift for all probability measures P_T on $\mathfrak{T}$, $t_1, ..., t_n \in \mathfrak{T}$ and $n \in \mathbb{N}$.*
Furthermore, if P_T has no null sets, then the notion of stationarity of X_t and $(p_\tau^{(t_1,...,t_n)}, P_T)$ having no drift for all $t_1, ..., t_n \in \mathfrak{T}$ and $n \in \mathbb{N}$, are equivalent.

The reason why having no drift does not imply stationarity comes from the fact that the latter is defined point-wise for all $\tau \in \mathfrak{T}$. It is therefore equivalent to the non-probabilistic definition of no drift (Gama et al., 2014), using the same transformation as used in Theorem 3. However additional assumptions could be added to induce such implication, e.g. by assuming that X_t has $\mathbb{P}$-a.s. (Corollary A.4) continuous paths[10].

2.6 Drift as dependency between data and time

In addition to these notions of drift from the literature, we will now discuss drift under a novel aspect, which will be particularly suited to derive efficient algorithms, namely in the context of independence of random variables.

In the classical machine learning setup one considers samples as realizations of (independent) identically distributed $\mathfrak{X}$-valued random variables. In the context of drift, this distribution changes, as discussed above. To put this into the context of dependence of variables, we can equip each sample with a timestamp of its occurrence: instead of $\mathfrak{X}$-valued random variables X, we consider $\mathfrak{X} \times \mathfrak{T}$-valued random variables (X, T). If there is no drift then the distribution of X should not depend on T, i.e. X and T should be independent:

Definition 8. Let (p_t, P_T) be a drift process and let $(X, T) \sim p_t \otimes P_T$ a pair of random variables. We say that p_t has *dependency drift* if X and T are not independent.

It turns out that this is an alternative characterization of proper drift:

[9] See Definition A.3 for details

[10] See Definition A.5

Theorem 4. *Let (p_t, P_T) be a drift process. Then p_t has proper drift if and only if it has dependency drift.*

This result allows us to reduce the problem of drift detection to the problem to test independence of variables. The latter problem is well investigated and highly efficient algorithms exist for independence tests.

3 APPLICATIONS

In Theorem 4 we showed that drift can be described as the dependency between data and time. We will now make use of this by construction two methods: A fast, ADWIN (Bifet and Gavaldà, 2007) based drift detector (SWIDD) and a drift explanation method (DriFDA).

3.1 Single Window Independence Drift Detection (SWIDD)

Drift detection refers to the task to determine whether there is a change in an observed data stream. Most drift detection methods (Bifet and Gavaldà, 2007; Gama et al., 2004; Baena-García et al., 2006; PAGE, 1954; Vorburger and Bernstein, 2006) detect drift using a two window approach; samples are hold in two windows that are assumed to be sampled from the same base distribution, so that drift may be detected using a two-sample test. This may be done directly as in (PAGE, 1954) or after a transformation, i.e. use the prediction error of a model (Bifet and Gavaldà, 2007; Gama et al., 2004; Baena-García et al., 2006).

We will rely on the pipeline as proposed in ADWIN as particularly popular method. ADWIN (Bifet and Gavaldà, 2007) stores the incoming prediction errors in a sliding, size-adaptive window that is successively split into two windows. Those two windows are then tested against each other by checking whether the absolute difference of the mean prediction error exceeds a predefined threshold. If so, a drift is indicated and all samples before that time are discarded.

We use Theorem 4 to extend this idea to detect general drift using a single window only: Instead of splitting our window we assign every sample with a time-stamp and apply a statistical test to determine whether time and data are dependent or not. This leads to Algorithm 1. Notice that this differs from the usual ADWIN only in line 4 where we add the timestamp to x_i of the moment of its arrival, rather than the prediction error and in line 6 where we use an independence test to check for drift (Theorem 4), rather than window splitting. We implement[11]SWIDD (Algorithm 1) based on the Hilbert-Schmidt Independence Criterion (HSIC) (Gretton et al., 2007).

Algorithm 1 SWIDD

1: **procedure** SWIDD: SINGLE WINDOW INDEPENDENCE DRIFT DETECTOR((x_i) data stream, p p-value for statistical test, $n_{\min}$ minimal number of samples in window)
2:     Initialize Window $W \leftarrow []$
3:     **while** Not at end of stream (x_i) **do**
4:         $W \leftarrow W \cup \{(x_i, t_i)\}$ ▷ i.e. add new sample x_i received at time t_i
5:         **repeat** Drop element from the tail of W
6:         **until** $|W| < n_{\min}$ or TEST(W, p) accepts H_0
7:     **end while**
8: **end procedure**

A beneficial property of SWIDD is that we may use a test for general or conditional independence; which allows us to apply it to virtual and real drift alike. Furthermore, since we do not depend on a model, we are not subject to its specific deficiencies (see Figure 2b).

SWIDD is superior to a fixed two-window approach when it comes to continuous, in particular fast and periodic drift. This is caused by the fact that two window approaches assume an identical distribution at least for a single window; a counter example would be $p_t = \mathcal{N}(\sin(t), \sigma)$ and a window size of $2n\pi$, where $\mathcal{N}(\mu, \sigma)$ denotes the normal distribution. Though this problem can be solved by dynamic window selection as used by ADWIN (see Corollary 1), it causes

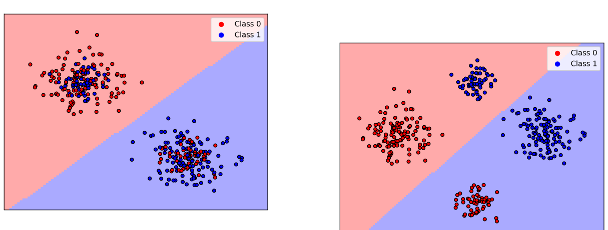
[11]The implementations of our proposed methods are available on GitHub - `https://github.com/FabianHinder/drifting-data-in-continuous-time`

a considerable amount of computation time.

3.1.1 Experiments

To demonstrate the generality of our approach, we created two artificial data sets which highlight typical challenges of existing approaches for drift detection. In addition, we compare SWIDD to the drift detection methods ADWIN (Bifet and Gavaldà, 2007) and DDM (Gama et al., 2004), which depend on the classification accuracy of a supervised model, as well as against the (unsupervised) statistical methods HDDDM (Ditzler and Polikar, 2011) and HDDDM where we replaced the Hellinger-distance and the t-test by the kernel two sample test (Gretton et al., 2006) (referred to as K2ST), on the SEA data set (Street and Kim, 2001) and the rotating hyperplane data set (Hulten et al., 2001) (referred to as RPLANE).

