Probabilistic Time Series Forecasting with Structured Shape and Temporal Diversity
Vincent Le Guen, Nicolas Thome

To cite this version:
Vincent Le Guen, Nicolas Thome. Probabilistic Time Series Forecasting with Structured Shape and Temporal Diversity. NeurIPS 2020, Dec 2020, Vancouver, Canada. hal-02992358

HAL Id: hal-02992358
https://hal.science/hal-02992358
Submitted on 6 Nov 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Probabilistic Time Series Forecasting with Structured Shape and Temporal Diversity

Vincent Le Guen 1,2 vincent.le-guen@edf.fr
Nicolas Thome 2 nicolas.thome@cnam.fr

1 EDF R&D, Chatou, France
2 Conservatoire National des Arts et Métiers, CEDRIC, Paris, France

Abstract

Probabilistic forecasting consists in predicting a distribution of possible future outcomes. In this paper, we address this problem for non-stationary time series, which is very challenging yet crucially important. We introduce the STRIPE model for representing structured diversity based on shape and time features, ensuring both probable predictions while being sharp and accurate. STRIPE is agnostic to the forecasting model, and we equip it with a diversification mechanism relying on determinantal point processes (DPP). We introduce two DPP kernels for modeling diverse trajectories in terms of shape and time, which are both differentiable and proved to be positive semi-definite. To have an explicit control on the diversity structure, we also design an iterative sampling mechanism to disentangle shape and time representations in the latent space. Experiments carried out on synthetic datasets show that STRIPE significantly outperforms baseline methods for representing diversity, while maintaining accuracy of the forecasting model. We also highlight the relevance of the iterative sampling scheme and the importance to use different criteria for measuring quality and diversity. Finally, experiments on real datasets illustrate that STRIPE is able to outperform state-of-the-art probabilistic forecasting approaches in the best sample prediction.

1 Introduction

Time series forecasting consists in analysing historical signal correlations to anticipate future outcomes. In this work, we focus on probabilistic forecasting in non-stationary contexts, i.e. we aim at producing plausible and diverse predictions where future trajectories can present sharp variations. This forecasting context is of crucial importance in many applicative fields, e.g. climate [60, 33, 14], optimal control or regulation [64, 40], traffic flow [38, 37], healthcare [7, 1], stock markets [13, 6], etc. Our motivation is illustrated in the example of the blue input in Figure 1(a): we aim at performing predictions covering the full distribution of future trajectories, whose samples are shown in green.

State-of-the-art methods for time series forecasting currently rely on deep neural networks, which exhibit strong abilities in modeling complex nonlinear dependencies between variables and time. Recently, increasing attempts have been made for improving architectures for accurate predictions [30, 52, 36, 41, 34] or for making predictions sharper, e.g. by explicitly modeling dynamics [8, 15, 49], or by designing specific loss functions addressing the drawbacks of blurred prediction with mean squared error (MSE) training [11, 46, 32, 56]. Although Figure 1(b) shows that such approaches produce sharp and realistic forecasts, their deterministic nature limits them to a single trajectory prediction without uncertainty quantification.
Figure 1: We address the probabilistic time series forecasting problem. (a) Recent deep learning models include a specific loss enabling sharp predictions [11, 46, 32, 56] (b), but are inadequate for producing diverse forecasts. On the other hand, probabilistic forecasting approaches based on generative models [63, 45] lose the ability to generate sharp forecasts (c). The proposed STRIPE model (d) produces both sharp and diverse future forecasts.

Methods targeting probabilistic forecasting enable to sample diverse predictions from a given input. This includes deterministic methods that predict the quantiles of the predictive distribution or probabilistic methods that sample future values from a learned approximate distribution, parameterized explicitly (e.g., Gaussian [51, 44, 50]), or implicitly with latent generative models [63, 28, 45]. These approaches are commonly trained using MSE or variants for probabilistic forecasts, e.g., quantile loss [27], and consequently often lose the ability to represent sharp predictions, as shown in Figure 1(c) for [63]. These generative models also lack an explicit structure to control the type of diversity in the latent space.

In this work, we introduce a model for including Shape and Time diversity in Probabilistic Forecasting (STRIPE). As shown in Figure 1(d), this enables to produce sharp and diverse forecasts, which fit well the ground truth distribution of trajectories in Figure 1(a).

STRIPE presented in section 3 is agnostic to the predictive model, and we use both deterministic or generative models in our experiments. STRIPE encompasses the following contributions. Firstly, we introduce a structured shape and temporal diversity mechanism based on determinantal point processes (DPP). We introduce two DPP kernels for modeling diverse trajectories in terms of shape and time, which are both differentiable and proved to be positive semi-definite (section 3.1). To have an explicit control on the diversity structure, we also design an iterative sampling mechanism to disentangle shape and time representations in the latent space (section 3.2).

Experiments are conducted in section 4 on synthetic datasets to evaluate the ability of STRIPE to match the ground truth trajectory distribution. We show that STRIPE significantly outperforms baseline methods for representing diversity, while maintaining the accuracy of the forecasting model. Experiments on real datasets further show that STRIPE is able to outperform state-of-the-art probabilistic forecasting approaches when evaluating the best sample (i.e., diversity), while being equivalent based on its mean prediction (i.e., quality).

