Evaluation of Trajectory Distribution Predictions with Energy Score

Novin Shahroudi [1] **Mihkel Lepson** [1] **Meelis Kull** [1]

Abstract

Predicting the future trajectory of surrounding objects is inherently uncertain and vital in the safe and reliable planning of autonomous systems such as in self-driving cars. Although trajectory prediction models have become increasingly sophisticated in dealing with the complexities of spatiotemporal data, the evaluation methods used to assess these models have not kept pace. "Minimum of N" is a common family of metrics used to assess the rich outputs of such models. We critically examine the Minimum of N within the proper scoring rules framework to show that it is not strictly proper and demonstrate how that could lead to a misleading assessment of multimodal trajectory predictions. As an alternative, we propose using Energy Score-based evaluation measures, leveraging their proven propriety for a more reliable evaluation of trajectory distribution predictions.

1. Introduction

Trajectory prediction is vital in various systems like autonomous vehicles. To plan safely, an autonomous agent has to forecast the future movements of surrounding objects amid inherent uncertainty. Traditional approaches (Sadeghian et al., 2019; Gao et al., 2020) generate single trajectory predictions, which are often deterministic point estimates, failing to represent the uncertainty. On the contrary, Multimodal Trajectory Prediction (MTP) or scenario forecasts (Morales et al., 2013) represent multiple plausible future trajectories that can naturally be viewed as a probability distribution over the forecast horizon, ideally capturing and matching the uncertainty in the data.

While modeling and representation of MTP have evolved (Gupta et al., 2018; Mangalam et al., 2020; Mohamed et al., 2020; Mangalam et al., 2021; Salzmann et al., 2020; Ma et al., 2021; Yue et al., 2022; Bae et al., 2022), evaluating such predictions has not kept pace. Minimum of N (MoN) is a common family of evaluation metrics inspired by variety loss (Gupta et al., 2018) is used to evaluate MTPs. As we demonstrate in our work, employing MoN is problematic as it can happen that a suboptimal prediction may get the same or lower error than the Bayes-optimal predictor, as shown in Fig. 1. Moreover, two models with comparable errors could have considerably different predictive distributions.

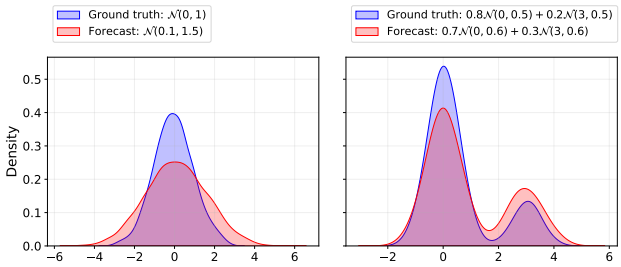


Figure 1. Gaussian and Mixture of Gaussian toy example in 1D that illustrates a suboptimal forecast (in red) performing better than the optimal forecast according to MoN despite a mismatch in mean, variance, and mixture weights. left: a unimodal example where the expected $minDE$ of optimal and suboptimal forecasts are 0.029 and 0.024, respectively. right: a bimodal example where similarly the expected $minDE$ is 0.024 and 0.021. Expected $minDE$ is calculated based on $K = 100$ trajectories, and $minDE$ is averaged across 2000 replicated experiments.

Similarly, (Thiede & Brahma, 2019) provides theoretical results that show MoN cannot be trusted as the only metric to compare models with, and they encourage the usage of the log-likelihood of the marginalized predictive distribution as a complementary metric. Despite that, since then, the literature has widely adopted MoN in various ways for the evaluation (Rudenko et al., 2020; Huang et al., 2023). In some instances, it is the main metric for evaluation. For example, state-of-the-art papers in the human trajectory prediction on the ETH/UCY dataset[1] are ranked primarily based on MoN. Many studies do accompany their evaluation with complementary metrics such as cross-entropy-

[1]Institute of Computer Science, University of Tartu, Tartu, Tartu County, Estonia. Correspondence to: Novin Shahroudi <novin.shahroudi@ut.ee>, Mihkel Lepson <mihkel.lepson@ut.ee>, Meelis Kull <meelis.kull@ut.ee>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

[1]https://paperswithcode.com/sota/trajectory-prediction-on-ethucy

based metrics (Rhinehart et al., 2018) or collision avoidance rate (Yue et al., 2022) and task-aware metrics (Ivanovic & Pavone, 2021). Complementing the evaluation with other metrics can be helpful if the new metric is able to identify the optimal prediction uniquely. However, using complementary metrics has its own challenges. For example, their use is often subjective to a particular downstream use case. Since many recent developments are based on some form of probabilistic modeling, it is prudent that we need evaluation metrics that are appropriate for assessing predicted probability distributions produced by these models.

We build on top of the theoretical results of (Thiede & Brahma, 2019) by studying MoN under the proper scoring rules framework (Gneiting & Raftery, 2007), focusing mainly on evaluating MTPs as probabilistic predictions. They show that the optimal solution of MoN loss is asymptotically a distribution with the PDF equal to the squared root of the ground truth's distribution when the number of trajectories is sufficiently large and propose a corrective procedure. Since the correction proposed by (Thiede & Brahma, 2019) does not eliminate the fact that MoN is not strictly proper, evaluating the predictive distribution of trajectories with MoN is problematic.

In summary, our contributions are threefold:

1. We critically examine MoN within the proper scoring rules framework, highlighting its lack of strict propriety and, therefore, being insufficient in assessing MTPs.
2. By that, we stress the need to introduce strictly proper scoring rules for evaluating trajectory predictions and propose the adoption of Energy Score-based metrics as an alternative to MoN for evaluating MTPs.
3. We show how Energy Score can be employed in various ways to evaluate MTPs for a more informed and reliable assessment.

The code for our experiments is available at https://github.com/novinsh/trajectory-prediction-eval-with-energy-score.

2. Related work

Evaluation metrics for MTP can be categorized into *Lower-bound-based*, *Probability-aware*, and *Distribution-aware* metrics (Huang et al., 2023). The common lower-bound-based instances of MoN are minimum displacement error (minDE) and minimum mean squared distance (minMSD). The first works to hint towards using such metrics for evaluation were (Alahi et al., 2016; Gupta et al., 2018; Sadeghian et al., 2019). Later, it got adopted and popularized more in other works (Lee et al., 2017; Rhinehart et al., 2018; Park et al., 2018; Bhattacharyya et al., 2018; Tang & Salakhutdinov, 2019; Schöller et al., 2020). Extension of lower-bound-based MoN to probability-aware variations can be achieved by applying MoN on *Most-likely* or Top-1-likely (Cui et al., 2019), $TopN\%$-likely or else *best-of-N* samples (Chen et al., 2021). $TopN\%$ is shown to be more stable than $Top1$ by (Bhattacharyya et al., 2020). Lower-bound metrics ignore the likelihood of the trajectories while implicitly being less sensitive to low-density regions – an indirect result of Theorem 1 from (Thiede & Brahma, 2019). Simultaneously, probability-aware instances of MoN assume that the ground-truth samples are realized from the highest probability regions of the ground truth's PDF.

Contrary to the previous categories, "distribution-aware" metrics are meant to evaluate the whole distribution. However, there is no agreed or formal notion of it. (Amirian et al., 2019) creates a synthetic setup where multiple observations are available from the ground truth, and it is possible to use Earth Mover's Distance (EMD) to evaluate models in a controlled environment. Others introduced Coverage-based metrics (Dendorfer et al., 2021; Li et al., 2022) that evaluate trajectories based on how close they fall from the ground truth as a measure to determine whether a sample is in- or out-of-distribution. They still employ MoN as part of their coverage-based metric. Lastly, the log-likelihood metric can be considered distribution-aware, which has been employed in many works in one way or the other (Rhinehart et al., 2018; Thiede & Brahma, 2019). However, estimating the likelihood is often challenging because existing models typically do not directly model the density, which means the likelihood function has to be estimated, which is often computationally expensive. Moreover, (Thiede & Brahma, 2019) in their Algorithm 1 rely on a marginalized log-likelihood metric to compensate for MoN's dilation effect. The marginalization in their work is due to computational considerations. More importantly, using *log-likelihood* can be seen in a broader view in that to employ *log score* as a strictly proper multivariate scoring rule. So, we extend their suggestion to use complementary metrics by saying that using complementary metrics is particularly beneficial if they are strictly proper. Also, our proposal to use Energy score-based evaluation allows for joint evaluation that might be relevant to many applications.

Many of the criticisms towards MoN in the literature, such as "only evaluating the best/most-likely sample" (Rhinehart et al., 2018), "robustness" (Bhattacharyya et al., 2020), "preferring higher entropy prediction" (Schöller et al., 2020) can be explained and studied more formally under the scoring rules framework. Empirically and analytically, we show why MoN is not desirable as an evaluation metric to assess the predictive distribution of trajectories or to be used as a loss function to issue probabilistic trajectory predictions, especially in safety-critical applications. However, its simplicity and lack of computational demand might be

attractive for certain resource-constrained settings, given that its lack of propriety is addressed, i.e., by compensating similar to (Thiede & Brahma, 2019) or ensuring it is not detrimental to the downstream use case.

3. Background

3.1. Definitions and notations

Trajectory A *trajectory* consists of a sequence of S-dimensional vectors where $S \in \{1, 2, 3, \dots\}$, encoding the coordinates of an agent throughout time. Let us consider trajectories of N agents and denote these trajectories as $\mathbf{y}_i$ where i is the index of the agent. The sequence can be divided into past $\mathbf{y}_i^{\text{past}}$ and future $\mathbf{y}_i^{\tau}$, where $\mathbf{y}_i^{\text{past}} = (y_i^{-T_{obs}+1}, \dots, y_i^0)$ is the observed trajectory with T_{obs} time steps, and $\mathbf{y}_i = (y_i^1, \dots, y_i^T)$ is the future trajectory with T time steps. In our work, typically $y_i^t \in \mathbb{R}^2$ because we consider $S = 2$. A trajectory can be seen as a multivariate time series, which, in our case, consists of two time series coupled based on a spatial dependency. In order to model uncertainties, we introduce spatiotemporal random vectors $\mathbf{Y}_i = (Y_i^1, \dots, Y_i^T)$ for $i = 1, \dots, N$ and assume that the future trajectory $\mathbf{y}_i$ of the i-th agent is a realization of this random vector $\mathbf{Y}_i$.

Trajectory Distribution Prediction In this work, we consider *trajectory distribution prediction* methods that for a given observed trajectory $\mathbf{y}_i^{\text{past}}$ output K future trajectories $\mathbf{x}_{i,1} \dots, \mathbf{x}_{i,K}$ drawn from some explicit or implicit[2] probability distribution $\mathbf{F}_{\mathbf{X}_i}$. Note that each predicted trajectory (also referred to as scenario) $\mathbf{x}_{i,k} \in \mathbb{R}^{T \times S}$ covers T timepoints and S spatial dimensions, and the point at time t is denoted as $x_{i,k}^t \in \mathbb{R}^S$.

Trajectory Distribution Evaluation *Trajectory distribution evaluation* is the process of assessing how well trajectory distribution predictions $\mathbf{X}_i \sim \mathbf{F}_{\mathbf{X}_i}$ follow the ground truth $\mathbf{Y}_i \sim \mathbf{F}_{\mathbf{Y}_i}$. A distribution-aware evaluation aims at calculating distance $d(\mathbf{F}_{\mathbf{X}_i}, \mathbf{F}_{\mathbf{Y}_i})$, which is some measure of distance between the predictive and ground truth distributions that informs on the quality of the predicted trajectory distribution.

3.2. Proper Scoring Rules

Typically, we would like to compare our predictions to the true generative distribution, but we do know the generative distribution, so a *scoring rule* allows us to instead compare the predictive distribution against a sample drawn from the generative distribution. A *scoring rule* provides a summary measure for evaluating probabilistic predictions that reports an empirical mean over the samples drawn from the ground truth distribution. The score could be used to compare a predictive distribution $\mathbf{F}_{\mathbf{X}_i}$ from a model with some ground truth distribution $\mathbf{F}_{\mathbf{Y}_i}$. The score is minimal when the prediction reports the true set of probabilities, i.e., $\mathbf{F}_{\mathbf{X}_i} = \mathbf{F}_{\mathbf{Y}_i}$. A scoring rule is proper if it satisfies this property as defined in Definition 3.1. We provide this definition based on (Gneiting & Raftery, 2007).

