Neural Goodness-of-fit Test for Time Dependent Data

Aoran Zhang [1] Wenbin Zhou [1] Liyan Xie[2] Shixiang Zhu [1]

[1]Carnegie Mellon University [2]University of Minnesota

Overview

Problem Setup:

Consider we have two sequentially obtained or generated data sequences $D_0 = \{x_1^{(0)}, \ldots, x_{n_0}^{(0)}\}$ and $D_1 = \{x_1^{(1)}, \ldots, x_{n_1}^{(1)}\}$. D_0 is obtained from observations of a real-world process. D_1 is a synthetic data sequence generated by a learned time series model.

$\mathbb{P}^*$ denotes the true (but unknown) probability distribution governing the real-world time series.

$\widehat{\mathbb{P}}$ denotes the probability distribution learned by the time series model.

We aim to assess whether the learned time series model $\widehat{\mathbb{P}}$ accurately captures the underlying distribution of the real-world data. Formally, we aim to test the following hypotheses:

$$H_0 : \mathbb{P}^* = \widehat{\mathbb{P}} \quad \text{versus} \quad H_1 : \mathbb{P}^* \neq \widehat{\mathbb{P}}. \tag{1}$$

Difficulties: Mainstream methods to measure the quality of time series models are goodness-of-fit (GOF) tests. However, for general time series models, especially generative time series models:

(i) Parametric GOF Tests rely on comparing the parameters of models. When parameters are specified, **prior knowledge and assumptions are required**. When parameters are estimated, particularly with complex data, the **estimation process can be prone to inaccuracies**.

(ii) Nonparametric GOF Tests often employ distance-based method (*e.g.* Maximum Mean Discrepancy). These approaches typically **ignore time dependence**.

(iii) Quality Measurements of Generative Models rely on heuristic metrics, *e.g.* Fréchet Inception Distance for images and BLEU score for text, and **cannot be readily applied to time series data**.

Neural Representation of History Embeddings

Neural Ordinary Differential Equations (ODEs): for a continuous-time series $\{x(t), t \geq 0\}$, we define a low-dimensional *history embedding* $h(t)$. The evolution of the embedding is governed by:

$$\frac{\mathrm{d}h(t)}{\mathrm{d}t} = f(h(t), x(t)) \text{ and with Euler Approximation: } h_{i+1} = h_i + f(h_i, x_i)\Delta t_i,$$

where f is the update function over continuous time. h_i and h_{i+1} are the history embeddings at times t_i and t_{i+1}, and $\Delta t_i = t_{i+1} - t_i$ for n discrete observations $\{t_i\}_{i=1}^n$.

Inspired by the model, we parameterize the history embedding updates as

$$h_{i+1} = \phi(x_i, h_i; \theta),$$

using an embedding function $\phi(\cdot, \cdot; \theta)$ modeled as a neural network with θ as network's weights.

Consistency Assumption: The learned embedding function $\phi(\cdot, \cdot; \widehat{\theta})$ is consistent, meaning it approximates the true underlying embedding function as closely as necessary. Therefore, the learned embedding h_{i+1} captures all relevant information from x_i and h_i.

Implications: The history embeddings $\{h_i\}$ possess the **Markov and homogeneity** property, *i.e.*:

$$\mathbb{P}(h_i | h_{i-1}, \ldots, h_1) = \mathbb{P}(h_i | h_{i-1}) \quad \text{and} \quad \mathbb{P}(h_i | h_{i-1}) = \mathbb{P}(h_{i-1} | h_{i-2}).$$

Reformulation of Goodness-of-fit Test: The distributional behavior of the learned history embeddings can be fully characterized by their one-step *transition* density function $Q : \mathscr{H} \times \mathscr{H} \mapsto \mathbb{R}_{\geq 0}$. Specifically, for any subset $B \subseteq \mathscr{H}$ and for all i, we have

$$\mathbb{P}\{h_{i+1} \in B | h_i = h\} = \int_{h' \in B} Q(h, h') dh',$$

where $Q(h, \cdot)$ serves as the conditional probability density function of h_{i+1} given $h_i = h$. By leveraging the Markov property, the original GOF test can be reformulated as:

$$H_0 : Q^\star = \widehat{Q} \quad \text{versus} \quad H_1 : Q^\star \neq \widehat{Q}, \tag{2}$$

where $Q^\star$ denotes the true transition density function of the history embeddings derived from real data, and $\widehat{Q}$ denotes the transition density function derived from