2 Related work

Deterministic time series forecasting

Traditional time series forecasting methods, including linear autoregressive models such as ARIMA [5] or exponential smoothing [26], handle linear dynamics and stationary time series (or made stationary by modeling trends and seasonality). Deep learning has become the state-of-the-art for automatically modeling complex long-term dependencies, with many works focusing on architecture design based on temporal convolution networks [4, 52], recurrent neural networks (RNNs) [30, 62, 43], or Transformer [55, 36]. Another crucial topic more recently studied in the non-stationary context is the choice of a suitable loss function. As an alternative to the mean squared error (MSE) largely used as a proxy, new differentiable loss functions were proposed to enforce more meaningful criteria such as shape and time [46, 11, 32, 56], e.g., soft-DTW based
on dynamic time warping [11] or the DILATE loss with a soft-DTW term for shape and a smooth
temporal distortion index (TDI) [19, 54] for accurate temporal localization. These works toward
sharper predictions were however only studied in the context of deterministic predictions and not for
multiple outcomes.

**Probabilistic forecasting** For describing the conditional distribution of future values given an
input sequence, a first class of deterministic methods add variance estimation with Monte Carlo
dropout [65, 31] or predict the quantiles of this distribution [59, 20, 58] by minimizing the pinball
loss [27, 48] or the continuous ranked probability score (CRPS) [22]. Other probabilistic methods
try to approximate the predictive distribution, *explicitly* with a parametric distribution (e.g. Gaussian
for DeepAR [51] and variants [44, 50]), or *implicitly* with a generative model with latent variables
(e.g. with conditional variational autoencoders (cVAEs) [63], conditional generative adversarial
networks (cGANs) [28], normalizing flows [45]). However, these methods lack the ability to produce
sharp forecasts by minimizing variants of the MSE (pinball loss, gaussian maximum likelihood),
at the exception of cGANs - but which suffer from mode collapse that limits predictive diversity.
Moreover, these generative models are generally represented by unstructured distributions in the
latent space (e.g. Gaussian), which do not allow to have an explicit control on the targeted diversity.

**Diverse predictions** For improving the diversity of predictions, several repulsive schemes were
studied such as the variety loss [25, 53] that consists in optimizing the best sample, or entropy
regularization terms [12, 57] that encourage a uniform distribution and thus more diverse samples.
Submodular distribution functions such as determinantal point processes (DPP) [29, 47, 39] are an
appealing probabilistic tool to enforce structured diversity via the choice of a positive semi-definite
kernel. DPPs has been successfully applied in various contexts, e.g. document summarization [23],
recommendation systems [21], object detection [2], and very recently to image generation [16] and
diverse trajectory forecasting [63]. GDPP [16] is based on matching generated and true sample
diversity by aligning the corresponding DPP kernels, and thus limits their use in datasets where the
full distribution of possible outcomes is accessible. In contrast, our approach is applicable in realistic
scenarii where only a single label is available for each training sample. Although we share with [63]
the goal to use DPP as diversification mechanism, the main limitation in [63] is to use the MSE loss
for training the prediction and diversification models, leading to blurred prediction, as illustrated in
Figure 1(c). Our approach is able to generate sharp and diverse predictions ; we also highlight the
importance in STRIPE to use different criteria for training the prediction model (quality) and the
diversification mechanism in order to make them cooperate.

### 3 Shape and time diversity for probabilistic time series forecasting

We introduce the STRIPE model for including shape and time diversity for probabilistic time series
forecasting, which is depicted in Figure 2. Given an input sequence \( x_{1:T} = (x_1, \ldots, x_T) \in \mathbb{R}^{p \times T} \), our
goal is to sample a set of \( N \) diverse and plausible future trajectories \( \hat{y}^{(i)} = (\hat{y}_{T+1}^{(i)}, \ldots, \hat{y}_{T+\tau}^{(i)}) \in \mathbb{R}^{d \times \tau} \from
the data future distribution \( \hat{y}^{(i)} \sim p(\cdot; x_{1:T}) \).

STRIPE builds upon a general Sequence To Sequence (Seq2Seq) architecture dedicated to multi-step
time series forecasting: the input time series \( x_{1:T} \) is fed into an encoder that summarizes the input
into a latent vector \( h \). Note that our method is agnostic to the specific choice of the forecasting model:
it can be a deterministic RNN, or a probabilistic conditional generative model (e.g. cVAE [63], cGAN
[28], normalizing flow [45]).

For training the predictor (upper part in Figure 2), we concatenate \( h \) with a vector \( \theta_k \in \mathbb{R}^k \) (free
space left for the diversifying variables) and a decoder produces a forecasted trajectory \( \hat{y}^{(0)} =
(\hat{y}_{T+1}^{(0)}, \ldots, \hat{y}_{T+\tau}^{(0)}) \). The predictor minimizes a quality loss \( \mathcal{L}_{\text{quality}}(\hat{y}^{(0)}, y^{(0)}) \) between the predicted
\( \hat{y}^{(0)} \) and ground truth future trajectory \( y^{(0)} \). In our non-stationary context, we train the STRIPE
predictor with \( \mathcal{L}_{\text{quality}} \) based on the recently proposed DILATE loss [32], that has proven successful
for enforcing sharp predictions with accurate temporal localization.