Definition 3.1. A (negatively-oriented) strictly proper scoring rule $\mathbf{S}$ maps a probability distribution $\mathbf{F}_{\mathbf{X}}$ and an observation $\mathbf{y}$ to a real number, i.e., $\mathbf{S}(\mathbf{F}_{\mathbf{X}}, \mathbf{y}) \in \mathbb{R}$. The expected value of $\mathbf{S}(\mathbf{F}_{\mathbf{X}}, .)$ under $\mathbf{F}_{\mathbf{Y}}$, is written as $\mathbf{S}(\mathbf{F}_{\mathbf{X}}, \mathbf{F}_{\mathbf{Y}}) = \mathbb{E}_{\mathbf{y}}[\mathbf{S}(\mathbf{F}_{\mathbf{X}}, \mathbf{y})]$ where $\mathbf{y} \sim \mathbf{F}_{\mathbf{Y}}$. A scoring rule is proper if $\mathbf{S}(\mathbf{F}_{\mathbf{X}}, \mathbf{F}_{\mathbf{Y}}) \geq \mathbf{S}(\mathbf{F}_{\mathbf{Y}}, \mathbf{F}_{\mathbf{Y}})$ for all $\mathbf{F}_{\mathbf{X}}$ and $\mathbf{F}_{\mathbf{Y}}$, and strictly proper when the equality holds if and only if $\mathbf{F}_{\mathbf{X}} = \mathbf{F}_{\mathbf{Y}}$.

Because of this property, a strictly proper scoring rule is not only useful for evaluation but also for learning and optimization of a probabilistic model as it encourages the model towards optimal prediction, i.e., a Bayes-optimal prediction of $\mathbf{F}_{\mathbf{Y}_i}$ that yields the best score.

3.3. Energy Score

Energy score is a measure used for evaluation of multivariate forecasts and shown to be a generalization of Continuous Ranked Probability Score (CRPS), which is a univariate score for evaluation of probabilistic forecasts (Gneiting & Raftery, 2007),

$$ES(\mathbf{F}_{\mathbf{X}}, \mathbf{y}) = \mathbb{E}\|\mathbf{X} - \mathbf{y}\|_p^\beta - \frac{1}{2}\mathbb{E}\|\mathbf{X} - \tilde{\mathbf{X}}\|_p^\beta$$

where $\|.\|_p$ indicates the *Lp-norm* (the original formulation has $p = 2$ but a generalization is provided in (Gneiting & Raftery, 2007)). Intuitively, the first term encourages the quality of individual predictions, while the second term encourages diversity. In expectation over many ground truth observations, the combination of these two terms achieves a balance that ensures strict propriety. Readers may refer to (Gneiting & Raftery, 2007; Székely & Rizzo, 2013) for more elaborate details. For $p = 2$, the norm is Euclidean, and the energy score is strictly proper for $0 < \beta < 2$ with smaller values suggested for heavy-tailed data to ensure detection of β moments. Furthermore, the energy Score is distantly related to permutational ANOVA (PANOVA) (Anderson, 2005). If $\beta = 2$ and $p = 2$, it will give us PANOVA with a minor difference that in PANOVA's original definition, the second term does not have the coefficient $1/2$.

[2] In general, the predictive distribution can be represented explicitly or implicitly. This choice has different computational and approximation implications depending on the modeling technique. For example, the sampling process itself could be costly, but this is not within the scope of our work. Regardless of the representation, in our work, the K trajectories are independent and identically distributed (i.i.d.) samples of the predictive distribution, which allows for a universal and non-parametric evaluation.

3.4. Displacement Error Metrics

Minimum Average Displacement Error (minADE) and Minimum Final Displacement Error (minFDE) defined in Eqs. 4 and 5, which are both based on Displacement Error (DE), are common instances of MoN.

$$DE(x, y) = \|x - y\|_2 \quad \forall x, y \in \mathbb{R}^S \tag{1}$$

$$ADE(\mathbf{X}_i, \mathbf{Y}_i) = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} DE(X_i^t, Y_i^t)\right] \tag{2}$$

$$FDE(\mathbf{X}_i, \mathbf{Y}_i) = \mathbb{E}\left[DE(X_i^T, Y_i^T)\right] \tag{3}$$

$$minADE(\mathbf{X}_i, \mathbf{Y}_i) = \mathbb{E}\left[\min_k \frac{1}{T}\sum_{t=1}^{T} DE(X_{i,k}^t, Y_i^t)\right] \tag{4}$$

$$minFDE(\mathbf{X}_i, \mathbf{Y}_i) = \mathbb{E}\left[\min_k DE(X_{i,k}^T, Y_i^T)\right] \tag{5}$$

where X_i^t and $X_{i,k}^t$ are random variables distributed identically to the timepoint t in $\mathbf{X}_i$. By using either $ADE(\mathbf{X}_i, \mathbf{Y}_i)$ or $ADE(\mathbf{F}_{\mathbf{X}_i}, \mathbf{F}_{\mathbf{Y}_i})$, we acknowledge that the scoring rule is based on the underlying distributions of $\mathbf{X}$ and $\mathbf{Y}$, whether they are expressed directly as random variables or through their cumulative distribution functions. The same applies to FDE, $minADE$, and $minFDE$. Here, we observe one sample $\mathbf{y}_i$ from the ground truth distribution, while observing K trajectories/samples $\mathbf{x}_{i,1} \dots, \mathbf{x}_{i,K}$ from the predictive distribution $\mathbf{F}_{\mathbf{X}_i}$. The above equations can be estimated as follows:

$$\widehat{ADE}(\mathbf{x}_i, \mathbf{y}_i) = \frac{1}{KT}\sum_{k=1}^{K}\sum_{t=1}^{T} DE(x_{i,k}^t, y_i^t)$$

$$\widehat{FDE}(\mathbf{x}_i, \mathbf{y}_i) = \frac{1}{K}\sum_{k=1}^{K} DE(x_{i,k}^T, y_i^T)$$

$$\widehat{minADE}(\mathbf{x}_i, \mathbf{y}_i) = \min_k \frac{1}{T}\sum_{t=1}^{T} DE(x_{i,k}^t, y_i^t)$$

$$\widehat{minFDE}(\mathbf{x}_i, \mathbf{y}_i) = \min_k DE(x_{i,k}^T, y_i^T)$$

We define the "L-lowest of N" (LoN) as a more general form of "Minimum of N".

$$ADE_{(L)}(\mathbf{X}_i, \mathbf{Y}_i) = \mathbb{E}\min_{\substack{\{k_1,\dots,k_L\} \\ k_i \neq k_j}} \frac{1}{LT}\sum_{l=1}^{L}\sum_{t=1}^{T} DE(X_{i,k_l}^t, Y_i^t) \tag{6}$$

$$FDE_{(L)}(\mathbf{X}_i, \mathbf{Y}_i) = \mathbb{E}\min_{\substack{\{k_1,\dots,k_L\} \\ k_i \neq k_j}} \frac{1}{L}\sum_{l=1}^{L} DE(X_{i,k_l}^T, Y_i^T) \tag{7}$$

It is noteworthy that $ADE_{(L=K)} \equiv ADE$ and similarly $FDE_{(L=K)} \equiv FDE$, at the same time, $ADE_{(L=1)} \equiv minADE$ and similarly $FDE_{(L=1)} \equiv minFDE$. In the remainder of the paper, we use these notations interchangeably.

4. MoN as a Scoring Rule

The following propositions study $ADE_{(L)}$ as an instance of MoN under the proper scoring rule framework and show that these measures are not strictly proper.

Proposition 4.1. *Average Displacement Error $ADE(\mathbf{X}_i, \mathbf{Y}_i)$ is improper, meaning there exist distributions $\mathbf{F}_{\mathbf{X}_i}$ and $\mathbf{F}_{\mathbf{Y}_i}$, for which*

$$ADE(\mathbf{X}_i, \mathbf{Y}_i) < ADE(\mathbf{Y}_i, \mathbf{Y}_i).$$

Proof. The proofs are provided in Appendix A. □

Because ADE is improper, FDE is improper too.

Proposition 4.2. *L-lowest Average Displacement Error $ADE_{(L)}(\mathbf{X}_i, \mathbf{Y}_i)$ is improper for any values of $L \leq 2$, meaning there exist distributions $\mathbf{F}_{\mathbf{X}_i}$ and $\mathbf{F}_{\mathbf{Y}_i}$, for which*

$$ADE_{(L)}(\mathbf{X}_i, \mathbf{Y}_i) < ADE_{(L)}(\mathbf{Y}_i, \mathbf{Y}_i).$$

Because $ADE_{(L)}$ is improper, $FDE_{(L)}$ is improper too.

Since ADE and $ADE_{(L)}$ are improper, they are unsuitable for identifying the optimal predictive distribution from the suboptimal one. We also empirically show these results in Section 6.1.

Another problem when using $ADE_{(L)}$ is that it depends on the number of trajectories K due to the minimum operation. This means that each extra trajectory we sample from the distribution only lowers the $ADE_{(L)}$ value, as we demonstrate in the following proposition.

Proposition 4.3. *Let $\mathbf{X}_i^{(K)} \sim \mathbf{F}_{\mathbf{X}_i}$ be a random vector of length K, $\mathbf{X}_i^{(K+1)} \sim \mathbf{F}_{\mathbf{X}_i}$ be a random vector of length $K+1$ and $\mathbf{Y}_i \sim \mathbf{F}_{\mathbf{Y}_i}$. Then*

$$ADE_{(L)}(\mathbf{X}_i^{(K)}, \mathbf{Y}_i) \geq ADE_{(L)}(\mathbf{X}_i^{(K+1)}, \mathbf{Y}_i)$$

We see that $ADE_{(L)}$ is decreasing as the sample size increases, and we also need to consider the fact that $ADE_{(L)}$ is bounded from below by zero. From this arises a question: For which distributions does it converge to zero? As it turns out, the $ADE_{(L)}$ converges to zero for all the distributions where the support covers the true distribution's support.

Proposition 4.4. *Let $\mathbf{X}_i \sim \mathbf{F}_{\mathbf{X}_i}$ of length K and $\mathbf{Y}_i \sim \mathbf{F}_{\mathbf{Y}_i}$. If $K \to \infty$, L is fixed and supp$(\mathbf{F}_{\mathbf{Y}_i}) \subset$ supp$(\mathbf{F}_{\mathbf{X}_i})$ then $ADE_{(L)}(\mathbf{X}_i, \mathbf{Y}_i) \to 0$.*

From the Proposition, we see that $ADE_{(L)}$ is a measure for which the value is not comparable across different values of K. As K increases, the value of $ADE_{(L)}$ decreases, even if the model does not improve, as we also empirically show in Section 6.2. From the interpretability point of view, this is somewhat unintuitive as one would expect that the scoring rule's report becomes more robust if more samples are available from the predictive distribution ($K \to \infty$).

Therefore, following Propositions 4.2 and 4.4, we suggest Energy Score as a strictly proper alternative for evaluating trajectory distribution predictions.

5. Distribution-aware Evaluation with Energy Score

Since trajectories are essentially a multivariate time series, we can adopt energy score as a multivariate scoring rule to evaluate the predicted probability distribution. 2D trajectories sampled from the predictive distribution are spatially coupled time series as depicted in Fig. 2. Therefore, the trajectories are multivariate in the temporal and spatial sense, which means that the interaction between spatiotemporal variables is of interest to be jointly assessed.

As far as we know, no other work has discussed the application of energy score in MTP evaluation. To this end, we introduce three variations of Energy Score for different use cases of evaluation in the context of trajectory prediction, which also applies to broader use cases in multivariate time series evaluation. The energy score, defined in Eq. 8, is applied to a probabilistic trajectory prediction with a predictive distribution $\mathbf{F}_{\mathbf{X}i}$ and a ground truth observation $\mathbf{y}_i$. The overall performance is then the average performance of all the instances in the dataset as shown in Eq. 9.

$$ES(\mathbf{F}_{\mathbf{X}i}, \mathbf{y}_i) = \overbrace{\mathbb{E}\|\mathbf{X}_i - \mathbf{y}_i\|_p^\beta}^{ED} - \frac{1}{2}\overbrace{\mathbb{E}\|\mathbf{X}_i - \tilde{\mathbf{X}}_i\|_p^\beta}^{EI} \quad (8)$$

$$\overline{ES} = \frac{1}{N}\sum_{i=1}^{N} ES(\mathbf{F}_{\mathbf{X}_i}, \mathbf{y}_i) \quad (9)$$

For brevity, hereafter, we will omit the subscript i when referring to each instance. The first term (ED) has a close similarity to the displacement error that measures the distance of the prediction with the observation. The second term (EI) is an intra-distance, which captures the uncertainty of $\mathbf{X}$ by calculating its dispersion. EI will be zero when the predictions are all the same, which is not attainable as long as the ground truth distribution has a variation. Intuitively, the ED term promotes each trajectory to be individually as good as possible, whereas the subtracted EI term promotes diversity. When averaged over many ground truth observations, the combination of these terms provides a balance under which the predictor is incentivized to match the ground truth distribution.