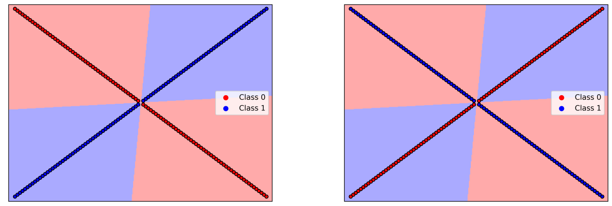
Comparison to supervised drift detection Methods such as ADWIN (Bifet and Gavaldà, 2007), DDM (Gama et al., 2004) and EDDM (Baena-García et al., 2006) use the classification error as an indicator of drift. They assume that drift leads to a change (e.g. decrease) in accuracy. We construct a scenario in which this assumption does not hold: We create a binary classification data set with two clusters. Both clusters are mixtures of samples from both classes, but with different dominance. A linear classifier yields a decision boundary as shown in Figure 2a. Drift is constructed by moving all samples



(a) A linear classifier fitted to the original data set.

(b) Drifted data set. Note that the accuracy does not change.

Figure 2: Fooling error based drift detection methods



(a) A quadratic discriminant model fitted to the original data set.

(b) Drifted data set. Mean, variance and feature-wise marginals did not change.

Figure 3: Fooling simple distributional drift detectors

from one class along the decision boundary in the direction of the upper right corner, whereby we do not cross the decision boundary. The final scenario is shown in Figure 2b. Error-based drift detectors do not detect this drift because the classification error does not change when moving the data points this way, unless the classifier is retrained. It would be possible to obtain a better (in the limit perfect) accuracy; yet, active drift learners would require a drift detection to do so. SWIDD detects this drift since it does not rely on the classification error.

Comparison to unsupervised drift detection Another class of methods for drift detection is based on distributional changes (Kifer et al., 2004; Matteson and James, 2014; Dette and Wied, 2016; Vorburger and Bernstein, 2006; Ditzler and Polikar, 2011; Dasu et al., 2006; Song et al., 2007; Gretton et al., 2006). These methods try to detect drift by detecting changes in the sampling distribution of the data stream. Many of these methods (Kifer et al., 2004; Matteson and James, 2014; Dette and Wied, 2016; Vorburger and Bernstein, 2006; Ditzler and Polikar, 2011) use some kind of windowing - split the data stream (or parts of it) into two windows and compute statistics on these windows. However, relying on two windows can be problematic because we have to select the right length of the window so that quickly occurring abrupt drifts are recognized – usually, it is assumed that the distribution of the samples in a window is fixed. Another problem of some of these methods is that they try to reduce computational complexity by assuming that the drift will show up in the mean, variance or feature-wise marginals (Ditzler and Polikar, 2011; Vorburger and Bernstein, 2006). This is problematic because one can construct drifting data sets where the mean, variance and the feature-wise marginal distribution do not change - such drifts can not be perceived by methods that make these simplifying assumptions. For instance we can construct a data set where the points are arranged like a cross so that each class has its own diagonal - see Figure 3a. If the cross is symmetric and if the samples are placed symmetrically around the center, then we can swap the labels of the two diagonals - see Figure 3b - but the mean, variance and the feature-wise marginal distributions do not change. Therefore, these methods do not recognize the drift. However, our method is able to detect this drift since it does not make any simplifying assumptions about the distributional changes.

Benchmarks We compare SWIDD on common benchmark data sets: the SEA data set (Street and Kim, 2001) and the rotating hyperplane data set (Hulten et al., 2001). We recorded the mean F1-score and mean computation time - over three runs with different random seeds - and compared SWIDD to HDDDM, K2ST, ADWIN and DDM. The results are shown in Table 1. SWIDD is best for SEA and second best for RPLANE, being reasonably fast in both cases.

Table 1: Mean F1-score and mean computation time of drift detectors on two common benchmark data sets.

	Data set		Computation time	
Method	RPLANE	SEA	RPLANE	SEA
SWIDD	0.86	0.63	2.59s	3.61s
HDDDM	0.44	0.44	0.02s	0.03s
K2ST	1.00	0.22	5.74s	19.87s
ADWIN	0.60	0.33	0.12s	0.14s
DDM	0.70	0.13	0.06s	0.06s

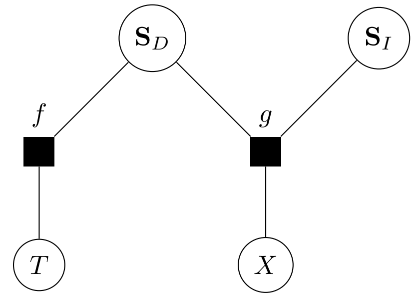


Figure 4: DriFDA factor graph

3.2 Drifting Feature Decomposition Analysis

Now we aim for a drift explanation method, i.e a technology which has the potential to uncover potentially semantically meaningful components from given data. Drifting Feature Decomposition Analysis (DriFDA) aims for a decomposition of the observed data X into a drifting part X_D and a non-drifting part X_I.

Let $(X,T) \sim p_t \otimes P_T$ with $\mathfrak{X} = \mathbb{R}^d$ and $\mathfrak{T} = \mathbb{R}_{\geq 0}$. By Theorem 4 drift is the same as dependency between X and T. If we model our data using independent, hidden source variables $\mathbf{S}_D$ and $\mathbf{S}_I$ that determine X and T, i.e. $f(\mathbf{S}_D) = T$ and $g(\mathbf{S}_D, \mathbf{S}_I) = X$, we arrive at the factor graph presented in Figure 4.

Therefore it is reasonable to define X_D resp. X_I as the best possible approximation of X using the information encoded in $\mathbf{S}_D$ resp. $\mathbf{S}_I$ only. In mathematical terms, we may express this idea using the notion of conditional expectation, i.e. we define

$$X_D := \mathbb{E}[g(\mathbf{S}_D, \mathbf{S}_I) \mid \mathbf{S}_D]$$
$$X_I := \mathbb{E}[g(\mathbf{S}_D, \mathbf{S}_I) \mid \mathbf{S}_I].$$

Since $\mathbf{S}_D$ and $\mathbf{S}_I$ are assumed to be independent and $\mathbf{S}_D$ determines T it follows that X_I must be independent of T and therefore it can not have drift. This on the other hand implies that X_D has to contain the entire drift information of X. Now by minimizing the information of $\mathbf{S}_D$ or maximizing the information of $\mathbf{S}_I$ (this depends on the chosen model), we force $\mathbf{S}_I$, and therefore X_I, to contain the entire non-drifting information of X. Concrete methods depend on the

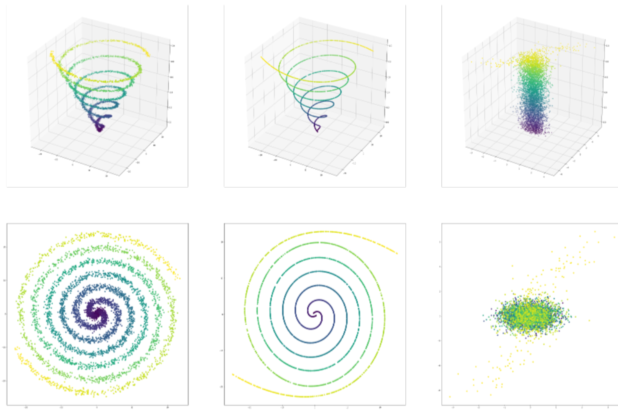


Figure 5: k-curve-DriFDA applied to twister data set. Original data (X), drift component (X_D), decomposition error ($X - X_D$) (Time is displayed as color and on Z-axis in upper row).

choice of the functional form of f and g.