For introducing structured diversity (lower part in Figure 2), we concatenate \( h \) with diversifying
latent variables \( z \in \mathbb{R}^k \) and produce \( N \) future trajectories \( \{\hat{y}^{(i)}\}_{i=1,\ldots,N} \). Our key idea is to augment
\( \mathcal{L}_{\text{quality}}(\cdot) \) with a diversification loss \( \mathcal{L}_{\text{diversity}}(\cdot; \mathcal{K}) \) parameterized by diversity kernel \( \mathcal{K} \) and
Figure 2: Our STRIPE model builds upon a Seq2Seq architecture trained with a quality loss $L_{\text{quality}}$ enforcing sharp predictions. Our contributions rely on the design of a diversity loss $L_{\text{diversity}}$ based on a specific Determinantal Point Processes (DPP). We design admissible shape and time DPP kernels, i.e. positive semidefinite, and differentiable for end-to-end training with deep models (section 3.1). We also introduce an iterative DDP sampling mechanism to generate disentangled latent codes between shape and time, supporting the use of different criteria for diversity and quality (section 3.2).

balanced by hyperparameter $\lambda \in \mathbb{R}$, leading to the overall objective training function:

$$L_{\text{STRIPE}}(\hat{y}^{(0)}, ..., \hat{y}^{(N)}; K) = L_{\text{quality}}(\hat{y}^{(0)}, y^{(0)}) + \lambda L_{\text{diversity}}(\hat{y}^{(1)}, ..., \hat{y}^{(N)}; K) \quad (1)$$

We highlight that STRIPE is applicable with a single target trajectory $y^{(0)}$, i.e. we do not require the full trajectory distribution. We now detail how the $L_{\text{diversity}}(\cdot; K)$ loss is designed to ensure diverse shape and time predictions.

### 3.1 STRIPE diversity module based on determinantal point processes

Our $L_{\text{diversity}}$ loss relies on determinantal point processes (DPP) that are a convenient probabilistic tool for enforcing structured diversity via adequately chosen positive semi-definite kernels. For comparing two time series $y_1$ and $y_2$, we introduce the two following kernels $K_{\text{shape}}$ and $K_{\text{time}}$, for finely controlling the shape and temporal diversity:

$$K_{\text{shape}}(y_1, y_2) = e^{-DTW_{\gamma}(y_1, y_2)} \quad (2)$$

$$K_{\text{time}}(y_1, y_2) = TDI_{\gamma}(y_1, y_2) = \frac{1}{Z} \sum_{A \in \mathcal{A}_{\tau,\tau}} \langle A, \Omega \rangle \exp \left(-\frac{\langle A \Delta(y_1, y_2) \rangle}{\gamma} \right) \quad (3)$$

where $DTW_{\gamma}(y_1, y_2) := -\gamma \log \left( \sum_{A \in \mathcal{A}_{\tau,\tau}} \exp \left(-\frac{\langle A \Delta(y_1, y_2) \rangle}{\gamma} \right) \right)$ is a smooth relaxation of Dynamic Time Warping (DTW) [11], and $K_{\text{time}}$ corresponds to a smooth Temporal Distortion Index (TDI) [19, 32]: $\gamma > 0$ denotes the smoothing coefficient, $A \subset \{0, 1\}^{\tau \times \tau}$ is a warping path between two time series of length $\tau$, $\mathcal{A}_{\tau,\tau}$ the set of all feasible warping paths and $\Delta(y_1, y_2) = [\delta((y_{1,i}, (y_{2,j})))_{1 \leq i, j \leq \tau}]$ is a pairwise cost matrix between time steps of both series with similarity measure $\delta : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$, $\Omega$ is a $\tau \times \tau$ matrix penalizing the deviation of warping paths from the main diagonal and $Z$ is the partition function. These kernels are derived from the two components of the DILATE loss [32]; however in contrast to the deterministic nature of DILATE, they are used in a probabilistic context for producing sharp and diverse forecasts.

$K_{\text{shape}}$ and $K_{\text{time}}$ are differentiable by design\(^1\), making them suitable for end-to-end training with back-propagation. We also derive the key following result for ensuring the submodularity properties of DPPs, that we prove in supplementary 1:

---

\(^1\)In the limit case $\gamma \rightarrow 0$, $DTW_{\gamma}$ (resp. TDI) recovers the standard DTW (resp. TDI).
Figure 3: At test time, STRIPE sequential shape and time sampling scheme that leverages the disentangled latent space. STRIPE-shape first proposes diverse shape latent variables. For each generated shape, STRIPE-time further enhances its temporal variability, leading to a final set of accurate predictions with shape and time diversity.