In a non-parametric case, the energy score operates on K uniform trajectories sampled from the predictive joint distribution $\mathbf{F}_{\mathbf{X}}$, whereas it can be calculated analytically in a parametric case. Since $\mathbf{X} \in \mathbb{R}^{K\times T\times S}$ and $\mathbf{y} \in \mathbb{R}^{1\times T\times S}$, there are many ways that distance between them can be calculated. Our variations outline three ways to calculate such a distance, and the difference is in how the Lp norm distance is calculated. For clarity, we provide the estimations for each term separately.

$$X_k^{t,s} = \begin{bmatrix} x_k^{11} & \dots & x_k^{1S} \\ \vdots & \ddots & \vdots \\ x_k^{T1} & \dots & x_k^{TS} \end{bmatrix} \qquad y^{t,s} = \begin{bmatrix} y^{11} & \dots & y^{1S} \\ \vdots & \ddots & \vdots \\ y^{T1} & \dots & y^{TS} \end{bmatrix}$$

Energy Score The distance between the matrices is an entry-wise matrix norm. With $p = 2$, it would be equal to Frobenius distance, and with $p > 2$, it would be equal to entry-wise Minkowski distance. This score evaluates the trajectories jointly on both temporal and spatial dimensions.

$$ES = \overbrace{\mathbb{E}_{k,t,s}\Big(\|\mathbf{X} - \mathbf{y}\|_p^\beta\Big)}^{ED} - \frac{1}{2}\overbrace{\mathbb{E}_{k,t,s}\Big(\|\mathbf{X} - \tilde{\mathbf{X}}\|_p^\beta\Big)}^{EI} \quad (10)$$

The Eq. 10 is estimated as follows:

$$\widehat{ED} = \frac{1}{K}\sum_{k=1}^{K}\Big(\sum_{t=1}^{T}\sum_{s=1}^{S}|X_k^{t,s} - y^{t,s}|^p\Big)^{\beta/p}$$

$$\widehat{EI} = \frac{1}{K^2}\sum_{k=1}^{K}\sum_{l=1}^{K}\Big(\sum_{t=1}^{T}\sum_{s=1}^{S}|x_k^{t,s} - \tilde{x}_l^{t,s}|^p\Big)^{\beta/p}$$

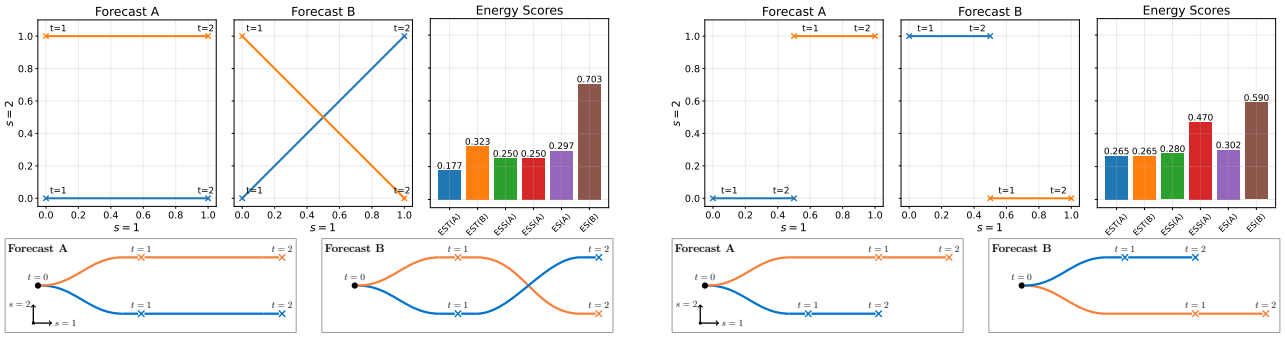
It is worth noting that the distance of each temporal and spatial variable is under the norm calculation.

Energy Score Temporal (EST) is a spatially marginalized energy score that calculates the distance column-wise, marginalizing the spatial dimension while calculating the distance jointly with respect to the temporal dimension. This variation is more sensitive to the temporal discrepancies. The distance calculation in this variation is akin to the Minkowski column distance between the two matrices.

$$EST = \overbrace{\mathbb{E}_{k,s}\Big(\|\mathbf{X} - \mathbf{y}\|_p^\beta\Big)}^{EDT} - \frac{1}{2}\overbrace{\mathbb{E}_{k,s}\Big(\|\mathbf{X} - \tilde{\mathbf{X}}\|_p^\beta\Big)}^{EIT} \quad (11)$$

$$\widehat{EDT} = \frac{1}{K}\sum_{k=1}^{K}\Big(\frac{1}{S}\sum_{s=1}^{S}\Big(\sum_{t=1}^{T}|x_k^{t,s} - y^{t,s}|^p\Big)^{\beta/p}\Big)$$

$$\widehat{EIT} = \frac{1}{K^2}\sum_{k=1}^{K}\sum_{l=1}^{K}\Big(\frac{1}{S}\sum_{s=1}^{S}\Big(\sum_{t=1}^{T}|x_k^{t,s} - \tilde{x}_l^{t,s}|^p\Big)^{\beta/p}\Big)$$



(a) (b)

Figure 2. **Top:** Toy examples of spatially and temporally indifferent cases respectively under (a) and (b) with their corresponding energy scores as bar plots. Forecasts A and B each issue two 2D scenarios (color-coded) for two time steps $t = \{1, 2\}$. **Bottom:** a comparable self-driving car demonstration for each case where the initial position of the target vehicle is marked by a solid black circle at $t = 0$.

Energy Score Spatial (ESS) is a temporally marginalized energy score that is used to calculate the distances row-wise, marginalizing over the temporal dimension while calculating the distance jointly with respect to the spatial dimension. This variation is more sensitive to the spatial discrepancies.

$$ESS = \overbrace{\mathbb{E}_{k,t}\Big(\|\mathbf{X} - \mathbf{y}\|_p^\beta\Big)}^{EDS} - \frac{1}{2}\overbrace{\mathbb{E}_{k,t}\Big(\|\mathbf{X} - \tilde{\mathbf{X}}\|_p^\beta\Big)}^{EIS} \quad (12)$$

$$\widehat{EDS} = \frac{1}{K}\sum_{k=1}^{K}\Big(\frac{1}{T}\sum_{t=1}^{T}\Big(\sum_{s=1}^{S}|x_k^{t,s} - y^{t,s}|^p\Big)^{\beta/p}\Big)$$

$$\widehat{EIS} = \frac{1}{K^2}\sum_{k=1}^{K}\sum_{l=1}^{K}\Big(\frac{1}{T}\sum_{t=1}^{T}\Big(\sum_{s=1}^{D}|x_k^{t,s} - \tilde{x}_l^{t,s}|^p\Big)^{\beta/p}\Big)$$

Figure 2 illustrates a toy example demonstrating the difference between the three variations. In this toy example, $T = 2$ and $S = 2$, with two forecasts, A and B, where forecast A matches the ground truth. Each forecast issues two equally likely trajectories/scenarios (color-coded in orange and blue) for the two time steps ahead in the two-dimensional space. In the first example (a), the forecasts are issued such that ESS would be indifferent between forecasts A and B while EST detects the difference, and vice-versa in the second example (b). ES detects the difference in both examples since it is calculated jointly over the temporal and spatial dimensions.

To illustrate the usefulness of EST and ESS, consider these examples to be about a self-driving car where the ego vehicle predicts the trajectory of a nearby car in a bird's-eye view or top-down view. In this context, case (a) exemplifies a lane-changing behavior for $t > 0$ where forecast A predicts no lane change while forecast B predicts a lane change. Despite the fact that both forecasts have similar spatial positions for each time step, that is $\{x_k^{t,s} | k \in \{1, 2\} \text{ from A}\} = \{x_k^{t,s} | k \in \{1, 2\} \text{ from B}\}$, they exhibit different temporal behavior. On the other hand, in case (b), forecasts A and B predict mirror images of each other with respect to axis $s = 2$, which exemplifies different longitudinal behavior. Despite the same temporal behavior, forecasts are different spatially. In this demonstration, if forecasts were governed by a constant velocity motion model $\boldsymbol{x} = \boldsymbol{v}t + \boldsymbol{x}^0$, then the differences between each scenario arise as a consequence of uncertainty about the initial position $\boldsymbol{x}^0$, the velocity $\boldsymbol{v}$, or both.

To emphasize, EST and ESS both assess temporal and spatial aspects. However, the joint information is lost due to marginalization over the spatial dimension for the EST and the temporal dimension for the ESS, making each insensitive to certain discrepancies. These variations of energy scores allow for various ways to diagnose forecasts. ES can be used as a primary evaluation metric and EST/ESS as a secondary measure to determine whether a forecast is better or worse than the other due to joint temporal or joint spatial information. Moreover, such marginalization allows for the different weighting of each dimension. For instance, in the context of self-driving cars, one can put more weight on $s = 1$ than $s = 2$ in the ESS, which emphasizes more on the longitudinal movement than latitudinal.

6. Experiments

We conduct a series of experiments highlighting the importance of adopting proper scoring rules as a distribution-aware evaluation of trajectory distribution predictions. We empirically demonstrate the consequence of adopting an improper score for evaluation and how it can go wrong in Section 6.1 through a showcase of propriety. We also empirically demonstrate the effect of the trajectory size K in Section 6.2. To see the energy score in action, we perform a real data experiment on the ETH/UCY dataset (Ess et al.,

2007) in Section 6.3. Additionally, we run a simulation to support our theoretical results in Appendix B.

6.1. Propriety Showcase

We generate trajectories through an autoregressive process with length $T = 4$ and set the second spatial dimension to all zeros. The generating process is effectively one-dimensional and can be summarized as follows:

$$y^t_{i,k} = y^{t-1}_{i,k} + \mathcal{N}(\mu^t + a^t, (\sigma^t + b^t)^2)$$
$$\text{where } t \in \{1,2,3\}, \quad i = [0, N), \quad k = [0, K), \quad y^0 = 0$$

a^t and b^t are deviations from the parameters of the normal distribution μ^t, and σ^t. This process mimics a motion dynamic where trajectories have varying velocities, as depicted in Fig. 3 with five trajectories generated from this process.

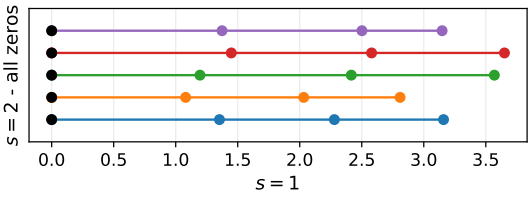


Figure 3. Generated trajectories with the same origin at $t = 0$ marked by a black dot. The second spatial dimension is set to zero for all, so they are plotted with a y-offset for visualization.

We set the ground truth parameters to be $\mu^t = 1$, $\sigma^t = 0.2$, $a^t = 0$, and $b^t = 0$ for $t = \{1, 2, 3\}$. Then, we generate $N = 5000$ observations and consider $K = \{10, 20, 50, 100, 300\}$ to generate predictions from the same process. We create sub-optimal predictions by creating deviations from the same data-generating process, setting $a^t = 0$ and using 21 equidistant values for b^t in the range $[-0.05, 0.05]$. We expect a metric to assign its lowest score to the prediction closest to the true value of the ground truth's parameter, that is, $b = 0$. For $|b| > 0$, we expect the score to reflect such discrepancy as predictions get further from the optimal. Figure 4 compares $minFDE$, $FDE_{(L)}$, and FDE, alongside with Final Energy Score (FES). Minimum of $minFDE$, $FDE_{(L)}$, and FDE are realized at a distribution with higher variance than the true distribution. In contrast, the FES is minimized near the true value, with some deviation for lower K due to noise. In the limit of the dataset size N and trajectory size K, the optimum will be realized by FES, as it is evident from the trend in Fig. 4(d), because it is a strictly proper score, whereas the other metrics get farther from the optimum. Moreover, lower variance is encouraged in the case of FDE since it evaluates the first moment.