3.2.1 Linear-DriFDA

A first approach to implement this method is by assuming that f and g are linear. Instead of estimating f, g, $\mathbf{S}_D$ and $\mathbf{S}_I$ all separately we may combine $\mathbf{S}_D$ and $\mathbf{S}_I$ resp. f and g into a single vector $\mathbf{S}$ resp. a single map represented by a matrix A. Under those assumptions we can compute X_D and X_I:

Lemma 1. *In the situation described above it holds*

$$
\begin{aligned}
X_D &= A\boldsymbol{S}_D + A\mathbb{E}[\boldsymbol{S}_I],\\
X_I &= A\boldsymbol{S}_I + A\mathbb{E}[\boldsymbol{S}_D],\\
X_D + X_I &= X + \mathbb{E}[X].
\end{aligned}
$$

Instead of forcing A to have a specific shape, we may simply train our model for a general linear form and apply feature selection, with respect to T, to determine whether a specific component of $\mathbf{S}$ belongs to $\mathbf{S}_D$ or $\mathbf{S}_I$. To assure that we can do this componentwise we need to assume the components of $\mathbf{S}$ to be independent. Note that this renders mutual information a particularly good choice as feature selection strategy, mirroring the assumed independence, i.e. non-redundancy of features (Hanchuan Peng et al., 2005).

To determine A and $\mathbf{S}$ we can use an independent component analysis (ICA) (Hyvärinen et al., 2001) to (X, T) - this leads to Algorithm 2.

Algorithm 2 Linear-DriFDA

procedure LINEAR-DRIFDA: DRIFTING FEATURE DECOMPOSITION ANALYSIS UNDER LINEARITY ASSUMPTION($(X, T) = (x(j), t(j))_{j=1\ldots N}$ data steam, n number of independent blind-sources, $I_{\min}$ minimal dependency)
    $(\mathbf{S}, A) \leftarrow \mathrm{ICA}((X, T), n)$
    for $i \in \{1, \ldots, n\}$ **do**
        $I_i \leftarrow I(S_i, T)$ ▷ Compute mutual information
        if $I_i \geq I_{\min}$ **then**
            $(S_D)_i \leftarrow S_i$
        else
            $(S_D)_i \leftarrow \frac{1}{N}\sum_{j=1}^{N} S_i(j)$ ▷ Mean value
        end if
    end for
    return $A\mathbf{S}_D$ ▷ Invert decomposition with drift relevant sources only
end procedure

3.2.2 k -curve-DriFDA

We model p_t as a mixture of drifting Gaussians, i.e.

$$p_t = \sum_{i=1}^{n} \lambda_i(t)\mathcal{N}(\mu_i(t), \sigma_i(t))$$

with $0 \leq \lambda_i(t)$, $\sum_i \lambda_i(t) = 1$ for all t. Then we can implement DriFDA as a generalized Gaussian-mixture clustering, where we estimate k curves that correspond to the means and variances. Under this assumption we may construct our model as follows: We choose

$$\mathbf{S}_D = (t, i)$$

where t is the time and i corresponds to the Gaussian generating the sample. It is natural to define

$$g((t, i), \mathbf{S}_I) := \sigma_i(t)\mathbf{S}_I + \mu_i(t).$$

So $\mathbf{S}_I$ generates row samples, which are then shaped using $\mathbf{S}_D$. Then it holds

$$X_D(t,i) = \mu_i(t) + \sigma_i(t)\mathbb{E}[\mathbf{S}_I],$$
$$\mathbf{S}_I \mid (t,i), X = \frac{1}{\sigma_i(t)}(X - \mu_i(t)).$$

This implies that adapting σ_i and μ_i corresponds to maximizing the Gaussianity, and therefore information, of $\mathbf{S}_I$ (cf. Hyvärinen et al. (2001)). Indeed, if σ_i and μ_i are adapted perfectly then $\mathbf{S}_I \sim \mathcal{N}(0,1)$, which then implies that $X_D(t,i) = \mu_i(t)$.

We may approximate the Gaussian clustering by a k-means algorithm for simplicity (cf. Bishop (2006)), i.e. we set $i = \text{argmin}_i \|\mu_i(t) - x\|$. The obtained algorithm can be found in the supplemental material (Algorithm 3). Notice that an incremental insertion of the data points may be used, rather than inserting them all at once, to reduce unwanted jumping behaviors of the mean-value-curves.

3.2.3 Experiments

We applied our two DriFDA variants to various artificial and real-world data sets. Since we want to decompose our data (X) into a drifting (X_D) and a non-drifting part we may quantify the reliability of our methods by measuring the dependency between the decomposition error $X - X_D$ and time T. To do so we use the prediction error with time as objective value and a k-nearest-neighbors model; the error on X serves as a baseline. For k-curves-DriFDA we used $k = 20$ RBF-networks with 10 prototypes, data was presented in 20 chunks. For Linear-DriFDA we used the mean overall mutual information as threshold. We use the Airlines, Electricity and Poker-Hand data sets from the MOA data set repository (Bifet et al., 2010). The artificial data sets *twister*: $p_t = \mathcal{N}(\alpha t(\sin(\beta t), \cos(\beta t)), \sigma)$, *spiral*: $p_t = \mathcal{N}(\alpha(\sin(\beta t), \cos(\beta t)), \sigma)$, Y: $p_t = \frac{1}{2}(\delta_{\max(0,t-\alpha)} + \delta_{-\max(0,t-\alpha)}) \times \mathcal{U}([0,1])$ and *square*: $p_t = \delta_{(\alpha t, \beta t)} * \mathcal{U}^2([0,1])$ were designed to provide ground truth, here δ denotes the Dirac measure and $\mathcal{U}$ the uniform measure.

Results are displayed in Table 2. Though Linear-DriFDA is only capable of finding linear relationships it works surprisingly well on a large fraction of the data sets. For k-curves-DriFDA, some results are excellent. The number of chunks used to present the data seems to be a very relevant hyper-parameter, hence an automatic optimization scheme or a robust selection technology would be helpful.

Table 2: Mean and variance (if ≥ 0.01) over 8 runs. $*$ no chunk wise adaption was used; † 40 chunks and $k = 4$ curves were used.