**Proposition 1.** Providing that $\kappa$ is a positive semi-definite (PSD) kernel such that $\kappa/1+\kappa$ is also PSD, if we define the cost matrix $\Delta$ with general term $\delta(y_i,y_j) = -\gamma \log \kappa(y_i,y_j)$, then $K_{\text{shape}}$ and $K_{\text{time}}$ defined respectively in Equations (2) and (3) are PSD kernels.

In practice, we choose $\kappa(u,v) = \frac{1}{2}e^{-\frac{(u-v)^2}{\sigma^2}}(1 - \frac{1}{2}e^{-\frac{(u-v)^2}{\sigma^2}})^{-1}$ that fulfills Prop 1 requirements.

**DPP diversity loss** We combine the two differentiable PSD kernels $K_{\text{shape}}$ and $K_{\text{time}}$ with the DPP diversity loss from [63] defined as the negative expected cardinality of a random subset $Y$ (of a ground set $\mathcal{Y}$ of $N$ items) sampled from the DPP of kernel $K$ (denoted as $K$ in matrix form of shape $N \times N$). This loss is differentiable and can be efficiently computed in closed-form:

$$L_{\text{diversity}}(Y;K) = -E_{Y \sim \text{DPP}(K)}|Y| = -\text{Trace}(I - (K + I)^{-1})$$

Intuitively, a larger expected cardinality means a more diverse sampled set according to kernel $K$. We provide more details on DPPs and the derivation of $L_{\text{diversity}}$ in supplementary 2.

### 3.2 STRIPE learning and sequential shape and temporal diversity sampling

To maximize shape and time diversity with Eq (1) and (4), a naive way is to consider the combined kernel $K_{\text{shape}} + K_{\text{time}}$ which is also PSD. However, this reduces to using the same criterion for quality and diversity, i.e. DILATE [32]. This directly makes $L_{\text{diversity}}$ conflicts with $L_{\text{quality}}$ and harms prediction performances, as shown in ablation studies (section 4.2). Another simple solution is to diversify using $K_{\text{shape}}$ and $K_{\text{time}}$ independently, which prevents from modeling joint shape and time variations, and intrinsically limits the expressiveness of the diversification scheme. In contrast, we propose a sequential shape and temporal diversity sampling scheme, which enables to jointly model variations in shape and time without altering prediction quality. We now detail how the STRIPE models are trained and then used at test time.

**STRIPE-shape and STRIPE-time learning** We start by independently training two proposal modules STRIPE-shape and STRIPE-time (and their respective encoders and decoders) by optimizing Eq (1) with $L_{\text{STRIPE}}(\cdot;K_{\text{shape}})$ (resp. $L_{\text{STRIPE}}(\cdot;K_{\text{time}})$). To this end, we complement the latent state $h$ of the forecaster with a diversifying latent variable $z \in \mathbb{R}^k$ decomposed into shape $z_s \in \mathbb{R}^{k/2}$ and temporal $z_t \in \mathbb{R}^{k/2}$ components: $z = (z_s,z_t) \in \mathbb{R}^k$. As illustrated in Figure 3, STRIPE-shape (the description of STRIPE-time is symmetric) is a proposal neural network that produces $N_s$ different shape latent codes $z_s^{(i)}$ (the output of the STRIPE-shape neural network is of shape $N_s \times k$). The decoder takes the concatenated state $(h,z_s^{(i)},z_t)$ for a fixed $z_t$ and produces $N_s$ future trajectories $\hat{y}^{(i)}$, whose diversity is maximized with $L_{\text{diversity}}(\hat{y}^{(1)},...\hat{y}^{(N_s)};K_{\text{shape}})$ in Eq (4).
Sequential sampling at test time Once the STRIPE-shape and STRIPE-time models (and their corresponding encoders and decoders) are learned, test-time sampling (illustrated in Figure 3 and detailed in Algorithm 1) consists in sequentially maximizing the shape diversity with STRIPE-shape (different guesses about the step amplitude in Figure 3) and the temporal diversity of each shape with STRIPE-time (the temporal localization of the step). Notice that the ordering shape+time is actually important since the notion of time diversity between two time series is only meaningful if they have a similar shape (so that computing the DTW optimal path has a sense).

As shown in our experiments, this two-steps scheme (denoted STRIPE S+T) leads to more diverse predictions with both shape and time criteria compared to using the shape or time kernels alone.

Algorithm 1: STRIPE S+T sampling at test time

<table>
<thead>
<tr>
<th>Sample an initial ( z_i^{(0)} \sim N(0, I) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( z_1^{(1)}, \ldots, z_{N_t}^{(1)} = )</td>
</tr>
<tr>
<td><strong>STRIPE-shape</strong>(( x_{1:T}, z_i^{(0)} ))</td>
</tr>
<tr>
<td><strong>for</strong> ( i = 1..N_s ) <strong>do</strong></td>
</tr>
<tr>
<td>( z_i^{(1)}, \ldots, z_i^{(i,N_t)} = )</td>
</tr>
<tr>
<td><strong>STRIPE-time</strong>(( x_{1:T}, z_i^{(i)} ))</td>
</tr>
<tr>
<td><strong>for</strong> ( j = 1..N_t ) <strong>do</strong></td>
</tr>
<tr>
<td>( \hat{y}_{T+1:t+\tau} = )</td>
</tr>
<tr>
<td><strong>Decoder</strong>(( x_{1:T}, (z_i^{(i)}, z_i^{(i,j)}) ))</td>
</tr>
<tr>
<td><strong>end</strong></td>
</tr>
<tr>
<td><strong>end</strong></td>
</tr>
</tbody>
</table>