Empirical results from this experiment confirm our Propositions 4.1 and 4.2. Additional figures and extended results can be found in Appendix C.

6.2. Effect of Sample Size

Similar to the experiment in Appendix F of (Bhattacharyya et al., 2020), we demonstrate how each metric is sensitive to the number of trajectories as shown in Table 1. Authors in (Bhattacharyya et al., 2020) previously showed that $TopK\%$ scores are more desirable over $TopK$ since they are more consistent. In this experiment, we show a similar comparison between $ADE_{(L>1)}$ and $ADE_{(L=1)}$.

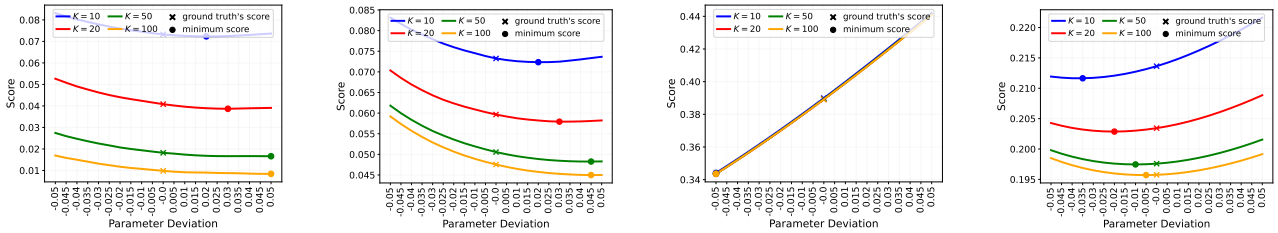
Table 1. Scores of an optimal prediction calculated over an expanding window in the range $t = [1, 3]$ for different values of K. All the scores reported in the table were multiplied by 100 for easier readability.

METRIC	K	$t=1$	$t=2$	$t=3$
$ADE_{(L=1)}$	20	1.20	4.00	6.50
	50	0.50	2.60	4.70
	100	0.30	1.90	3.70
$ADE_{(L=0.1K)}$	20	1.70	5.00	7.70
	50	1.40	4.60	7.20
	100	1.40	4.50	7.00
ES	20	11.70	19.20	25.10
	50	11.30	18.60	24.40
	100	11.20	18.40	24.10

We use the same data-generating process as in Section 6.1. Table 1 clearly shows that $ADE_{(L=1)}$ is not comparable across different K. More precisely, it decreases by increasing K while the prediction has not changed. On the other hand, $ADE_{(L>1)}$ is more consistent across different values of K. Despite $ADE_{(L>1)}$ seems to be more consistent than $ADE_{(L=1)}$, due to being improper, as we discussed in Proposition 4.2 and showed empirically in Section 6.1, in general, we do not recommend it. At the same time, the ES defined in Eq. 10 is a measure where the score is cross-comparable across different values of K because it estimates the same expected value. For extended results of this experiment, refer to the Appendix D.

6.3. Real Data

For the real data experiment, we adopt the pre-trained models from (Bae et al., 2022) and evaluate their models with Energy Score (ES) alongside MoN alternatives such as minADE and minFDE. Table 2 partially reproduces Table 2 of (Bae et al., 2022). It contains two types of trajectory prediction models, namely the Gaussian distribution-based (STGCNN, SGCN) and CVAE-based (PECNET), with three different sampling procedures: Monte-Carlo (MC), Quasi-monte Carlo (QMC) and Non-probability Sampling Network (NPSN). In total, we evaluated nine different models. With this experiment, we wanted to show that the ES ranks the models differently than its counterpart,



(a) $FDE_{(L=1)}$ (b) $FDE_{(L=0.1K)}$ (c) $FDE_{(L=K)}$ (d) FES

Figure 4. Empirically showing how propriety manifests itself. The X-axis represents predictions with different deviations from the optimal parameters. A strictly proper metric gets minimized at the optimal parameter (deviation=0) as K increases. As a reminder: $FDE_{(L=1)} = minFDE$ and $FDE_{(L=K)} = FDE$.

Table 2. baselines: *-mc. The reported numbers are ADE/FDE/ES, which are averages over instances within each dataset. The last column is the average of all datasets. Bold: best model, underline: second best model

	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
STGCNN-MC	0.65/1.10/1.44	0.50/0.86/1.05	0.44/0.80/0.96	0.34/0.53/0.77	0.30/0.48/0.67	0.45/0.75/0.98
PECNET-MC	0.61/1.07/1.64	0.22/0.39/0.70	0.33/0.56/0.89	0.25/0.45/0.74	0.19/0.33/0.65	0.32/0.56/0.92
SGCN-MC	0.57/1.00/1.34	0.31/0.53/0.73	0.37/0.67/0.85	0.29/0.51/0.68	0.22/0.42/0.53	0.35/0.63/0.82
STGCNN-QMC	0.61/1.03/1.30	0.34/0.52/0.98	0.36/0.63/0.89	0.32/0.53/0.74	0.29/0.50/0.65	0.38/0.64/0.91
PECNET-QMC	0.60/1.05/1.62	0.21/0.37/0.68	0.33/0.54/0.88	0.24/0.43/0.72	0.18/0.31/0.62	0.31/0.54/0.91
SGCN-QMC	0.49/0.81/**1.23**	0.21/0.31/0.66	0.31/0.56/**0.78**	0.25/0.45/**0.63**	0.19/0.36/**0.49**	0.29/0.50/**0.76**
STGCNN-NPSN	0.44/0.65/1.48	0.21/0.34/0.88	0.28/0.44/0.88	0.25/0.43/0.83	0.22/0.38/0.73	0.28/0.45/0.96
PECNET-NPSN	0.55/0.88/1.60	0.19/0.29/0.63	0.29/0.44/0.88	0.21/0.33/0.70	0.16/**0.25**/0.56	0.28/0.44/0.87
SGCN-NPSN	**0.36/0.59/1.23**	**0.16/0.25/0.62**	**0.23/0.39**/0.79	**0.18/0.32**/0.66	**0.14/0.25**/0.50	**0.21/0.36/0.76**

MoN. When taking the arithmetic average (*AVG*) across five datasets, ES considers SGCN-NPSN to be equally good as SGCN-QMC, while $minADE/minFDE$ considers SGCN-NPSN alone to be the best model. At the same time, ES and $minADE/minFDE$ disagree on 3 out of 5 datasets (*UNIV, ZARA1, ZARA2*) on the best model while agreeing on *HOTEL* and having a tie between best and second best on *ETH*. Therefore, when it comes to a majority decision, ES favors SGCN-QMC over SGCN-NPSN, in contrast to $minADE/minFDE$, which prefers SGCN-NPSN consistently across all datasets. This implies that there are aspects in which the predictions of SGCN-NPSN are worse than SGCN-QMC on 3 out of 5 datasets. The Energy Score captures these aspects, but $minADE$ and $minFDE$ do not. Therefore, in a practical application, one should not blindly use SGCN-NPSN but also consider SGCN-QMC because it can be better in some scenarios, depending on the downstream task. Therefore, whether to consider SGCN-NPSN as the best depends on the downstream use case and objective. Moreover, the reason why $minADE/minFDE$ consistently ranks SGCN-NPSN as the best method can be partially attributed to the fact that the NPSN method optimizes for $minADE$ as part of the objective (Bae et al., 2022). Thus, it is expected that the $minADE/minFDE$ evaluations would show the lowest scores for *-NPSN models. Further results and comparison across other metrics can be found in Appendix E.

7. Conclusion

We extended the previous analysis on MoN (Thiede & Brahma, 2019) by studying it under the proper scoring rules framework. We showed that MoN is not strictly proper, meaning it cannot uniquely identify the optimal predictive distribution that matches the ground truth. If used for developing and tuning prediction methods, it gives wrong incentives. We adopted the energy score as a strictly proper alternative to evaluating the MTPs, which evaluates the distribution of the trajectory predictions. Moreover, Energy Score can be used for both parametric and non-parametric PDFs, making it more flexible than MoN, which is a non-parametric approach and lacks an analytical derivation. We also proposed three different ways to use the Energy Score for the evaluation of multimodal trajectory predictions that allow investigation of the impact of joint information by looking at marginalized versions of the energy score and comparing them with the joint energy score, which can be a useful tool for diagnosing evaluation of trajectory distribution predictions.

In summary, we think that distribution-aware evaluation of trajectory predictions can benefit from principles of proper scoring rules, and it should be adopted as a more reliable

alternative for evaluating trajectory distribution predictions. However, the choice of the metric should ideally be informed by the downstream task, and thus, there are surely tasks where an MoN metric would be the right metric to use. Therefore, making a final call about which forecast or model to choose over another amidst different rankings reported by different metrics requires further research with a holistic view encompassing the downstream objectives.

8. Limitations and future work

Calculating Energy Score has a computational complexity of $\mathcal{O}(K^2)$; however, there exist estimations with the complexity of $\mathcal{O}(K)$ that can be considered (Ziel & Berk, 2019). Moreover, Energy Score has been successfully applied for relatively large K, e.g., $K = 100$ in other application domains (Dumas et al., 2022). In our work, we considered K as low as 10 and 20 for the synthetic and real data, respectively, which is common for the trajectory prediction tasks. However, for applications that require even lower values, its impact on the variance of the scores merits further investigation.

Acknowledgements

This work was supported by the Estonian Research Council grant PRG1604, the European Social Fund via IT Academy program and partly by the Estonian Centre of Excellence in AI, funded by the Estonian Ministry of Education and Research. We would also like to express our sincere gratitude to Karl Riis and Arun Singh for their invaluable discussions.

Impact Statement

This paper seeks to advance the field of machine learning with a focus on reliability and performance evaluation. While there are societal and ethical implications inherent in our research, we have not identified any specific consequences beyond those already recognized within the broader field.

References

Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., and Savarese, S. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Amirian, J., Hayet, J.-B., and Pettre, J. Social ways: Learning multi-modal distributions of pedestrian trajectories with gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

Anderson, M. J. Permutational multivariate analysis of variance. *Department of Statistics, University of Auckland, Auckland*, 26:32–46, 2005.

Arnold, B. C., Balakrishnan, N., and Nagaraja, H. N. *A First Course in Order Statistics*. Society for Industrial and Applied Mathematics, 2008. doi: 10.1137/1.9780898719062. URL https://epubs.siam.org/doi/abs/10.1137/1.9780898719062.

Bae, I., Park, J.-H., and Jeon, H.-G. Non-probability sampling network for stochastic human trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6477–6487, June 2022.

Bhattacharyya, A., Schiele, B., and Fritz, M. Accurate and diverse sampling of sequences based on a "best of many" sample objective. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Bhattacharyya, A., Hanselmann, M., Fritz, M., Schiele, B., and Straehle, C.-N. Conditional Flow Variational Autoencoders for Structured Sequence Prediction, August 2020. URL http://arxiv.org/abs/1908.09008. arXiv:1908.09008 [cs, stat].

Chen, G., Li, J., Zhou, N., Ren, L., and Lu, J. Personalized trajectory prediction via distribution discrimination. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15580–15589, October 2021.

Cui, H., Radosavljevic, V., Chou, F.-C., Lin, T.-H., Nguyen, T., Huang, T.-K., Schneider, J., and Djuric, N. Multimodal trajectory predictions for autonomous driving using deep convolutional networks. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 2090–2096, 2019. doi: 10.1109/ICRA.2019.8793868.

Dendorfer, P., Elflein, S., and Leal-Taixé, L. Mg-gan: A multi-generator model preventing out-of-distribution samples in pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13158–13167, October 2021.

Dumas, J., Wehenkel, A., Lanaspeze, D., Cornélusse, B., and Sutera, A. A deep generative model for probabilistic energy forecasting in power systems: normalizing flows. *Applied Energy*, 305:117871, 2022.

Ess, A., Leibe, B., and Van Gool, L. Depth and Appearance for Mobile Scene Analysis. In *2007 IEEE 11th International Conference on Computer Vision*, pp. 1–8, October 2007. doi: 10.1109/ICCV.2007.4409092. ISSN: 2380-7504.

Gao, J., Sun, C., Zhao, H., Shen, Y., Anguelov, D., Li, C., and Schmid, C. Vectornet: Encoding hd maps and agent dynamics from vectorized representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Gneiting, T. and Raftery, A. E. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378, March 2007. ISSN 0162-1459. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437. Publisher: Taylor & Francis _eprint: https://doi.org/10.1198/016214506000001437.

Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., and Alahi, A. Social GAN: Socially Acceptable Trajectories With Generative Adversarial Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2255–2264, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Gupta_Social_GAN_Socially_CVPR_2018_paper.html.