Data set	k-curves-DriFDA	Lin.-DriFDA
Airlines	0.67^*	$0.92(\pm 0.01)$
Electricity	0.54	$0.70(\pm 0.01)$
Poker-Hand	0.21	$0.22(\pm 0.01)$
twister	$0.13^\dagger$	0.92
spiral	0.02^*	0.13
Y	0.06^*	0.02
square	0.04^*	0.02

4 DISCUSSION

We have presented formal definitions of drift in continuous time, this way substantiating common practice such as learning on time windows, drift detection by referring to model errors, or change point detection by a mathematical justification. In addition, we derived a particularly elegant novel characterization in terms of independence of observations and time, which opens the way towards efficient and flexible algorithms which are based on classical independence tests. We have demonstrated this potential by a novel drift detection method, and a novel decomposition method which can disentangle drifting and non-drifting part of observed signals. The latter has so far been tested in first benchmarks only, displaying a robust and surprisingly efficient behavior. The suitability to uncover semantically meaningful signals in the context of larger applications and specific domain expertise is subject of ongoing work.

Acknowledgement

Funding by the VW Foundation in the frame of the project IMPACT, and by the BMBF for the project ITS_ML, grant number 01IS18041, is gratefully acknowledged.

References

Alippi, C., Boracchi, G., and Roveri, M. (2017). Hierarchical change-detection tests. *IEEE Trans. Neural Netw. Learning Syst.*, 28(2):246–258.

Aminikhanghahi, S. and Cook, D. J. (2017). A survey of methods for time series change point detection. *Knowl. Inf. Syst.*, 51(2):339–367.

Baena-García, M., Campo-Ávila, J., Fidalgo-Merino, R., Bifet, A., Gavald, R., and Morales-Bueno, R. (2006). Early drift detection method.

Bifet, A. and Gavaldà, R. (2007). Learning from time-changing data with adaptive windowing. In *Proceedings of the Seventh SIAM International Conference on Data Mining, April 26-28, 2007, Minneapolis, Minnesota, USA*, pages 443–448.

Bifet, A., Holmes, G., Kirkby, R., Pfahringer, B., and Braun, M. (2010). Moa: Massive online analysis. *Journal of Machine Learning Research 11: 1601-1604*.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

Dasu, T., Krishnan, S., Venkatasubramanian, S., and Yi, K. (2006). An information-theoretic approach to detecting changes in multi-dimensional data streams. In *In Proc. Symp. on the Interface of Statistics, Computing Science, and Applications*.

Dette, H. and Wied, D. (2016). Detecting relevant changes in time series models. *Journal of the Royal Statistical Society Series B*, 78(2):371–394.

Ditzler, G. and Polikar, R. (2011). Hellinger distance based drift detection for nonstationary environments. In *2011 IEEE Symposium on Computational Intelligence in Dynamic and Uncertain Environments, CIDUE 2011, Paris, France, April 13, 2011*, pages 41–48.

Ditzler, G., Roveri, M., Alippi, C., and Polikar, R. (2015). Learning in nonstationary environments: A survey. *IEEE Comp. Int. Mag.*, 10(4):12–25.

Gama, J., Medas, P., Castillo, G., and Rodrigues, P. P. (2004). Learning with drift detection. In *Advances in Artificial Intelligence - SBIA 2004, 17th Brazilian Symposium on Artificial Intelligence, São Luis, Maranhão, Brazil, September 29 - October 1, 2004, Proceedings*, pages 286–295.

Gama, J. a., Žliobaitė, I., Bifet, A., Pechenizkiy, M., and Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Comput. Surv.*, 46(4):44:1–44:37.

Goldenberg, I. and Webb, G. I. (2019). Survey of distance measures for quantifying concept drift and shift in numeric data. *Knowl. Inf. Syst.*, 60(2):591–615.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. J. (2006). A kernel method for the two-sample-problem. In *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*, pages 513–520.

Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. J. (2007). A kernel statistical test of independence. In *Advances in Neural Information Processing Systems 20, Proceedings of the Twenty-First Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 3-6, 2007*, pages 585–592.

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., and Schölkopf, B. (2009). *Covariate shift and local learning by distribution matching*, pages 131–160. MIT Press, Cambridge, MA, USA.

Hanchuan Peng, Fuhui Long, and Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238.

Hulten, G., Spencer, L., and Domingos, P. M. (2001). Mining time-changing data streams. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, San Francisco, CA, USA, August 26-29, 2001*, pages 97–106.

Hyvärinen, A., Karhunen, J., and Oja, E. (2001). *Independent Component Analysis.* Wiley.

Kifer, D., Ben-David, S., and Gehrke, J. (2004). Detecting change in data streams. In *Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30*, VLDB '04, pages 180–191. VLDB Endowment.

Matteson, D. S. and James, N. A. (2014). A nonparametric approach for multiple change point analysis of multivariate data. *Journal of the American Statistical Association*, 109(505):334–345.

PAGE, E. S. (1954). CONTINUOUS INSPECTION SCHEMES. *Biometrika*, 41(1-2):100–115.

Park, K. I. (2018). Fundamentals of probability and stochastic processes with applications to communications.

Parthasarathy, K. R. (1967). *Probability measures on metric spaces*, volume 3 of *Probability and mathematical statistics ; 3*. Acad. Pr., New York [u.a.].

Roveri, M. (2019). Learning discrete-time markov chains under concept drift. *IEEE Trans. Neural Netw. Learning Syst.*, 30(9):2570–2582.

Song, X., Wu, M., Jermaine, C., and Ranka, S. (2007). Statistical change detection for multi-dimensional data. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, pages 667–676, New York, NY, USA. ACM.

Street, W. N. and Kim, Y. (2001). A streaming ensemble algorithm (SEA) for large-scale classification. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, San Francisco, CA, USA, August 26-29, 2001*, pages 377–382.

Vorburger, P. and Bernstein, A. (2006). Entropy-based concept shift detection. In *Proceedings of the 6th IEEE International Conference on Data Mining (ICDM 2006), 18-22 December 2006, Hong Kong, China*, pages 1113–1118.

Webb, G. I., Lee, L. K., Petitjean, F., and Goethals, B. (2017). Understanding concept drift. *CoRR*, abs/1704.00362.

A SUPPLEMENTAL MATERIAL

A.1 Theorems and proofs

We will now give additional definition, remarks, theorems, lemmas and corollaries. In particular we will provide proofs for the theorems given in the paper. Note that we will include the definitions and theorems given in the paper using the same numeration as before.

A.1.1 Definition of a drift process

Definition A.1. Let $(\mathfrak{T},\mathcal{B}),(\mathfrak{X},\mathcal{A})$ be two measurable spaces. A *Markov kernel* is a map $\kappa : \mathfrak{T}\times\mathcal{A}\to\mathbb{R}$ such that:

1. $t\mapsto\kappa(t,A)$ is measurable for all $A\in\mathcal{A}$,
2. $\kappa(t,\cdot)$ is a probability measure for all $t\in\mathfrak{T}$.

Definition 1. Let $(\mathfrak{T},\mathcal{B})$ and $(\mathfrak{X},\mathcal{A})$ be two measurable spaces. A *drift process* (p_t,P_T) is a Markov kernel p_t from $\mathfrak{T}$ to $\mathfrak{X}$ and a probability measure P_T on $\mathfrak{T}$.