4 Experiments

To illustrate the relevance of STRIPE, we carry out experiments in two different settings: in the first one, we compare the ability of forecasting methods to capture the full predictive distribution of future trajectories on a synthetic dataset with multiple possible futures for each input. To validate our approach in realistic settings, we evaluate STRIPE on 2 standard real datasets (traffic & electricity) where we evaluate the best (resp. the mean) sample metrics as a proxy for diversity (resp. quality).

Implementation details: To handle the inherent ambiguity of the synthetic dataset (multiple targets for one input), our STRIPE model is based on a natively stochastic model (cVAE). Since this situation does not arise exactly for real-world datasets, we choose in this case a deterministic Seq2Seq predictor with 1 layer of 128 Gated Recurrent Units (GRU) [9]. In our experiments, all methods produce \( N=10 \) future trajectories that are compared to the unique (or multiple) ground truth(s). For a fair comparison, STRIPE S+T generates \( N_s \times N_t = 10 \times 10 = 100 \) predictions and we randomly sample \( N=10 \) predictions for evaluation. Further neural network architectures and implementation details are described in supplementary 3.1. Our PyTorch code implementing STRIPE is available at https://github.com/vincent-leguen/STRIPE.

4.1 Synthetic dataset with multiple futures

We use a synthetic dataset similar to [32] that consists in predicting step functions based on a two-peaks input signal (see Figure 1). For each input series of 20 timesteps, we generate 10 different future series of length 20 by adding noise on the step amplitude and localisation. The dataset is composed of \( 100 \times 10 = 1000 \) time series for each train/valid/test split (further dataset description in supplementary 3.1).

Metrics: In this multiple futures context, we define two specific discrepancy measures \( H_{\text{quality}}(\ell) \) and \( H_{\text{diversity}}(\ell) \) for assessing the divergence between the predicted and true distributions of futures trajectories for a given loss \( \ell \) (\( \ell = \text{MSE} \) or \( \text{DILATE} \) in our experiments):

\[
H_{\text{quality}}(\ell) = \mathbb{E}_{\mathbf{x} \in \mathcal{D}_{\text{test}}} \mathbb{E}_{y} \left[ \inf_{\hat{y} \in F(\mathbf{x})} \ell(\hat{y}, y) \right] \quad H_{\text{diversity}}(\ell) = \mathbb{E}_{\mathbf{x} \in \mathcal{D}_{\text{test}}} \mathbb{E}_{y \in F(\mathbf{x})} \left[ \inf_{\hat{y}} \ell(\hat{y}, y) \right]
\]

\( H_{\text{quality}} \) penalizes forecasts \( \hat{y} \) that are far away from a ground truth future of \( \mathbf{x} \) denoted \( y \in F(\mathbf{x}) \) (similarly to the precision concept in pattern recognition) whereas \( H_{\text{diversity}} \) penalizes when a true future is not covered by a forecast (similarly to recall). We also use the continuous ranked probability score (CRPS)\(^2\) which is a standard proper scoring rule [22] for assessing probabilistic forecasts [20].

---

\(^2\)An intuitive definition of the CRPS is the pinball loss integrated over all quantile levels. The CRPS is minimized when the predicted future distribution is identical to the true future distribution.
Table 1: Forecasting results on the synthetic dataset with multiple futures for each input, averaged over 5 runs (mean ± standard deviation). Best equivalent method(s) (Student t-test) shown in bold. Metrics are scaled (MSE × 1000, DILATE × 100, CRPS × 1000) for readability.