Huang, R., Xue, H., Pagnucco, M., Salim, F., and Song, Y. Multimodal Trajectory Prediction: A Survey, February 2023. URL http://arxiv.org/abs/2302.10463. arXiv:2302.10463 [cs].

Ivanovic, B. and Pavone, M. Rethinking Trajectory Forecasting Evaluation, July 2021. URL http://arxiv.org/abs/2107.10297. arXiv:2107.10297 [cs, eess].

Lee, N., Choi, W., Vernaza, P., Choy, C. B., Torr, P. H. S., and Chandraker, M. Desire: Distant future prediction in dynamic scenes with interacting agents. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Li, L., Pagnucco, M., and Song, Y. Graph-based spatial transformer with memory replay for multi-future pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2231–2241, June 2022.

Ma, Y. J., Inala, J. P., Jayaraman, D., and Bastani, O. Likelihood-based diverse sampling for trajectory forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13279–13288, October 2021.

Mangalam, K., Girase, H., Agarwal, S., Lee, K.-H., Adeli, E., Malik, J., and Gaidon, A. It Is Not the Journey But the Destination: Endpoint Conditioned Trajectory Prediction. In Vedaldi, A., Bischof, H., Brox, T., and Frahm, J.-M. (eds.), *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, pp. 759–776, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58536-5. doi: 10.1007/978-3-030-58536-5_45.

Mangalam, K., An, Y., Girase, H., and Malik, J. From Goals, Waypoints & Paths To Long Term Human Trajectory Forecasting. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15213–15222, Montreal, QC, Canada, October 2021. IEEE. ISBN 978-1-66542-812-5. doi: 10.1109/ICCV48922.2021.01495. URL https://ieeexplore.ieee.org/document/9709992/.

Mohamed, A., Qian, K., Elhoseiny, M., and Claudel, C. Social-STGCNN: A Social Spatio-Temporal Graph Convolutional Neural Network for Human Trajectory Prediction. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14412–14420, Seattle, WA, USA, June 2020. IEEE. ISBN 978-1-72817-168-5. doi: 10.1109/CVPR42600.2020.01443. URL https://ieeexplore.ieee.org/document/9156583/.

Morales, J. M., Conejo, A. J., Madsen, H., Pinson, P., and Zugno, M. *Integrating renewables in electricity markets: operational problems*, volume 205. Springer Science & Business Media, 2013.

Park, S. H., Kim, B., Kang, C. M., Chung, C. C., and Choi, J. W. Sequence-to-sequence prediction of vehicle trajectory via lstm encoder-decoder architecture. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1672–1678, 2018. doi: 10.1109/IVS.2018.8500658.

Rhinehart, N., Kitani, K. M., and Vernaza, P. R2p2: A reparameterized pushforward policy for diverse, precise generative path forecasting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

Rudenko, A., Palmieri, L., Herman, M., Kitani, K. M., Gavrila, D. M., and Arras, K. O. Human motion trajectory prediction: a survey. *The International Journal of Robotics Research*, 39(8):895–935, July 2020. ISSN 0278-3649. doi: 10.1177/0278364920917446. URL https://doi.org/10.1177/0278364920917446. Publisher: SAGE Publications Ltd STM.

Sadeghian, A., Kosaraju, V., Sadeghian, A., Hirose, N., Rezatofighi, H., and Savarese, S. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Salzmann, T., Ivanovic, B., Chakravarty, P., and Pavone, M. Trajectron++: Dynamically-Feasible Trajectory Forecasting with Heterogeneous Data. In Vedaldi, A., Bischof, H., Brox, T., and Frahm, J.-M. (eds.), *Computer Vision – ECCV 2020*, Lecture Notes in Computer Science, pp.

683–700, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58523-5. doi: 10.1007/978-3-030-58523-5_40.

Schöller, C., Aravantinos, V., Lay, F., and Knoll, A. What the Constant Velocity Model Can Teach Us About Pedestrian Motion Prediction. *IEEE Robotics and Automation Letters*, 5(2):1696–1703, April 2020. ISSN 2377-3766. doi: 10.1109/LRA.2020.2969925. Conference Name: IEEE Robotics and Automation Letters.

Székely, G. J. and Rizzo, M. L. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, August 2013. ISSN 0378-3758. doi: 10.1016/j.jspi.2013.03.018. URL https://www.sciencedirect.com/science/article/pii/S0378375813000633.

Tang, C. and Salakhutdinov, R. R. Multiple futures prediction. *Advances in neural information processing systems*, 32, 2019.

Thiede, L. A. and Brahma, P. P. Analyzing the variety loss in the context of probabilistic trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9954–9963, 2019.

Yue, J., Manocha, D., and Wang, H. Human Trajectory Prediction via Neural Social Physics. In Avidan, S., Brostow, G., Cissé, M., Farinella, G. M., and Hassner, T. (eds.), *Computer Vision – ECCV 2022*, Lecture Notes in Computer Science, pp. 376–394, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19830-4. doi: 10.1007/978-3-031-19830-4_22.

Ziel, F. and Berk, K. Multivariate Forecasting Evaluation: On Sensitive and Strictly Proper Scoring Rules. *Papers*, October 2019. URL https://ideas.repec.org//p/arx/papers/1910.07325.html. Number: 1910.07325 Publisher: arXiv.org.

Appendix

A. Theory

A.1. Proof of Proposition 4.1

Proof. We need to show that the Average Displacement Error is improper. For this, we need to find two distributions that violate properness. For simplicity, we show it for one-dimensional random variables. Let $\mathbf{Y} = (Y^1, \dots, Y^T)$ be a random vector, where $Y^t \sim N(\mu_{Y^t}, \sigma^2_{Y^t})$ is distributed normally for $t = 1, \dots, T$. We need to show that there exists a random vector $\mathbf{X} = (X^1, \dots, X^T)$ such that $ADE(\mathbf{X}, \mathbf{Y}) < ADE(\mathbf{Y}, \mathbf{Y})$. From the definition of ADE, we know that

$$ADE(\mathbf{X}, \mathbf{Y}) = \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} DE(X^t, Y^t)\right] = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[DE(X^t, Y^t)]$$

Let's choose $X^t \sim N(\mu_{X^t}, \sigma^2_{X^t})$, with values of μ_{X^t} and $\sigma^2_{X^t}$ specified later. Then $DE(X^t, Y^t) = |X^t - Y^t|$. We know that the difference of normally distributed random variables is also with normal distribution and $X^t - Y^t \sim N(\mu_{X^t} - \mu_{Y^t}, \sigma^2_{X^t} + \sigma^2_{Y^t})$. Also, we know that the absolute value of a normally distributed random variable is with the folded normal distribution. By using the known formula for the mean of a folded normal distribution, we get that

$$\mathbb{E}[DE(X^t, Y^t)] = \mathbb{E}[|X^t - Y^t|] = \sqrt{\sigma^2_{X^t} + \sigma^2_{Y^t}}\frac{\sqrt{2}}{\sqrt{\pi}} e^{-\frac{(\mu_{X^t} - \mu_{Y^t})^2}{2(\sigma^2_{X^t} + \sigma^2_{Y^t})}} + (\mu_{X^t} - \mu_{Y^t})\left(1 - 2\Phi\left(-\frac{\mu_{X^t} - \mu_{Y^t}}{\sqrt{\sigma^2_{X^t} + \sigma^2_{Y^t}}}\right)\right)$$

To construct an example where $ADE(\mathbf{X}, \mathbf{Y}) < ADE(\mathbf{Y}, \mathbf{Y})$ we pick $\mathbf{X}$ such that $\mu_{X^t} = \mu_{Y^t}$ and $\sigma^2_{X^t} \to 0$. Then

$$\mathbb{E}[DE(X^t, Y^t)] = \sqrt{\sigma^2_{X^t} + \sigma^2_{Y^t}}\frac{\sqrt{2}}{\sqrt{\pi}} \to \sqrt{\sigma^2_{Y^t}}\frac{\sqrt{2}}{\sqrt{\pi}}$$

From this, we get that

$$ADE(\mathbf{X}, \mathbf{Y}) \to \frac{1}{T}\sum_{t=1}^{T}\sqrt{\sigma^2_{Y^t}}\frac{\sqrt{2}}{\sqrt{\pi}} < \frac{1}{T}\sum_{t=1}^{T}\sqrt{\sigma^2_{Y^t} + \sigma^2_{Y^t}}\frac{\sqrt{2}}{\sqrt{\pi}} = ADE(\mathbf{Y}, \mathbf{Y})$$

□

A.2. Proof of Proposition 4.2

Proof. Let's look at a set of trajectories $\mathbf{X} = X_1, \dots, X_K \overset{i.i.d}{\sim} Ber(p_X)$ and a random variable $Y \sim Ber(p_Y)$ in one dimensional space. We fix $T = 1$ for simplicity of proof. We define a new random variable $Z_{i:K}$ as i-th order statistic of $\{|X_1 - y|, \dots, |X_K - y|\}$. Then

$$ADE_{(L)}(\mathbf{X}, Y) = \mathbb{E}\min_{\substack{\{k_1,\dots,k_L\}\\ k_i \neq k_j}} \frac{1}{LT}\sum_{l=1}^{L}\sum_{t=1}^{T} DE(X^t_{k_l}, Y^t) = \mathbb{E}[\mathbb{E}[\frac{1}{L}\sum_{l=1}^{L} Z_{l:K}|Y = y]] = \mathbb{E}[\frac{1}{L}\sum_{l=1}^{L}\mathbb{E}[Z_{l:K}|Y = y]]$$

To find the distribution of $Z_{i:K}$, we must first consider the distribution of $|X_i - y|$. We get that $|X_i - y| \sim Ber(|y - p_X|), \forall i \in \{1, \dots, K\}$. Since $\{|X_1 - y|, \dots, |X_K - y|\}$ are independent and with the same Bernoulli distribution when y is fixed, then we know what the expected value in this case is (Arnold et al., 2008):

$$\mu_{l:K} = \mathbb{E}[Z_{l:K}|Y = y] = \sum_{r=K-l+1}^{K}\binom{K}{r}|y - p_X|^r(1 - |y - p_X|)^{K-r}$$

From this, we get that

$$ADE_{(L)}(\mathbf{X},Y) = \mathbb{E}[\frac{1}{L}\sum_{l=1}^{L}\mathbb{E}[Z_{l:K}|Y=y]] = \mathbb{E}[\frac{1}{L}\sum_{l=1}^{L}\sum_{r=K-l+1}^{K}\binom{K}{r}|y-p_X|^r(1-|y-p_X|)^{K-r}] =$$
$$(1-p_Y)[\frac{1}{L}\sum_{l=1}^{L}\sum_{r=K-l+1}^{K}\binom{K}{r}p_X^r(1-p_X)^{K-r}] + p_Y[\frac{1}{L}\sum_{l=1}^{L}\sum_{r=K-l+1}^{K}\binom{K}{r}(1-p_X)^r p_X^{K-r}] =$$
$$\frac{1}{L}\sum_{l=1}^{L}\sum_{r=K-l+1}^{K}\binom{K}{r}[(1-p_Y)p_X^r(1-p_X)^{K-r} + p_Y(1-p_X)^r p_X^{K-r}]$$

Using this formula, we can calculate the $ADE_{(L)}$ for different values of L and K. When we fix $L=1$, which is $minADE$, then

$$ADE_{(1)}(\mathbf{X},Y) = minADE(\mathbf{X},Y) = (1-p_Y)p_X^K + p_Y(1-p_X)^K$$

Taking the derivative with respect to p_X and setting it equal to zero, we can find the value of p_X, which gives the optimal result:

$$\frac{d}{dp_X}((1-p_Y)p_X^K + p_Y(1-p_X)^K) = 0$$
$$(1-p_Y)Kp_X^{K-1} - p_Y K(1-p_X)^{K-1} = 0$$
$$\frac{(1-p_X)^{K-1}}{p_X^{K-1}} = \frac{1-p_Y}{p_Y}$$
$$\frac{1-p_X}{p_X} = \sqrt[K-1]{\frac{1-p_Y}{p_Y}}$$
$$p_X = \frac{1}{\sqrt[K-1]{\frac{1}{p_Y}-1}+1}$$

From the equation, we see that in Bernoulli's case, it is strictly proper only when $K=2$. When $K=1$ and $p_X > 0.5$, then $p_X = 1$ gives the lowest ADE. When $p_Y < 0.5$, then lowest ADE is obtained by $p_X = 0$. When $K >= 3$ and the $p_Y \neq 0.5$, the lowest ADE is obtained by p_X value that is between p_Y and 0.5. We also see that when $K \to \infty$, the optimal $p_X \to 0.5$.