Remark A.1. Notice that

1. Markov kernels are exactly the measurable maps $\kappa : \mathfrak{T}\to\mathbf{P}(\mathfrak{X})$, where $\mathbf{P}(\mathfrak{X})$ is the set of all probability measures on $\mathfrak{X}$ equipped with the initial σ-algebra induced by all evaluation maps $\Phi_A : P\mapsto P(A)$ for $A\in\mathcal{A}$.
2. If we assume that $\mathfrak{T}$ is a topological space, then every continuous map $\kappa : \mathfrak{T}\to\mathbf{P}(\mathfrak{X})$ is a Markov kernel, here we equip $\mathbf{P}(\mathfrak{X})$ with the topology induced by the total variation norm. This follows by writing $\|P-Q\|_{\mathrm{TV}}=\sup_{A\in\mathcal{A}}|P(A)-Q(A)|$ implying $t\mapsto\kappa(t,A)$ is continuous, and hence measurable, for all $A\in\mathcal{A}$.

Definition A.2. Let $\mathfrak{X}$ be some set and $\mathcal{A}_0\subset 2^{\mathfrak{X}}$ be a set of subsets of $\mathfrak{X}$. Then *the σ-algebra generated by* $\mathcal{A}_0$, denoted by $\sigma(\mathcal{A}_0)$, is defined as the (with respect to inclusion) smallest σ-algebra on $\mathfrak{X}$ that contains $\mathcal{A}_0$.

Remark A.2. It can be shows that

$$\sigma(\mathcal{A}_0)=\bigcap_{\mathcal{S}\in\mathcal{F}(\mathcal{A}_0)}\mathcal{S},\qquad\text{where}$$
$$\mathcal{F}(\mathcal{A}_0)=\left\{\mathcal{S}\subset 2^X\ \middle|\ {\scriptstyle \mathcal{S}\text{ is }\sigma\text{-algebra on }\mathfrak{X} \atop \text{and }\mathcal{A}_0\subset\mathcal{S}}\right\}.$$

Definition A.3. Let $(\mathfrak{X},\mathcal{A})$ be a measurable space. We say that $\mathcal{A}_0\subset\mathcal{A}$ is *a generator of* $\mathcal{A}$ iff $\sigma(\mathcal{A}_0)=\mathcal{A}$. We say that $\mathcal{A}_0$ is *stable under finite intersections* iff for all $A,B\in\mathcal{A}_0$ it holds $A\cap B\in\mathcal{A}_0$.

We will make heavy use of the following, well known theorem:

Theorem A.1. *Let $(\mathfrak{X},\mathcal{A})$ be a measurable space and P and Q be probability measures on $\mathfrak{X}$. Let $\mathcal{A}_0\subset\mathcal{A}$ be a generating set, stable under finite intersections.*

Suppose that $P(A)=Q(A)$ for all $A\in\mathcal{A}_0$, then it holds $P=Q$, i.e. $P(A)=Q(A)$ for all $A\in\mathcal{A}$.

Proof. Well known. □

Definition A.4. Let $(\mathfrak{T},\mathcal{B}),(\mathfrak{X},\mathcal{A})$ be two measurable spaces. Let P_T and P_X be probability measures on $\mathfrak{T}$ resp. $\mathfrak{X}$. We call a probability measure P on $(\mathfrak{T}\times\mathfrak{X},\sigma(\mathcal{B}\otimes\mathcal{A}))$, where $\mathcal{B}\otimes\mathcal{A}=\{B\times A\mid A\in\mathcal{A},B\in\mathcal{B}\}$, such that

$$P_T(B)P_X(A)=P(B\times A)$$

for all $A\in\mathcal{A},B\in\mathcal{B}$ the *product measure* of P_T and P_X and denote it by $P_T\times P_X$.

Remark A.3. It can be shown that product measures always exist and that they are uniquely determined, justifying the notation above.

Remark A.4 (Fubini's theorem for Markov kernels). Let $(\mathfrak{T},\mathcal{B}),(\mathfrak{X},\mathcal{A})$ be two measurable spaces. Let p_t be a Markov kernel form $\mathfrak{T}$ to $\mathfrak{X}$, and P_T a measure on $\mathfrak{T}$. There exists a unique probability measure P on $(\mathfrak{T}\times\mathfrak{X},\sigma(\mathcal{B}\otimes\mathcal{A}))$, such that for all $A\in\mathcal{A},B\in\mathcal{B}$ it holds

$$P(B\times A)=\int_B p_t(A)P_T(\mathrm{d}t).$$

We denote this uniquely determined measure by

$$p_t\otimes P_T:=P.$$

Definition 2. Let (p_t, P_T) be a drift process. We say that p_t has *no drift* or does not drift if $p_t = p_s$ holds P_T-a.s., i.e. $(P_T \times P_T)(\{(s,t) \in \mathfrak{T} \times \mathfrak{T} \mid p_t = p_s\}) = 1$. We say that p_t has *drift* or is drifting if it is not the case that it does not drift.

Lemma A.1. *Let $\mathfrak{T}$ be a countable index set and $(\mathfrak{X}, \mathcal{A})$ be a measurable space. Then there exists a σ-algebra $\mathcal{B}$ on $\mathfrak{T}$ and a probability measure P_T on $(\mathfrak{T}, \mathcal{B})$ such that for every non-probabilistic drift process p_t it holds: (p_t, P_T) has drift if and only if there exists $t, s \in \mathfrak{T}$ such that $p_t \neq p_s$.*

Proof. Choose $\mathcal{B}$ as the power set of $\mathfrak{T}$ and let $f : \mathbb{N} \to \mathfrak{T}$ be a counting function. Now define

$$P_{\mathbb{N}}(A) = \frac{6}{\pi^2} \sum_{i \in A} \frac{1}{i^2}$$

and P_T as the image measure of $P_{\mathbb{N}}$ under f. Since P_T has no null sets the statement follows. $\square$

Definition 3. Let (p_t, P_T) be a drift process. We say that p_t has *no proper drift* iff there exists a drift process (p'_t, P'_T), such that $p_t \otimes P_T = p'_t \otimes P'_T$, that does not drift. We say that p_t has *proper drift* iff it is not the case that it has no proper drift.

We make use of the following, well known lemma:

Lemma A.2. *Let $(p_t, P_T), (p'_t, P'_T)$ be two drift processes. If $\mathcal{B}$ has a countable generating set, stable under finite intersection, then it holds*

$$p_t \otimes P_T = p'_t \otimes P'_T$$

if and only if

$$P_T = P'_T \qquad \text{and} \qquad p_t = p'_t \quad P_T - a.s..$$

Proof. Recall that P_T, P'_T are probability measures and that p_t, p'_t are Markov kernels. Then this is well known (and easy). $\square$

A.1.2 Drift as change of distribution

Definition 4. Let (p_t, P_T) be a drift process. We say that p_t is *constant* if there exists a probability measure $P_X \in \mathbf{P}(\mathfrak{X})$ such that $p_t = P_X$ for P_T-a.s. all $t \in \mathfrak{T}$. We say that p_t has a *change of distribution* or is changing iff it is not the case that p_t is constant.