<table>
<thead>
<tr>
<th>Methods</th>
<th>H_{quality} (\downarrow)</th>
<th>H_{diversity} (\downarrow)</th>
<th>CRPS (\downarrow)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MSE DILATE</td>
<td>MSE DILATE</td>
<td></td>
</tr>
<tr>
<td>Deep AR [51]</td>
<td>26.6 ± 6.4</td>
<td>67.0 ± 12.0</td>
<td>45.4 ± 4.3</td>
</tr>
<tr>
<td>cVAE MSE</td>
<td>11.8 ± 0.5</td>
<td>48.8 ± 3.2</td>
<td>85.4 ± 7.0</td>
</tr>
<tr>
<td>variety loss [53]</td>
<td>13.1 ± 2.7</td>
<td>50.9 ± 4.7</td>
<td>19.6 ± 1.1</td>
</tr>
<tr>
<td>Entropy regul. [12]</td>
<td>12.7 ± 0.7</td>
<td>51.5 ± 2.9</td>
<td>19.7 ± 0.7</td>
</tr>
<tr>
<td>Diverse DPP [63]</td>
<td>15.9 ± 2.6</td>
<td>56.6 ± 2.8</td>
<td>16.5 ± 1.5</td>
</tr>
<tr>
<td>GDPP [16] kernel MSE</td>
<td>11.7 ± 1.3</td>
<td>47.5 ± 3.1</td>
<td>19.5 ± 0.4</td>
</tr>
<tr>
<td>STRIPE S+T (ours)</td>
<td>12.4 ± 1.0</td>
<td>48.7 ± 0.7</td>
<td>18.1 ± 1.6</td>
</tr>
<tr>
<td>cVAE DILATE</td>
<td>11.6 ± 1.8</td>
<td>28.3 ± 2.9</td>
<td>22.2 ± 2.5</td>
</tr>
<tr>
<td>variety loss [53]</td>
<td>14.9 ± 3.3</td>
<td>33.5 ± 1.9</td>
<td>23.8 ± 3.9</td>
</tr>
<tr>
<td>Entropy regul. [12]</td>
<td>12.7 ± 0.6</td>
<td>29.9 ± 3.2</td>
<td>23.5 ± 2.6</td>
</tr>
<tr>
<td>Diverse DPP [63]</td>
<td>11.1 ± 1.6</td>
<td>30.2 ± 2.9</td>
<td>20.7 ± 2.3</td>
</tr>
<tr>
<td>GDPP [16] kernel DILATE</td>
<td>10.6 ± 1.6</td>
<td>28.7 ± 4.1</td>
<td>21.7 ± 2.1</td>
</tr>
<tr>
<td>STRIPE S+T (ours)</td>
<td>10.8 ± 0.4</td>
<td>30.7 ± 0.9</td>
<td>14.5 ± 0.6</td>
</tr>
</tbody>
</table>

**Results** We compare our method to 4 recent competing diversification mechanisms (variety loss [53], entropy regularisation [12], diverse DPP [63] and GDPP [16]) based two different forecasting backbones: a conditional variational autoencoder (cVAE) trained with MSE and with DILATE. Results in Table 1 show that our model STRIPE S+T based on a cVAE DILATE obtains the global best performances by improving the diversity by a large margin ($H_{diversity}(DILATE) = 35.5$ vs. 67.8), significantly outperforming other methods. This highlights the relevance of the structured shape and time diversity in STRIPE. It is worth mentioning that STRIPE also presents the best performances in quality. In contrast, other diversification mechanisms (variety loss, entropy regularisation, diverse DPP) based on the same backbone (cVAE DILATE) improve the diversity in DILATE but at the cost of a loss in quality in MSE and/or DILATE. Although GDPP does not deteriorate quality, it is significantly worse than STRIPE in diversity, and the approach requires full future distribution supervision, which it not applicable in real dataset (see section 2).

Similar conclusions can be drawn for the cVAE MSE backbone: the different diversity mechanisms improve the diversity but at the cost of a loss of quality. For example, Diverse DPP MSE [63] improves diversity ($H_{diversity}(DILATE) = 59.6$ vs. 85.4) but loose in quality ($H_{quality}(DILATE) = 56.6$ vs. 48.8). In contrast, STRIPE S+T again both improves diversity ($H_{diversity}(DILATE) = 62.0$ vs. 85.4) with equivalent quality ($H_{quality}(DILATE) = 48.7$ vs. 48.8). We further highlight that STRIPE S+T gets the best results evaluated in CRPS, confirming its ability to better recover the true future distribution.

4.2 Ablation study

To analyze the respective roles of the quality and diversity losses, we perform an ablation study on the synthetic dataset with the cVAE backbone trained with the quality loss DILATE and different DPP diversity losses. For a finer analysis, we report in Table 2 the shape (DTW) and time (TDI) components of the DILATE loss [32].

Table 2: Ablation study on the synthetic dataset. We train a backbone cVAE with the DILATE quality loss and compare different DPP kernels for diversity. Metrics are scaled for readability. Results averaged over 5 runs (mean ± std). Best equivalent method(s) (Student t-test) shown in bold.