Let $L=2$. Then

$$ADE_{(2)}(\mathbf{X},Y) = \frac{1}{2}\sum_{l=1}^{2}\sum_{r=K-l+1}^{K}\binom{K}{r}[(1-p_Y)p_X^r(1-p_X)^{K-r} + p_Y(1-p_X)^r p_X^{K-r}] =$$
$$\frac{1}{2}[2[(1-p_Y)p_X^K + p_Y(1-p_X)^K] + K(1-p_Y)p_X^{K-1}(1-p_X) + Kp_Y(1-p_X)^{K-1}p_X]$$

To show that p_Y is not optimal (we assume that $p_Y \neq 0.5$), we take $p_X = p_Y$ and show that the derivative is not equal to zero. We get that

$$ADE_{(2)}(Y,Y) = \frac{1}{2}[2[(1-p_Y)p_Y^K + p_Y(1-p_Y)^K] + K(1-p_Y)p_Y^{K-1}(1-p_Y) + Kp_Y(1-p_Y)^{K-1}p_Y] =$$
$$\frac{1}{2}[2[(1-p_Y)p_Y^K + p_Y(1-p_Y)^K] + K(1-p_Y)^2 p_Y^{K-1} + Kp_Y^2(1-p_Y)^{K-1}]$$

Taking the derivative with respect to p_Y, we get

$$\begin{aligned}
\frac{d}{dp_Y}[(1-p_Y)p_Y^K + p_Y(1-p_Y)^K + 0.5K(1-p_Y)^2p_Y^{K-1} + 0.5Kp_Y^2(1-p_Y)^{K-1}] = \\
-p_Y^K + K(1-p_Y)p_Y^{K-1} + (1-p_Y)^K - Kp_Y(1-p_Y)^{K-1} - K(1-p_Y)p_Y^{K-1} \\
+0.5K(K-1)(1-p_Y)^2p_Y^{K-2} + Kp_Y(1-p_Y)^{K-1} - 0.5K(K-1)p_Y^2(1-p_Y)^{K-2} = \\
-p_Y^K + (1-p_Y)^K + 0.5K(K-1)(1-p_Y)^2p_Y^{K-2} - 0.5K(K-1)p_Y^2(1-p_Y)^{K-2} = \\
-p_Y^K + (1-p_Y)^K + 0.5K(K-1)((1-p_Y)^2p_Y^{K-2} - p_Y^2(1-p_Y)^{K-2})
\end{aligned}$$

Then, if $K = 2$, we get that

$$-p_Y^K + (1-p_Y)^K + 0.5K(K-1)((1-p_Y)^2p_Y^{K-2} - p_Y^2(1-p_Y)^{K-2}) = 2 - 4p_Y$$

From this, we see that $p_X = p_Y$ does not give the optimal value since the derivative is not zero. If $K = 3$, we get $4p_Y^3 - 6p_Y^2 + 1$, which is positive, when $p_Y < 0.5$ and negative when $p_Y > 0.5$. □

A.3. Proof of Proposition 4.3

Proof. Let $\mathbf{F_X}$ be the predicted distribution of $\mathbf{F_Y}$ and $\mathbf{X}^{(K)}$ is random vector containing K elements and $\mathbf{X}^{(K+1)}$ is random vector containing $K+1$ elements. $\mathbf{Y}$ is a random vector from $\mathbf{F_Y}$. Let's fix L, so that $L \leq K$, and $\mathbf{Y} = \mathbf{y}$. First, we find the displacement error for each trajectory in $\mathbf{X}^{(K)}$. We get that $Z_1 = \frac{1}{T}\sum_{t=1}^{T} DE(X_1^t, y^t), Z_2 = \frac{1}{T}\sum_{t=1}^{T} DE(X_2^t, y^t), \ldots, Z_K = \frac{1}{T}\sum_{t=1}^{T} DE(X_K^t, y^t)$. If we order them, we obtain order statistics $Z_{1:K} \leq Z_{2:K} \leq \ldots \leq Z_{K:K}$. We also do the same for $\mathbf{X}^{(K+1)}$. To show that

$$ADE_{(L)}(\mathbf{X}^{(K)}, \mathbf{Y}) = \mathbb{E}[\mathbb{E}[\frac{1}{L}\sum_{l=1}^{L} Z_{l:K}|\mathbf{Y} = \mathbf{y}]] \geq \mathbb{E}[\mathbb{E}[\frac{1}{L}\sum_{l=1}^{L} Z_{l:K+1}|\mathbf{Y} = \mathbf{y}]] = ADE_{(L)}(\mathbf{X}^{(K+1)}, \mathbf{Y})$$

we need to show that $\mathbb{E}[\frac{1}{L}\sum_{l=1}^{L} Z_{l:K}|\mathbf{Y} = \mathbf{y}] \geq \mathbb{E}[\frac{1}{L}\sum_{l=1}^{L} Z_{l:K+1}|\mathbf{Y} = \mathbf{y}]$ for every $\mathbf{y}$. Since we fixed $\mathbf{y}$ arbitrarily, we achieve it by showing that $\mathbb{E}[Z_{i,K}|\mathbf{Y} = \mathbf{y}] \geq \mathbb{E}[Z_{i,K+1}|\mathbf{Y} = \mathbf{y}] \quad \forall i \in [1, K]$. It is known that for $1 \leq i \leq K$

$$i \cdot \mathbb{E}[Z_{i+1:K+1}|\mathbf{Y} = \mathbf{y}] + (K+1-i) \cdot \mathbb{E}[Z_{i:K+1}|\mathbf{Y} = \mathbf{y}] = (K+1) \cdot \mathbb{E}[Z_{i:K}|\mathbf{Y} = \mathbf{y}]$$

For proof, see "A First Course in Order Statistics" Theorem 5.3.1 (Arnold et al., 2008). Due to the fact that expected values are always non-negative, in our case, we get that

$$\begin{aligned}
i \cdot \mathbb{E}[Z_{i+1:K+1}|\mathbf{Y} = \mathbf{y}] + (K+1-i)\mathbb{E}[Z_{i:K+1}|\mathbf{Y} = \mathbf{y}] = (K+1) \cdot \mathbb{E}[Z_{i:K}|\mathbf{Y} = \mathbf{y}] \Leftrightarrow \\
i \cdot \mathbb{E}[Z_{i+1:K+1}|\mathbf{Y} = \mathbf{y}] - i \cdot \mathbb{E}[Z_{i:K+1}|\mathbf{Y} = \mathbf{y}] + (K+1) \cdot \mathbb{E}[Z_{i:K+1}|\mathbf{Y} = \mathbf{y}] = (K+1) \cdot \mathbb{E}[Z_{i:K}|\mathbf{Y} = \mathbf{y}] \Rightarrow \\
\mathbb{E}[Z_{i:K+1}|\mathbf{Y} = \mathbf{y}] \leq \mathbb{E}[Z_{i:K}|\mathbf{Y} = \mathbf{y}]
\end{aligned}$$

because $i \cdot \mathbb{E}[Z_{i+1:K+1}|\mathbf{Y} = \mathbf{y}] - i \cdot \mathbb{E}[Z_{i:K+1}|\mathbf{Y} = \mathbf{y}] \geq 0$. This proves that $ADE_{(L)}(\mathbf{X}^{(K)}, \mathbf{Y}) \geq ADE_{(L)}(\mathbf{X}^{(K+1)}, \mathbf{Y})$. □

A.4. Proof of Proposition 4.4

Proof. Let $\mathbf{F_X}$ be the predicted distribution of $\mathbf{F_Y}$ and $\mathbf{X}^{(K)}$ is random vector containing K elements and $\mathbf{Y}$ is a random vector from $\mathbf{F_Y}$. We also know that $\text{supp}(\mathbf{F_Y}) \subset \text{supp}(\mathbf{F_X})$. Similarly to the last proof, we obtain the order statistic $Z_{i:K}$ from it. We need to show that

$$\lim_{K\to\infty} ADE_{(L)}(\mathbf{X}^{(K)}, \mathbf{Y}) = \lim_{K\to\infty} \mathbb{E}[\mathbb{E}[\frac{1}{L}\sum_{l=1}^{L} Z_{l:K}|\mathbf{Y} = \mathbf{y}]] \to 0$$

We obtain it by showing that $\mathbb{E}[Z_{L:K}|\mathbf{Y} = \mathbf{y}] \to 0$ for every $\mathbf{y}$, because $\mathbb{E}[Z_{i:K}|\mathbf{Y} = \mathbf{y}] \leq \mathbb{E}[Z_{i+1:K}|\mathbf{Y} = \mathbf{y}]$ for every $i \in [1, K-1]$. Let us fix $\mathbf{y}$. Let $L = K \cdot p$, where $0 < p < 1$. Then, in the limit, we know that

$$\mathbb{E}[Z_{L:K}|\mathbf{Y} = \mathbf{y}] = \mathbb{E}[Z_{Kp:K}|\mathbf{Y} = \mathbf{y}] \simeq F^{-1}(p)$$

where F denotes the cumulative distribution function of the $\frac{1}{T}\sum_{t=1}^{T} DE(X^t, y^t)$. For proof about the limit, see "A First Course in Order Statistics" (Arnold et al., 2008). When $K \to \infty$, then $p \to 0$, because L is fixed. Since $\text{supp}(\mathbf{F_Y}) \subset \text{supp}(\mathbf{F_X})$ then $F(\varepsilon) = P(\frac{1}{T}\sum_{t=1}^{T} DE(X^t, y^t) < \varepsilon) > 0$ for every $\mathbf{y}$ value and $\varepsilon > 0$. This means that $F^{-1}(p) \to 0$ and we get that $\mathbb{E}[Z_{L:K}|\mathbf{Y} = \mathbf{y}] \to 0$.

□

B. Bernoulli Simulation

In this section, we confirm our theoretical results empirically through a simulation. Ground truth variable $Y \sim Ber(p_y)$ and prediction variable $X \sim Ber(p_x)$ are defined as follows:

$$P_Y(y) = \begin{cases} 1 - p_y & y = 0 \\ p_y & y = 1 \end{cases} \qquad P_X(x) = \begin{cases} 1 - p_x & x = 0 \\ p_x & x = 1 \end{cases}$$

We consider five prediction cases as depicted in Fig. 5. Each case represents a prototypical case of a probabilistically optimal prediction that matches the ground truth, a random (uniform) prediction, and an overconfident or underconfident prediction. We set the number of observations to $N = 5000$ and estimate $ADE_{(L)}$ under different values of L and K as reported in Tables 3 to 5.

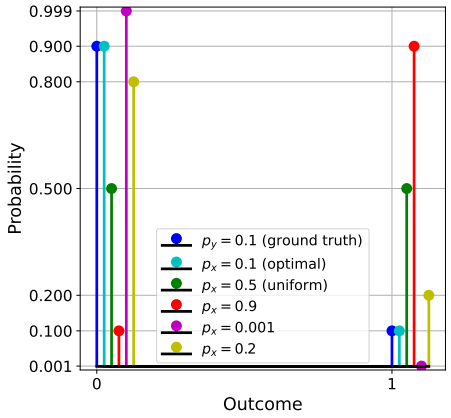


Figure 5. Probability mass function of five different Bernoulli predictions p_x and ground truth p_y

Table 3. Results obtained for $ADE_{(L=1)}$

	K=1	K=2	K=3	K=10	K=20	K=50	K=100
$p_x = 0.1$ (GROUND TRUTH)	0.175	**0.091**	0.072	0.041	0.014	0.001	**0.000**
$p_x = 0.5$ (UNIFORM)	0.497	0.253	0.127	**0.001**	**0.000**	**0.000**	**0.000**
$p_x = 0.9$	0.824	0.728	0.663	0.317	0.108	0.005	**0.000**
$p_x = 0.01$	**0.108**	0.102	0.092	0.093	0.083	0.059	0.036
$p_x = 0.2$	0.264	0.101	**0.059**	0.011	0.002	**0.000**	**0.000**

Table 3 confirms the theoretical results from Appendix A.2. For the case of $L = 1$, when $K = 1$, prediction $p_x = 0.001$ yields the lowest error. This is expected as this prediction has a higher probability than the ground truth for the majority class (or a lower probability for the minority class) – in other words, it has sharper mass. For $K = 2$, as expected, we see that the optimal prediction was successfully detected, and as K increases, the uniform prediction yields the lowest ADE in the limit. In our example, this is already realized for $K = 10$, but it would be higher values depending on how complicated the distribution is. Also, yet again, we observe that as K increases, the error tends to converge to zero, which, besides our other empirical results from Section 6.2, confirms our Proposition 4.4.