Lemma A.3. *Let (p_t, P_T) be a drift process. Then p_t is constant if and only if there exists a $t_0 \in \mathfrak{T}$ such that $p_t = p_{t_0}$ for P_T-a.s. all $t \in \mathfrak{T}$. In particular we may choose $P_X = p_{t_0}$.*

Proof. For "$\Leftarrow$" choose $P_X = p_{t_0}$, for "$\Rightarrow$" note that there exists a $t_0 \in \mathfrak{T}$ such that $p_{t_0} = P_X$ and hence $p_{t_0} = p_t$ for P_T-a.s. all $t \in \mathfrak{T}$. $\square$

Corollary A.1. *Let (p_t, P_T) be a drift process. Assume that p_t is constant. Then it holds $P_X = \int p_t \mathrm{d}P_T$.*

Proof. Let $t_0 \in \mathfrak{T}$ as in Lemma A.3. Then it holds $\int p_t \mathrm{d}P_T = \int p_{t_0} \mathrm{d}P_T = p_{t_0} \int 1 \mathrm{d}P_T = P_X$. $\square$

We will now give a proof of Theorem 1:

Theorem 1. *Let (p_t, P_T) be a drift process. Then p_t is constant if and only if p_t has no drift.*

Proof. **"$\Rightarrow$":** Denote by $C = \{t | p_t = P_X\}$ and by $D = \{(t,s) | p_t = p_s\}$. Obviously it holds $C \times C \subset D$. Since p_t is constant we have $P_T(C) = 1$ and hence

$$1 = (P_T \times P_T)(C \times C) \leq (P_T \times P_T)(D).$$

"$\Leftarrow$": Since P_T is finite we may write $(P_T \times P_T)(A) = \int P_T(A^x) P_T(\mathrm{d}x)$, where $A^x = \{y | (x,y) \in A\}$. This implies that $P_T(\{s \in T | p_s = p_t\}) = 1$ for P_T-a.s. all $t \in \mathfrak{T}$. Therefore the statement follows by Lemma A.3. $\square$

Lemma A.4. *Let (p_t, P_T) be a drift process. Then p_t has no proper drift if and only if it exists a probability measure P_X such that*

$$p_t \otimes P_T = P_X \times P_T.$$

If P_X exists, then it is unique with this property.

Proof. **"$\Rightarrow$":** Suppose p_t has no proper drift, let p'_t be the not drifting drift process as in the definition. By Theorem 1 there exists a P_X such that

$$P_X \times P_T = p'_t \otimes P_T = p_t \otimes P_T.$$

"$\Leftarrow$": We may consider $t \mapsto P_X$ as a constant kernel, i.e. (P_X, P_T) is a drift process. Clearly P_X does not drift and hence p_t has no proper drift.

Uniqueness: for all $A \in \mathcal{A}$ we have

$$\begin{aligned} P'_X(A) &= (P'_X \times P_T)(A \times \mathfrak{T}) \\ &= (P_X \times P_T)(A \times \mathfrak{T}) = P_X(A). \end{aligned}$$

□

A.1.3 Drift as change of model

Definition 5. Let (p_t, P_T) be a drift process. For a P_T-non-null set $A \in \mathcal{B}$ we define the *$\mathfrak{T}$-invariant model* of p_t over A as the marginalization of $(p_t \otimes P_T)(\cdot|\mathfrak{X} \times A)$ onto $\mathfrak{X}$ or equivalent

$$p_A := \frac{1}{P_T(A)} \int_A p_t \mathrm{d}P_T.$$

Remark A.5. We would like to point out that this notion is by far less theoretical than p_t, since single points tend to be null sets, i.e. even with an infinite amount of data, the probability to observe even a single sample at time t is still zero and therefore we cannot estimate p_t directly, even though we have an infinite amount of samples to estimate p_A.

In addition notice that (by Corollary A.1) a drift process is constant if $p_t = p_T$ for P_T-a.s. all $t \in \mathfrak{T}$, so the notion of model we consider is turned into the classical model, if we assume that no drift takes place.

Lemma A.5. *Let (p_t, P_T) be a drift process and let $A, B, C \in \mathcal{B}$ be pair wise disjoint, non-null sets. Suppose that*

$$p_A = p_{B \cup C} \qquad \text{and} \qquad p_{A \cup C} = p_B$$

then it holds $p_A = p_B = p_C$.

Proof. Its a computation: Denote by $\lambda := \frac{P_T(A)}{P_T(A)+P_T(C)}, \mu := \frac{P_T(B)}{P_T(B)+P_T(C)}$. It holds

$$\begin{aligned} &0 < P_T(A), P_T(B), P_T(C) < 1 \\ &\Rightarrow 0 < \lambda, \mu < 1, \\ p_{A\cup C} &= \frac{P_T(A)p_A + P_T(C)p_C}{P_T(A) + P_T(C)} \\ &= \lambda p_A + (1-\lambda)p_C, \\ p_{B\cup C} &= \frac{P_T(B)p_B + P_T(C)p_C}{P_T(B) + P_T(C)} \\ &= \mu p_B + (1-\mu)p_C. \end{aligned}$$

Solving the last two equations for p_C it holds

$$\begin{aligned} \frac{p_A - \lambda p_B}{1-\lambda} &= p_C = \frac{p_B - \mu p_A}{1-\mu} \\ \Leftrightarrow p_A - \lambda p_B - \mu p_A + \lambda\mu p_B &= p_B - \lambda p_B - \mu p_A + \lambda\mu p_A \\ \Leftrightarrow (1-\lambda\mu)p_A &= (1-\lambda\mu)p_B \\ &\overset{\mu\lambda\neq 1}{\Leftrightarrow} p_A = p_B \\ \Rightarrow p_C = \frac{p_B - \lambda p_B}{1-\lambda} &= p_B \end{aligned}$$

as stated. □

Definition 6. We say that a pair of P_T-non-null sets $A, B \in \mathcal{B}$ are *alternating sets* iff $p_A \neq p_B$. If alternating sets exist, then we say that p_t has *model drift.*

Corollary A.2. *Let (p_t, P_T) be a drift process and let $A, B \in \mathcal{B}$ be disjoint, alternating sets. Then A, A^C or B, B^C are alternating, too.*

Proof. Let $C = (A \cup B)^C$. If C is a null set, then $p_A \neq p_B = p_{B\cup C}$ and hence we have that A, A^C are alternating. If C is not a null set, then $p_A \neq p_{B\cup C}$ or $p_B \neq p_{A\cup C}$, by Lemma A.5, and hence A, A^C resp. B, B^C are alternating. □

Corollary A.3. *Let (p_t, P_T) be a drift process and let $A, B \in \mathcal{B}$ such that $P_T(A), P_T(B) \in (0, 1)$. If*

$$p_A = p_{A^C} \qquad \text{and} \qquad p_B = p_{B^C}$$

then it holds $p_A = p_B$.