<table>
<thead>
<tr>
<th>cVAE DILATE diversity</th>
<th>H_{quality} (\downarrow)</th>
<th>H_{diversity} (\downarrow)</th>
<th>CRPS (\downarrow)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MSE DILATE</td>
<td>MSE DILATE</td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>11.6 ± 1.8</td>
<td>28.3 ± 2.9</td>
<td></td>
</tr>
<tr>
<td>DILATE</td>
<td>11.1 ± 1.6</td>
<td>30.2 ± 2.8</td>
<td></td>
</tr>
<tr>
<td>MSE</td>
<td>10.9 ± 1.5</td>
<td>30.2 ± 2.9</td>
<td></td>
</tr>
<tr>
<td>shape (ours)</td>
<td>11.0 ± 1.4</td>
<td>30.2 ± 1.2</td>
<td></td>
</tr>
<tr>
<td>time (ours)</td>
<td>11.9 ± 0.5</td>
<td>31.2 ± 1.3</td>
<td></td>
</tr>
<tr>
<td>S+T (ours)</td>
<td>10.8 ± 0.4</td>
<td>30.7 ± 0.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>22.2 ± 2.5</td>
<td>18.5 ± 1.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>20.7 ± 2.3</td>
<td>18.6 ± 1.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>20.1 ± 2.2</td>
<td>18.5 ± 1.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>15.5 ± 1.04</td>
<td>16.4 ± 1.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>16.1 ± 0.70</td>
<td>17.6 ± 0.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>14.5 ± 0.6</td>
<td>16.1 ± 1.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>67.8 ± 7.8</td>
<td>62.0 ± 4.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>48.6 ± 2.2</td>
<td>62.6 ± 11.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>41.9 ± 8.8</td>
<td>61.7 ± 9.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>15.4 ± 4.2</td>
<td>37.8 ± 3.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>15.1 ± 3.1</td>
<td>38.9 ± 3.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>13.2 ± 1.7</td>
<td>35.5 ± 1.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>62.2 ± 4.2</td>
<td>60.7 ± 1.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>62.1 ± 0.9</td>
<td>63.2 ± 1.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>62.3 ± 1.4</td>
<td>60.5 ± 0.4</td>
<td></td>
</tr>
</tbody>
</table>
Figure 4: Qualitative predictions for Traffic and Electricity datasets. Input series in blue are not shown entirely for readability. We display 10 future predictions of STRIPE S+T that are both sharp and accurate compared to the ground truth (GT) future in green.

Results presented in Table 2 first reveal the crucial importance to define different criteria for quality and diversity. With the same loss for quality and diversity (as this is the case in [63]), we observe here that the DILATE DPP kernel does not bring a statistically significant diversity gain compared to the cVAE DILATE baseline (without diversity loss). By choosing the MSE kernel instead, we even get a small diversity and quality improvement.

In contrast, our introduced shape and time kernels $K_{\text{shape}}$ and $K_{\text{time}}$ largely improve the diversity in DILATE without deteriorating precision. As expected, each kernel brings its own benefits: $K_{\text{shape}}$ brings the best improvement in the shape metric DTW ($H_{\text{diversity}}(\text{DTW}) = 16.4$ vs. $18.8$) and $K_{\text{shape}}$ the best improvement in the time metric TDI ($H_{\text{diversity}}(\text{TDI}) = 15.1$ vs. $48.6$). With our sequential shape and time sampling scheme described in section 3.2, STRIPE S+T gathers the benefits of both criteria and gets the global best results in diversity and equivalent results in quality.

4.3 State-of-the-art comparison on real-world datasets

We evaluate here the performances of STRIPE on two challenging real-world datasets commonly used as benchmarks in the time series forecasting literature [61, 51, 30, 44, 32, 52]: Traffic: consisting in hourly road occupancy rates (between 0 and 1) from the California Department of Transportation, and Electricity: consisting in hourly electricity consumption measurements (kWh) from 370 customers. For both datasets, models predict the 24 future points given the past 168 points (past week). Although these datasets present daily, weakly, yearly periodic patterns, we are more interested here in modeling finer intraday temporal scales, where these signals present sharp fluctuations that are crucial for many applications, e.g. short-term renewable energy forecasts for load adjustment in smart-grids [33].

Contrary to the synthetic dataset, we only dispose of one future trajectory sample $y_{T+1:T+\tau}$ for each input series $x_{1:T}$. In this case, the metrics $H_{\text{quality}}$ (resp. $H_{\text{diversity}}$) defined in section 4.1 reduce to the mean sample (resp. best sample), which are common for evaluating stochastic forecasting models [3, 18]. We also report the CRPS in supplementary 3.2.

Results in Table 3 reveal that STRIPE S+T outperforms all other methods in terms of the best sample trajectory evaluated in MSE and DILATE for both datasets, while being equivalent in the mean sample in 3/4 cases. Interestingly, STRIPE S+T provides better best trajectories (evaluated in MSE Table 3: Forecasting results on the Traffic and Electricity datasets, averaged over 5 runs (mean ± std). Metrics are scaled for readability. Best equivalent method(s) (Student t-test) shown in bold.