Table 4. Results obtained for $ADE_{(L=2)}$

	K=2	K=3	K=10	K=20	K=50	K=100
$p_x = 0.1$ (GROUND TRUTH)	0.177	**0.103**	0.058	0.025	0.002	**0.000**
$p_x = 0.5$ (UNIFORM)	0.502	0.315	**0.007**	**0.000**	**0.000**	**0.000**
$p_x = 0.9$	0.819	0.766	0.487	0.230	0.018	**0.000**
$p_x = 0.01$	**0.105**	**0.103**	0.100	0.091	0.079	0.054
$p_x = 0.2$	0.256	0.125	0.025	0.005	**0.000**	**0.000**

Similar observations can be made in Tables 4 and 5 for cases $L = 2$ and $L = 10$. The minimum K in each case is 2 and 10, respectively. As a reminder, when $L = K$, we have $ADE_{(L)} \equiv ADE$. Hence, the observation from Appendix A.1 and Section 6.1 that ADE prefers sharper predictions can be once more established, but for the Bernoulli example.

Table 5. Results obtained for $ADE_{(L=10)}$

	K=10	K=20	K=50	K=100
$p_x = 0.1$ (GROUND TRUTH)	0.178	0.079	0.053	0.012
$p_x = 0.5$ (UNIFORM)	0.498	0.086	**0.000**	**0.000**
$p_x = 0.9$	0.823	0.721	0.447	0.104
$p_x = 0.01$	**0.104**	0.096	0.100	0.092
$p_x = 0.2$	0.259	**0.058**	0.012	**0.000**

C. Propriety Showcase – Extended Results

We provide the results for the same setup as in Section 6.1 but for all the other metrics as depicted in Fig. 6.

As we can see in Figures 6(c) and 6(f), FDE and ADE are not proper as they assign a lower error to a prediction that has a lower variance (solid circles) compared to the optimal prediction (crosses). They encourage lower variance to the extent of mode collapse. Comparing three FDE variations in Figures 6(a) to 6(c), they clearly exhibit signs of impropriety. ADE variations in Figures 6(d) to 6(f) show subtle signs of impropriety for this particular example. At the same time, all ES variations become closer to the truth as the sample size K increases. Due to randomness and our limited number of repetitions, they still do not match the optimal parameter value.

Please note that the y-axis scale differs across all the metrics in Figure 6. In this experiment, we wanted to show that by increasing K, the energy scores identify the optimum asymptotically while MoN variations do not. This fact does not depend on the scale of the y-axis.

D. Effect of Sample Size – Extended Results

Effect of trajectory size K reported across all metrics in Tables 6 and 7. Since in our synthetic setup $S = 1$, the only difference between EST and ESS is the marginalization of the temporal dimension. In Table 7, we can see that both metrics are the same at $t = 1$ but as the horizon expands, i.e., $t = 2, 3$, EST reports larger errors than ESS since ESS marginalizes over the temporal steps. These experiments are run with $N = 5000$.

Figures 7 and 8 empirically demonstrate the theoretical results from Proposition 4.4 in that as $K \to \infty$ the MoN-based measures converge to zero. In contrast, in the case of energy score, the score has faster convergence with higher K. Another important observation is that MoN-based measures with different values of K are not cross-comparable, meaning that one has to treat each as a separate metric.

Table 6. The error for each step is calculated based on an expanding window over the horizon (temporal steps). $DE_{(L=1)}$ metrics are sensitive to the trajectory size K. Their error decreases as K increases, while $DE_{(L>1)}$ and energy-score-based metrics are stable with respect to the increase of K. All the scores reported in the table were multiplied by 100 for better readability.

METRIC	K	$t=0$	$t=1$	$t=2$	$t=3$
$FDE_{(L=1)}$	10	0.00	4.20	5.90	7.30
	20	0.00	2.30	3.30	4.10
	50	0.00	1.00	1.50	1.80
	100	0.00	0.60	0.80	1.00
	300	0.00	0.20	0.30	0.40
$FDE_{(L=0.1K)}$	10	0.00	4.20	5.90	7.30
	20	0.00	3.40	4.80	6.00
	50	0.00	2.90	4.10	5.10
	100	0.00	2.70	3.90	4.80
	300	0.00	2.60	3.70	4.50
FES	10	0.00	12.20	17.40	21.40
	20	0.00	11.70	16.60	20.30
	50	0.00	11.30	16.10	19.70
	100	0.00	11.20	15.90	19.50
	300	0.00	11.20	15.90	19.40

METRIC	K	$t=0$	$t=1$	$t=2$	$t=3$
$ADE_{(L=1)}$	10	0.00	2.10	5.50	8.30
	20	0.00	1.20	4.00	6.50
	50	0.00	0.50	2.60	4.70
	100	0.00	0.30	1.90	3.70
	300	0.00	0.10	1.10	2.60
$ADE_{(L=0.1K)}$	10	0.00	2.10	5.50	8.30
	20	0.00	1.70	5.00	7.70
	50	0.00	1.40	4.60	7.20
	100	0.00	1.40	4.50	7.00
	300	0.00	1.30	4.40	6.90
ES	10	0.00	12.20	20.00	26.30
	20	0.00	11.70	19.20	25.10
	50	0.00	11.30	18.60	24.40
	100	0.00	11.20	18.40	24.10
	300	0.00	11.20	18.30	24.00

Table 7. Similar table as in Table 6 but for ESS and EST

METRIC	K	$t=0$	$t=1$	$t=2$	$t=3$
EST	10	0.00	6.10	10.60	13.90
	20	0.00	5.90	10.10	13.20
	50	0.00	5.70	9.80	12.80
	100	0.00	5.60	9.70	12.70
	300	0.00	5.60	9.70	12.60
ESS	10	0.00	6.10	9.90	12.70
	20	0.00	5.90	9.50	12.20
	50	0.00	5.70	9.20	11.80
	100	0.00	5.60	9.10	11.70
	300	0.00	5.60	9.00	11.60

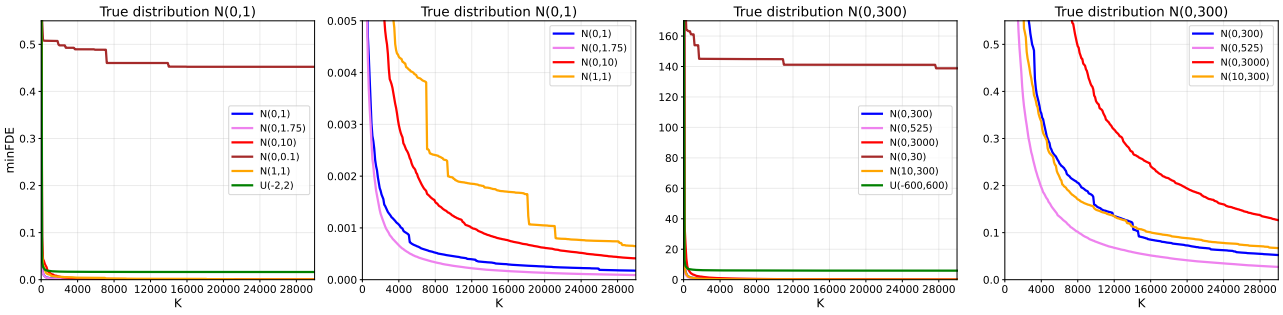


Figure 7. Estimated minFDE values for different distributions when the number of trajectories increases. In the figure, the second plot is the same as the first, but the y-axis is changed to better illustrate the differences for lower values. The same is true for the third and fourth plots.

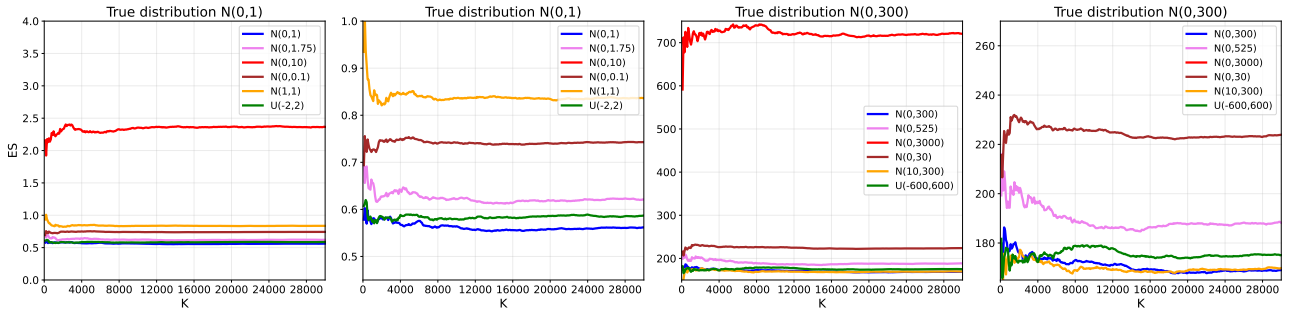


Figure 8. Estimated energy score values for different distributions when the number of trajectories increases. In the figure, the second plot is the same as the first, but the y-axis is changed to better illustrate the differences for lower values. The same is true for the third and fourth plots.

E. Real Data – Extended Results

We report an extension of the results reported earlier for cross-comparison between $minFDE$ and FES in Table 8, $minADE$ and ES in Table 9, ES, EST, and ESS in Table 10 and finally $minADE$, EST and ESS in Table 12. From Table 8, we see that except on ZARA1, $minFDE$ and $minFES$ agree on the best model, and while $minFDE$ chooses NPSN variations of STGCNN and PECNET as the second best model, the FES picks SGCN-QMC as the second best consistently.

Table 8. Reported values: minFDE/FES, which are averages over instances within each dataset. AVG is the arithmetic average over all datasets. **Bold**: best model, underline: second best model. Baselines: *-MC.

	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
STGCNN-MC	1.10/1.51	0.86/1.22	0.80/1.07	0.53/0.84	0.48/0.73	0.75/1.07
PECNET-MC	1.07/1.69	0.39/0.76	0.56/0.96	0.45/0.81	0.33/0.70	0.56/0.99
SGCN-MC	1.00/1.40	0.53/0.82	0.67/0.94	0.51/0.75	0.42/0.58	0.63/0.90
STGCNN-QMC	1.03/1.44	0.52/1.08	0.63/0.98	0.53/0.83	0.50/0.73	0.64/1.01
PECNET-QMC	1.05/1.67	0.37/0.74	0.54/0.95	0.43/0.79	0.31/0.68	0.54/0.97
SGCN-QMC	0.81/1.30	0.31/0.71	0.56/0.87	0.45/**0.71**	0.36/0.55	0.50/0.83
STGCNN-NPSN	0.65/1.54	0.34/0.95	0.44/0.95	0.43/0.90	0.38/0.78	0.45/1.02
PECNET-NPSN	0.88/1.64	0.29/**0.68**	0.44/0.94	0.33/0.76	**0.25**/0.61	0.44/0.93
SGCN-NPSN	**0.59/1.29**	**0.25/0.68**	**0.39/0.85**	**0.32**/0.72	**0.25/0.54**	**0.36/0.82**

Table 9 is the same table as in the body of the paper. In three out of five cases (UNIV, ZARA1, ZARA2), ES ranks SGCN-QMC as the best model, while in one case (ETH), there is a tie between SGCN-QMC and SGCN-NPSN, and finally on average (AVG), these two models do not seem to be that different according to ES. At the same time, $minADE$ picks SGCN-NPSN as the best model, and similar to $minFDE$, it chooses between STGCNN-NPSN and PECNET-NPSN as the second best model.

Table 9. Reported values: minADE/ES, which are averages over instances within each dataset. AVG is the arithmetic average over all datasets. **Bold**: best model, underline: second best model. Baselines: *-MC.