Proof. We may find pair wise disjoint sets $D, E, F \in \mathcal{B}$, such that $A, A^C, B, B^C \in \sigma(\{D, E, F\})$. Without loss of generality we may assume $A = D \cup E, A^C = F, B = D, B^C = E \cup F$. If E is a null set, then trivially it holds $p_A = p_B$ otherwise we may apply Lemma A.5 to see that $p_D = p_E$ and hence $p_A = p_B$. □

We will now give a proof of Theorem 2:

Theorem 2. *Let (p_t, P_T) be a drift process and let $\mathcal{B}_0 \subset \mathcal{B}$ a generating set (i.e. $\sigma(\mathcal{B}_0) = \mathcal{B}$), which is stable under finite intersections. Then the following properties are equivalent:*

1. p_t has proper drift,

2. p_t has model drift,

3. there exist alternating sets A, A^C, with $A \in \mathcal{B}_0$.

Proof. **"3. $\Rightarrow$ 2.":** is clear. **"2. $\Rightarrow$ 1.":** Let A, B be alternating sets. Assume that p_t has no proper drift. By Lemma A.4 we have $p_t \otimes P_T = P_X \times P_T$ and hence $p_A = P_X = p_B$ which is a contradiction.

"1. $\Rightarrow$ 3.": Assume that for all $A \in \mathcal{B}_0$ with $P_T(A) \in (0,1)$ it holds $p_A = p_{A^C}$. Then it follows by Corollary A.3 that $p_A = p_B$ for all $A, B \in \mathcal{B}_0$ with $P_T(A), P_T(B) \in (0,1)$. Since adding a null set to A wount change p_A we have that $p_A = p_B$ for all non-null sets $A, B \in \mathcal{B}_0$.

Hence $P_X = p_A$ is well defined for any non-null set $A \in \mathcal{B}_0$. Now it holds

$$\begin{aligned}(p_t \otimes P_T)(B \times C) &\overset{\text{def. } p_B}{=} P_T(B) p_B(C) \\ &= P_T(B) P_X(C)\end{aligned}$$

for all $B \in \mathcal{B}_0, C \in \mathcal{A}$. Since $\sigma(\mathcal{B}_0 \otimes \mathcal{A}) = \sigma(\mathcal{B} \otimes \mathcal{A})$ we have that $p_t \otimes P_T = P_X \times P_T$ which by Lemma A.4 implies that p_t has no proper drift. This is a contradiction. ☐

We will now give a proof of Corollary 1:

Corollary 1. *Let (p_t, P_T) be a drift process and suppose that $\mathfrak{T} \in \{[a,b], \mathbb{R}_{\geq 0}, \mathbb{R}\}, \mathcal{B} = \mathfrak{B}(\mathfrak{T})$. If p_t has proper drift, then there exists a change-point, i.e. it exists a $t_0 \in \mathbb{R}$ such that $p_{\{t<t_0\}} \neq p_{\{t \geq t_0\}}$.*

Proof. Recall that $\{(-\infty, a) | a \in \mathbb{Q}\}$ is a generator of $\mathfrak{B}(\mathbb{R})$; therefore we may find a $t_0 \in \mathbb{R}$ such that $A = (-\infty, t_0) \cap \mathfrak{T}, A^C$ are alternating sets (Theorem 2). ☐

A.1.4 Drift as non-stationarity of a stochastic process

Definition A.5. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $(\mathfrak{X}, \mathcal{A})$ be a measurable space and $\mathfrak{T}$ be an index set. A *stochastic process* is a collection of $\mathfrak{T}$-indexed, $\mathfrak{X}$-valued random variables $(X_t)_{t \in \mathfrak{T}}$. By fixing a $\omega \in \Omega$ we obtain a map $X_\bullet(\omega) : \mathfrak{T} \to \mathfrak{X}, t \mapsto X_t(\omega)$; we refer to those maps as *the paths of* X_t. We say that X_t has $\mathbb{P}$*-a.s. continuous paths* iff $t \mapsto X_t(\omega)$ is continuous for $\mathbb{P}$-a.s. all $\omega \in \Omega$.

Definition 7. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A stochastic process $X_t : \Omega \times \mathfrak{T} \to \mathfrak{X}$ is *stationary* if

$$\mathbb{P} \circ (X_{t_1}, ..., X_{t_n})^{-1} = \mathbb{P} \circ (X_{t_1+\tau}, ..., X_{t_n+\tau})^{-1}$$

for all $t_1, ..., t_n \in \mathfrak{T}$, $\tau \in \mathfrak{T}$ and $n \in \mathbb{N}$.

We will now give a proof of Theorem 3:

Theorem 3. *Let X_t be a stochastic process. For every sequence $t_1, ..., t_n \in \mathfrak{T}$ we obtain a non-probabilistic drift process by setting $p_\tau^{(t_1,...,t_n)} = \mathbb{P} \circ (X_{t_1+\tau}, ..., X_{t_n+\tau})^{-1}$. If X_t is stationary then $(p_\tau^{(t_1,...,t_n)}, P_T)$ has no drift for all probability measures P_T on $\mathfrak{T}$, $t_1, ..., t_n \in \mathfrak{T}$ and $n \in \mathbb{N}$.*
Furthermore, if P_T has no null sets, then the notion of stationarity of X_t and $(p_\tau^{(t_1,...,t_n)}, P_T)$ having no drift for all $t_1, ..., t_n \in \mathfrak{T}$ and $n \in \mathbb{N}$, are equivalent.

Proof. Using Lemma A.1 the problem boils down to remarking that the empty set always has measure zero. ☐

Corollary A.4. *Let X_t be a $\mathbb{R}^d$-valued stochastic process with $\mathbb{P}$-a.s. continuous paths. Let P_T be a probability measure on $\mathfrak{T}$. Suppose that Lebesgue-measure is absolutely continuous with respect to P_T. Then X_t is stationary if and only if $(p_\tau^{(t_1,...,t_n)}, P_T)$ has no drift for every sequence $t_1, ..., t_n \in \mathfrak{T}$ and $n \in \mathbb{N}$..*

Proof. "$\Leftarrow$" follows by Theorem 3 and Theorem 1. Show "$\Rightarrow$": Since $p_\tau^{(t_1,...,t_n)}$ has no drift we may find a τ_0 such that $p_{\tau_0}^{(t_1,...,t_n)} = p_\tau^{(t_1,...,t_n)}$ for P_T-a.s. all $\tau \in \mathfrak{T}$ (Corollary A.3).