<table>
<thead>
<tr>
<th>Method</th>
<th>MSE</th>
<th>Traffic</th>
<th>DILATE</th>
<th>Electricity</th>
<th>DILATE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nbeats [41] MSE</td>
<td>-</td>
<td>7.8 ± 0.3</td>
<td>-</td>
<td>22.1 ± 0.8</td>
<td>-</td>
</tr>
<tr>
<td>Nbeats [41] DILATE</td>
<td>-</td>
<td>17.1 ± 0.8</td>
<td>-</td>
<td>17.8 ± 0.3</td>
<td>-</td>
</tr>
<tr>
<td>Deep AR [51]</td>
<td>15.1 ± 1.7</td>
<td>6.6 ± 0.7</td>
<td>30.3 ± 1.9</td>
<td>16.9 ± 0.6</td>
<td>67.6 ± 5.1</td>
</tr>
<tr>
<td>cVAE DILATE</td>
<td>10.0 ± 1.7</td>
<td>8.8 ± 1.6</td>
<td>19.1 ± 1.2</td>
<td>17.0 ± 1.1</td>
<td>28.9 ± 0.8</td>
</tr>
<tr>
<td>Variety loss [53]</td>
<td>9.8 ± 0.8</td>
<td>7.9 ± 0.8</td>
<td>18.9 ± 1.4</td>
<td>15.9 ± 1.2</td>
<td>29.4 ± 1.0</td>
</tr>
<tr>
<td>Entropy regul. [12]</td>
<td>11.4 ± 1.3</td>
<td>10.3 ± 1.4</td>
<td>19.1 ± 1.4</td>
<td>16.8 ± 1.3</td>
<td>34.4 ± 4.1</td>
</tr>
<tr>
<td>Diverse DPP [63]</td>
<td>11.2 ± 1.8</td>
<td>6.9 ± 1.0</td>
<td>20.5 ± 1.0</td>
<td>14.7 ± 1.0</td>
<td>31.5 ± 0.8</td>
</tr>
<tr>
<td>STRIPE S+T</td>
<td>10.1 ± 0.4</td>
<td>6.5 ± 0.2</td>
<td>19.2 ± 0.8</td>
<td>14.2 ± 0.2</td>
<td>29.7 ± 0.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>23.4 ± 0.2</td>
<td>24.4 ± 0.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>16.9 ± 0.2</td>
<td></td>
</tr>
</tbody>
</table>
and DILATE) than the recent state-of-the-art N-Beats algorithm [41] (either trained with MSE or DILATE), which is dedicated to producing high quality deterministic forecasts. This confirms that STRIPE’s structured quality and diversity framework enables to obtain very accurate best predictions. Finally when compared to the state-of-the art probabilistic deep AR method [51], STRIPE S+T is consistently better in diversity and quality.

We display a few qualitative forecasting examples of STRIPE S+T on Figure 4 and additional ones in supplementary 3.3. We observe that STRIPE predictions are both sharp and accurate: both the shape diversity (amplitude of the peaks) and temporal diversity match the ground truth future.

4.4 Model analysis

![Figure 5: Influence of the number N of trajectories on quality (higher is better) and diversity for the synthetic dataset.](image)

![Figure 6: Scatterplot of 50 predictions in the plane (DTW,TDI), comparing STRIPE S+T v.s. Diverse DPP DILATE [63].](image)

We analyze in Figure 5 for the synthetic dataset the evolution of performances when increasing the number $N$ of sampled future trajectories from 5 to 100: we observe that this results in higher normalized DILATE diversity ($H_{diversity}(5)/H_{diversity}(N)$) for STRIPE S+T without deteriorating quality (which even increases slightly). In contrast, deepAR [51], which does not have control over the targeted diversity, increases diversity with $N$ but at the cost of a loss in quality. This again confirms the relevance of our approach that effectively combines an adequate quality loss function and a structured diversity mechanism.

We provide an additional analysis to highlight the importance to separate the criteria for enforcing quality and diversity. In Figure 6, we represent 50 predictions from the models Diverse DPP DILATE [63] and STRIPE S+T in the plane (DTW,TDI). Diverse DPP DILATE [63] uses a DPP diversity loss based on the DILATE kernel, which is the same than for quality. We clearly see that the two objectives conflict: this model increases the DILATE diversity (by increasing the variance in the shape (DTW) or the time TDI components) but a lot of these predictions have a high DILATE loss (worse quality). In contrast, STRIPE S+T predictions are diverse in DTW and TDI, and maintain an overall low DILATE loss. STRIPE S+T succeeds in recovering a set of good tradeoffs between shape and time leading a low DILATE loss.

5 Conclusion and perspectives

We present STRIPE, a probabilistic time series forecasting method that introduces structured shape and temporal diversity based on determinantal point processes. Diversity is controlled via two proposed differentiable positive semi-definite kernels for shape and time and exploits a forecasting model with a disentangled latent space. Experiments on synthetic and real-world datasets confirm that STRIPE leads to more diverse forecasts without sacrificing on quality. Ablation studies also reveal the crucial importance to decouple the criteria used for quality and diversity.

A future perspective would be to incorporate seasonality and extrinsic prior knowledge (such as special events) [31, 41] to better model the non-stationary abrupt changes and their impact on diversity and model confidence [10]. Other appealing directions include diversity-promoting forecasting for exploration in reinforcement learning [42, 17, 35], and extension of structured diversity to spatio-temporal or video prediction tasks [60, 18, 24].
**Broader Impact**

Probabilistic time series forecasting, especially in the non-stationary contexts, is a paramount research problem with immediate and large impacts in the society. A wide range of sensitive applications heavily rely on accurate forecasts of uncertain events with potentially sharp variations for making crucial decisions: in weather and climate science, better anticipating floods, hurricanes, earthquakes or other extreme events evolution could help taking emergency measures on time and save lives; in medicine, better predictions of an outbreak’s evolution is a particularly actual topic. We believe that introducing meaningful criteria such as shape and time, which are more related to application-specific evaluation metrics, is an important step toward more reliable and interpretable forecasts for decision makers.

**References**