	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
STGCNN-MC	0.65/1.44	0.50/1.05	0.44/0.96	0.34/0.77	0.30/0.67	0.45/0.98
PECNET-MC	0.61/1.64	0.22/0.70	0.33/0.89	0.25/0.74	0.19/0.65	0.32/0.92
SGCN-MC	0.57/1.34	0.31/0.73	0.37/0.85	0.29/0.68	0.22/0.53	0.35/0.82
STGCNN-QMC	0.61/1.30	0.34/0.98	0.36/0.89	0.32/0.74	0.29/0.65	0.38/0.91
PECNET-QMC	0.60/1.62	0.21/0.68	0.33/0.88	0.24/0.72	0.18/0.62	0.31/0.91
SGCN-QMC	0.49/**1.23**	0.21/0.66	0.31/**0.78**	0.25/**0.63**	0.19/**0.49**	0.29/**0.76**
STGCNN-NPSN	0.44/1.48	0.21/0.88	0.28/0.88	0.25/0.83	0.22/0.73	0.28/0.96
PECNET-NPSN	0.55/1.60	0.19/0.63	0.29/0.88	0.21/0.70	0.16/0.56	0.28/0.87
SGCN-NPSN	**0.36/1.23**	**0.16/0.62**	**0.23**/0.79	**0.18**/0.66	**0.14**/0.50	**0.21/0.76**

Table 10 compares reports marginalized versions of ES namely the EST and ESS alongside ES itself. The rankings do not always agree across the three. Recall that EST marginalizes over spatial dimensions and ESS marginalizes over temporal dimension. Hence, they ignore some spatial and temporal information, respectively. We listed possible cases of agreement or disagreement between the three together with a possible explanation for each case in Table 11. For the top-ranked model, in two datasets ETH and HOTEL, we have ALL three agreeing, while for the second-ranked model, in two datasets ZARA1 and ZARA2 we have a case where the three disagree. For the first- and second-ranked models, we do not have a case where ES & EST agree while ESS disagrees. However, we have ES and ESS agree while EST disagrees on the second-ranked model on AVG and similarly but less strongly on the first-ranked model on ZARA1 and ZARA2. We have numerous instances in which EST and ESS agree while ES disagrees. For the top-ranked model, that is the case under datasets UNIV, ZARA1 and ZARA2 and similarly for the second-ranked model under datasets ETH, HOTEL, and UNIV.

Table 10. Reported values: ES/EST/ESS, which are averages over instances within each dataset. AVG is the arithmetic average over all datasets. **Bold**: best model, underline: second best model. Baselines: *-MC.

	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
STGCNN-MC	1.44/1.02/0.75	1.05/0.83/0.61	0.96/0.72/0.52	0.77/0.57/0.41	0.67/0.49/0.36	0.98/0.73/0.53
PECNET-MC	1.64/1.11/0.79	0.70/0.52/0.36	0.89/0.66/0.46	0.74/0.54/0.37	0.65/0.47/0.31	0.92/0.66/0.46
SGCN-MC	1.34/0.93/0.66	0.73/0.55/0.39	0.85/0.63/0.44	0.68/0.50/0.34	0.53/0.39/0.26	0.82/0.60/0.42
STGCNN-QMC	1.30/1.04/0.72	0.98/0.74/0.55	0.89/0.68/0.49	0.74/0.58/0.41	0.65/0.52/0.36	0.91/0.71/0.51
PECNET-QMC	1.62/1.09/0.78	0.68/0.50/0.35	0.88/0.65/0.46	0.72/0.52/0.36	0.62/0.45/0.30	0.91/0.64/0.45
SGCN-QMC	**1.23**/0.90/0.63	0.66/0.48/0.35	**0.78**/0.61/0.42	**0.63**/0.49/**0.33**	**0.49**/0.37/**0.25**	**0.76**/0.57/**0.39**
STGCNN-NPSN	1.48/1.03/0.73	0.88/0.64/0.49	0.88/0.66/0.48	0.83/0.60/0.45	0.73/0.52/0.39	0.96/0.69/0.51
PECNET-NPSN	1.60/1.08/0.77	0.63/**0.46/0.33**	0.88/0.64/0.45	0.70/0.51/0.34	0.56/0.41/0.27	0.87/0.62/0.43
SGCN-NPSN	**1.23/0.89/0.62**	**0.62/0.46/0.33**	0.79/**0.58/0.41**	0.66/**0.48/0.33**	0.50/**0.36/0.25**	**0.76/0.55/0.39**

Table 11. Possible explanations for different cases of agreement or disagreement between ES, EST, and ESS

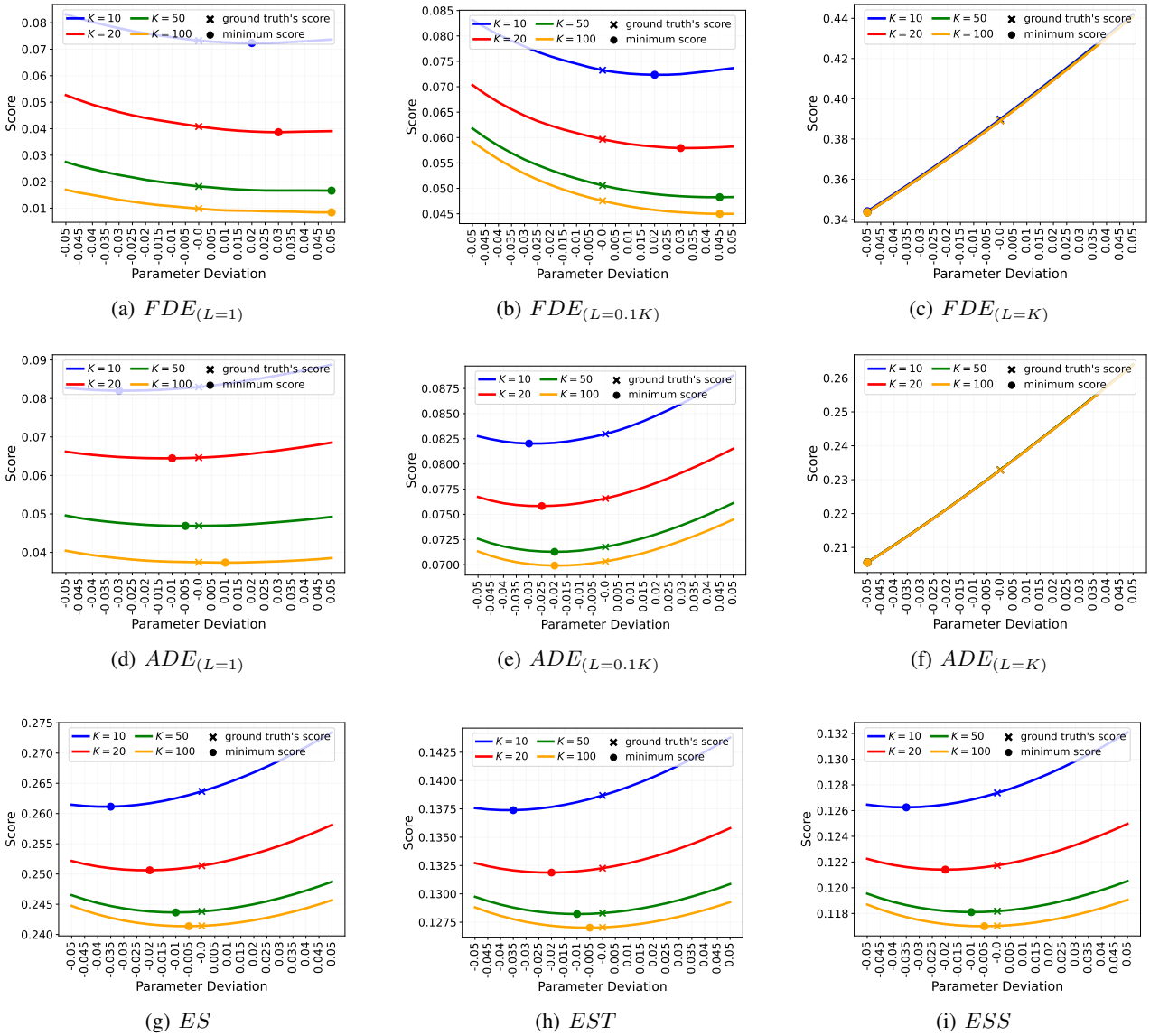
AGREEMENT	POSSIBLE EXPLANATION
ALL	Joint spatiotemporal aspects evaluated by ES make little difference.
NONE	Spatial and temporal aspects, individually as well as jointly, make a significant difference.
ES & EST	Temporal aspects dominate the joint evaluation by ES.
ES & ESS	Spatial aspects dominate the joint evaluation by ES.
EST & ESS	There are aspects captured by ES that EST and ESS are not able to capture, i.e., the interaction between space and time.

Overall, followed by the fact that the three scores ALL agree on the top-ranked model, we can conclude that the joint spatiotemporal aspects of the predictions are not that different on the ETH and HOTEL datasets. The same could be said about AVG but less strongly since there is a tie between the top two models SGCN-NPSN and SGCN-QMC. NONE agreeing, in other words, the disagreement between all three scores occurring on limited cases of the second-ranked model, under ZARA1, ZARA2 and AVG suggest that the spatial and temporal aspects are such that they individually or jointly can make a difference. Which score to rely on is subjective, depending on the downstream use case. Since we do not observe agreement between ES & EST, for the first- and second-best models, the temporal aspects do not seem to have affected the joint evaluation by ES. At the same time, the agreement between ES & ESS suggests that spatial aspects seem to dominate evaluation by ES for the top two models under AVG and less strongly under datasets ZARA1 and ZARA2. Lastly, numerous instances for the case EST & ESS suggest that there are aspects captured by ES that could not be captured by either EST or ESS. Since EST and ESS agree, these aspects must be related to the interaction between space and time. In summary, the analysis of these results highlights the use of EST and ESS alongside ES and underlies the importance of a joint evaluation with ES.

In Table 12 we provide a similar comparison between $minADE$, EST and ESS. There is more agreement between $minADE$ and its ES counterparts on the top-ranked model, while there is more disagreement on the second-based model. We attribute the former partly to the fact that $minADE$ is marginalized over the temporal dimension similar to the ESS and the latter to the lack of propriety. From the marginalization point of view, $minADE$ is comparable with ESS since both are averaged over the temporal steps. However, their ranking is considerably different, which we attribute to the fact that one is proper and the other is not.

Table 12. Reported values: minADE/EST/ESS, which are averages over instances within each dataset. AVG is the arithmetic average over all datasets. **Bold**: best model, underline: second best model. Baselines: *-MC.

	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
STGCNN-MC	1.10/1.02/0.75	0.86/0.83/0.61	0.80/0.72/0.52	0.53/0.57/0.41	0.48/0.49/0.36	0.75/0.73/0.53
PECNET-MC	1.07/1.11/0.79	0.39/0.52/0.36	0.56/0.66/0.46	0.45/0.54/0.37	0.33/0.47/0.31	0.56/0.66/0.46
SGCN-MC	1.00/0.93/0.66	0.53/0.55/0.39	0.67/0.63/0.44	0.51/0.50/0.34	0.42/0.39/0.26	0.63/0.60/0.42
STGCNN-QMC	1.03/1.04/0.72	0.52/0.74/0.55	0.63/0.68/0.49	0.53/0.58/0.41	0.50/0.52/0.36	0.64/0.71/0.51
PECNET-QMC	1.05/1.09/0.78	0.37/0.50/0.35	0.54/0.65/0.46	0.43/0.52/0.36	0.31/0.45/0.30	0.54/0.64/0.45
SGCN-QMC	0.81/0.90/0.63	0.31/0.48/0.35	0.56/0.61/0.42	0.45/0.49/**0.33**	0.36/0.37/**0.25**	0.50/0.57/**0.39**
STGCNN-NPSN	0.65/1.03/0.73	0.34/0.64/0.49	0.44/0.66/0.48	0.43/0.60/0.45	0.38/0.52/0.39	0.45/0.69/0.51
PECNET-NPSN	0.88/1.08/0.77	0.29/**0.46**/**0.33**	0.44/0.64/0.45	0.33/0.51/0.34	**0.25**/0.41/0.27	0.44/0.62/0.43
SGCN-NPSN	**0.59**/**0.89**/**0.62**	**0.25**/**0.46**/**0.33**	**0.39**/**0.58**/**0.41**	**0.32**/**0.48**/**0.33**	**0.25**/**0.36**/**0.25**	**0.36**/**0.55**/**0.39**



(a) $FDE_{(L=1)}$ (b) $FDE_{(L=0.1K)}$ (c) $FDE_{(L=K)}$

(d) $ADE_{(L=1)}$ (e) $ADE_{(L=0.1K)}$ (f) $ADE_{(L=K)}$

(g) ES (h) EST (i) ESS

Figure 6. Plot of different metrics across different predictions that deviate on the variance parameter. The lowest reported score versus the unbiased prediction is depicted in the circle and cross, respectively.