It remains to prove that $\tau \mapsto p_\tau^{(t_1,...,t_n)}$ is continuous with respect to total variation norm, then $t \mapsto \|p_\tau^{(t_1,...,t_n)} - p_{\tau_0}^{(t_1,...,t_n)}\|$ is continuous. Since it is equals 0 P_T-a.s. and continuous it follows that it is 0 everywhere, since Lebesgue-measure is absolutely continuous with respect to P_T, which implies $p_\tau^{(t_1,...,t_n)} = p_{\tau_0}^{(t_1,...,t_n)}$ for all $\tau \in \mathfrak{T}$.

By the triangle inequality $\tau \mapsto (X_{t_1+\tau}, ..., X_{t_n+\tau})$ is $\mathbb{P}$-a.s. continuous, it is therefore enough to show that a stochastic process with a.s. continuous paths has continuously changing marginal distributions, but this is well known (also known as: sample continuity implies continuity in distribution). $\square$

A.1.5 Drift as dependency between data and time

Definition 8. Let (p_t, P_T) be a drift process and let $(X,T) \sim p_t \otimes P_T$ a pair of random variables. We say that p_t has *dependency drift* if X and T are not independent.

We will now give a proof of Theorem 4:

Theorem 4. *Let (p_t, P_T) be a drift process. Then p_t has proper drift if and only if it has dependency drift.*

Proof. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the underlying probability space, i.e. X and T are measurable maps from Ω to $\mathfrak{X}$ resp. $\mathfrak{T}$. X and T are independent if and only if

$$(p_t \otimes P_T)(A \times B) = \mathbb{P}_{X,T}(A \times B) = \mathbb{P}_X(A)\mathbb{P}_T(B)$$

holds for all $A \in \mathcal{A}, B \in \mathcal{B}$. By setting $A = \mathfrak{X}$ we obtain $\mathbb{P}_T = P_T$ and therefore $p_t \otimes P_T = \mathbb{P}_X \times P_T$ which, by Lemma A.4, holds if and only if p_t has no proper drift. $\square$

A.1.6 Linear-DriFDA

We will now give a proof of Lemma 1:

Lemma 1. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $\boldsymbol{S}_D$ and $\boldsymbol{S}_I$ be independent, $\mathbb{R}^{n_D}$- resp. $\mathbb{R}^{n_I}$-valued random variables. Let $A : \mathbb{R}^{n_D+n_I} \to \mathbb{R}^d$ be a linear map.*

Denote by $\boldsymbol{S} := (\boldsymbol{S}_D, \boldsymbol{S}_I)$, $X := A\boldsymbol{S}$, $X_D = \mathbb{E}[X|\boldsymbol{S}_D]$ and $X_I = \mathbb{E}[X|\boldsymbol{S}_I]$. Then it holds

$$\begin{aligned}
X_D &= A\boldsymbol{S}_D + A\mathbb{E}[\boldsymbol{S}_I],\\
X_I &= A\boldsymbol{S}_I + A\mathbb{E}[\boldsymbol{S}_D],\\
X_D + X_I &= X + \mathbb{E}[X].
\end{aligned}$$

Proof. Without loss of generality we may assume that $\mathbf{S} = \mathbf{S}_D + \mathbf{S}_I$, i.e. $\mathbf{S}_D$ and $\mathbf{S}_I$ "use different dimension". Now its a computation

$$\begin{aligned}
X_D &= \mathbb{E}[A\mathbf{S}|\mathbf{S}_D]\\
&= A(\mathbb{E}[\mathbf{S}_D|\mathbf{S}_D] + \mathbb{E}[\mathbf{S}_I|\mathbf{S}_D])\\
&= A\mathbf{S}_D + A\mathbb{E}[\mathbf{S}_I]\\
X_I &= \cdots = A\mathbf{S}_I + A\mathbb{E}[S_D]\\
\Rightarrow X &= A\mathbf{S}\\
&= A\mathbf{S}_D + A\mathbf{S}_I\\
&= (X_D - A\mathbb{E}[\mathbf{S}_I]) + (X_I - A\mathbb{E}[\mathbf{S}_D])\\
&= X_D + X_I - \mathbb{E}[A\mathbf{S}]\\
&= X_D + X_I - \mathbb{E}[X].
\end{aligned}$$

Note that we dropped the dimensions containing T for simplicity. $\square$

A.2 Algorithms

Algorithm 3 k-curve-DriFDA

```
procedure k-CURVE-DRIFDA: DRIFTING FEATURE DECOMPOSITION ANALYSIS VIA k-CURVES((x_j, t_j) data steam, k number of curves)
    D ← ∅
    Initialize μ_i, i = 1..., k using k-means.
    while Not at end of stream do
        Draw next batch D_new from stream
        D ← D ∪ D_new
        while μ_i not converged do
            for i = 1, ..., k do
                D_i ← ∅
            end for
            for (x, t) ∈ D do
                i* ← argmin_i ‖x − μ_i(t)‖
                D_{i*} ← D_{i*} ∪ {(x, t)}
            end for
            for i = 1, ..., k do
                Retrain μ_i using D_i
            end for
        end while
    end while
    return (μ_i)_{i=1,...,k}
end procedure
```

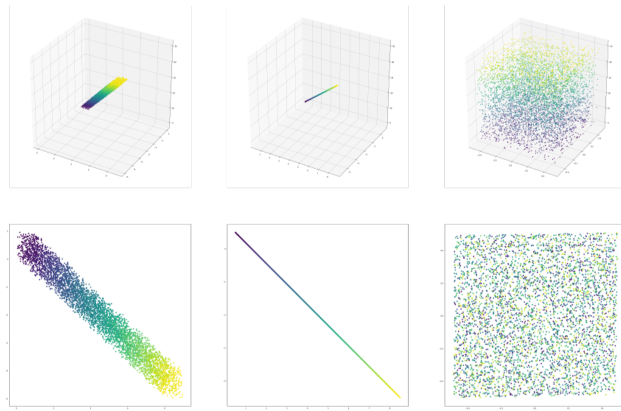


Figure 6: Linear-DriFDA applied to square data set

A.3 Visualization of DriFDA

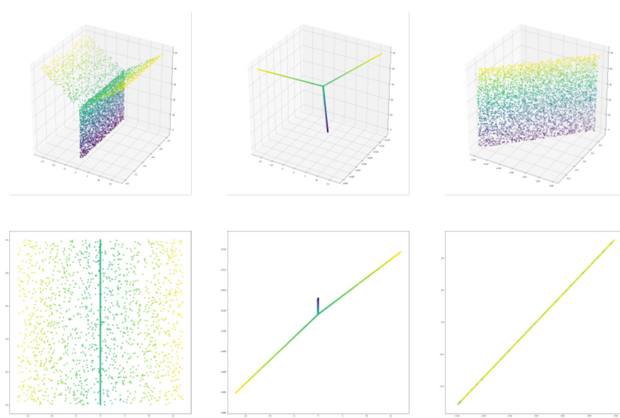


Figure 7: Linear-DriFDA applied to Y data set