Proposed Algorithm

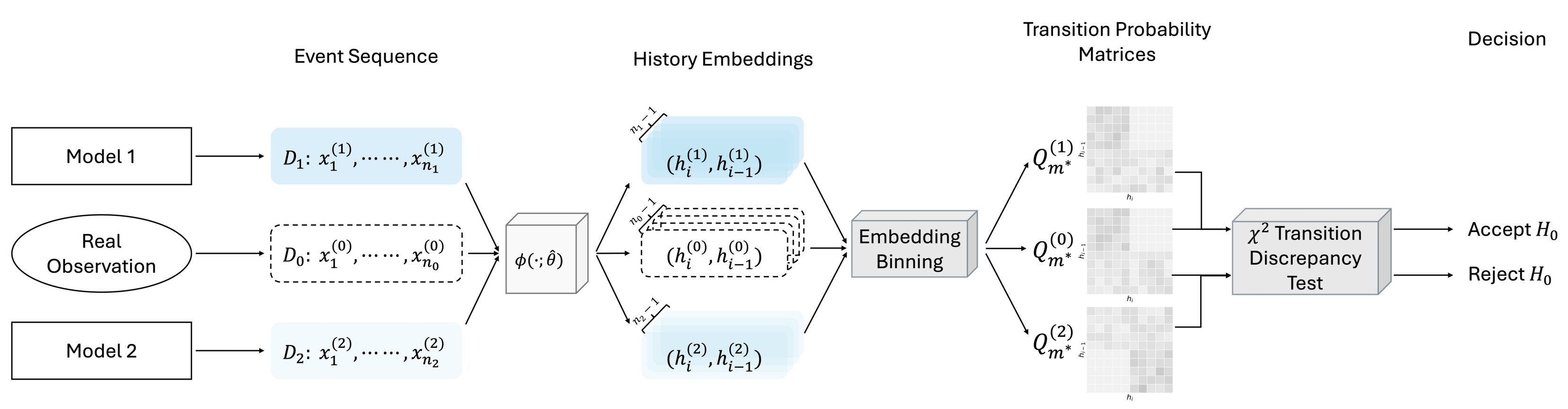


Figure 1. Overview of Testing Procedure

Step 0: Training and Extraction We begin by learning the embedding function ϕ using the real-world observations D_0. After training, we apply the learned embedding function to both datasets D_0 and D_1 and obtain their embedding sequences $\{h_i^{(0)}\}$ and $\{h_i^{(1)}\}$.

Step 1: Embedding Binning We partition the embedding space $\mathcal{H}$ into m bins of equal size, denoted as $\mathcal{H}_1, \ldots, \mathcal{H}_m$ to approximate Q with a discrete transition matrix.

1. A sequence of bin indices $\{y_i\}_{i=1}^n$ where $y_i = \sum_{u=1}^m u \cdot \mathbb{1}\{h_i \in \mathcal{H}_u\}$.
2. The transition count matrix $C_m = (c_{uv})_{u,v=1}^m$ where $c_{uv} = \sum_{i=1}^{n-1} \mathbb{1}\{h_i \in \mathcal{H}_u \text{ and } h_{i+1} \in \mathcal{H}_v\}$.
3. The empirical transition probability matrix $Q_m := (q_{uv})_{u,v=1}^m$ where
$$q_{uv} = \frac{c_{uv}}{\sum_{v'=1}^m c_{uv'}}.$$

We denote the empirical transition count and probability matrices obtained from D_0 as $C_m^{(0)}$ and Q_m^*, and the corresponding matrices from D_1 as $C_m^{(1)}$ and $\widehat{Q}_m$.

To determine the optimal number of bins m, we formulate the following optimization problem:

$$\max_{m \geq 1} \left\{ \|Q_m^\star - \widehat{Q}_m\|_F + \lambda \left(S(Q_m^\star) + S(\widehat{Q}_m) \right) \right\},$$

where $\|\cdot\|_F$ denotes the Frobenius norm, λ is a user-defined smoothing constraint, and $S(\cdot)$ is the smoothness measure of the transition matrix defined as:

$$S(Q_m) = -\sqrt{\sum_{u=2}^{m-1} \sum_{v=2}^{m-1} (\nabla^2 q_{uv})^2}.$$

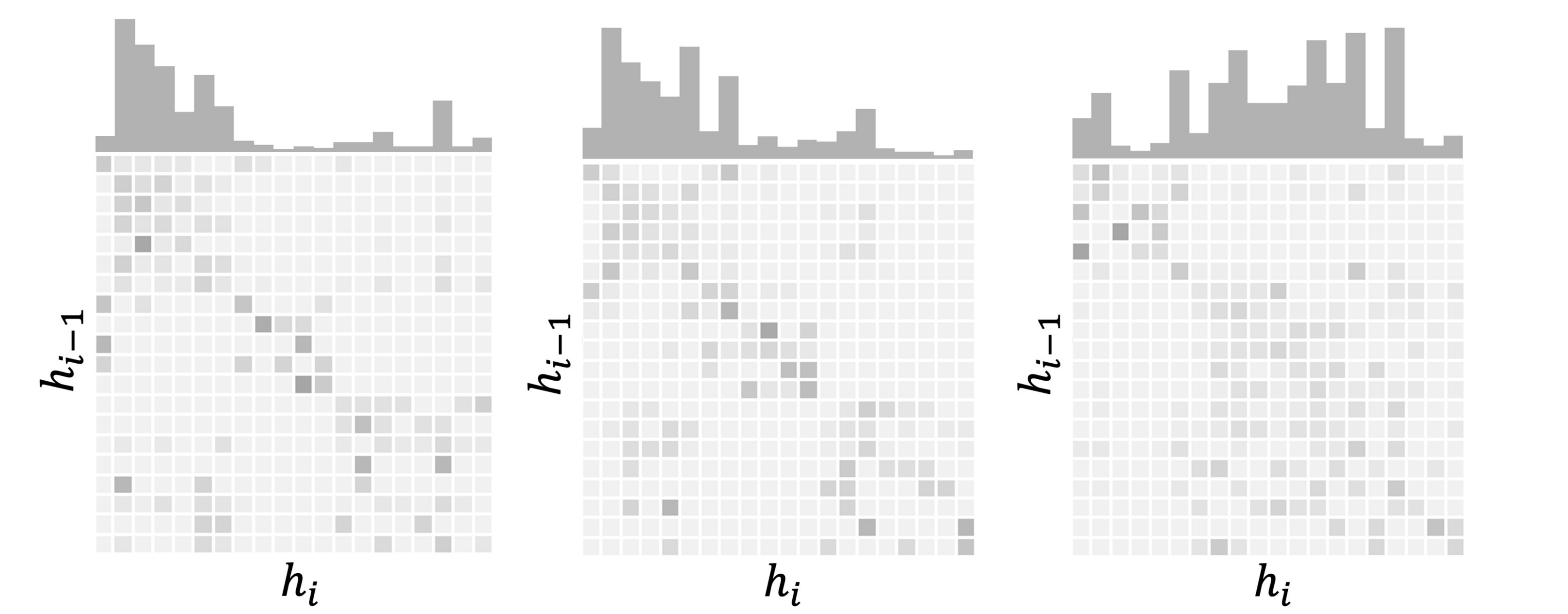
where $\nabla^2 q_{uv} = q_{u+1,v} + q_{u-1,v} + q_{u,v+1} + q_{u,v-1} - 4q_{uv}$ denotes the second derivatives.

Step 2: χ^2 Transition Discrepancy Test A chi-square test statistics W_m is given by

$$W_m = \sum_{u=1}^m \sum_{v=1}^m \frac{c_u^{(0)} c_u^{(1)}}{c_{uv}^{(0)} + c_{uv}^{(1)}} (q_{uv}^\star - \widehat{q}_{uv})^2 \tag{3}$$

where $c_u^{(k)} = \sum_{v=1}^m c_{uv}^{(k)}$ is the the total count of transitions from state u.

Results



(a) Real $Q^\star$ (b) $\widehat{Q}_1$ (0.99) (c) $\widehat{Q}_2$ (≈ 0)

Figure 2. Transition probability matrices of history embeddings Q from (a) real data, (b) data generated by Model 1, and (c) data generated by Model 2. Model 1 exhibits a better fit compared to Model 2, as evidenced by the closer resemblance between the histograms in (a) and (b). The number in the parentheses indicates the corresponding testing score.

Table 1. Testing accuracy on synthetic dataset with $\alpha = 0.05$. "–" indicates the method is not applicable.

Methods	Time Series						TPP					STPP				
$P^\star$ / $\widehat{P}$	Average	ARMA(2,1) / ARMA(2,1)	ARMA(2,1) / ARMA(2,2)	ARMA(2,2) / ARMA(2,2)	ARMA(2,1) / GARCH(1,1)	ARMA(2,2) / GARCH(1,1)	Average	SE / SE	SC / SC	SE / SC	SC / SE	Average	STD / STD	GAU / GAU	STD / GAU	GAU / STD
`EL`	0.37	0.08	0.33	0.23	0.76	0.64	–	–	–	–	–	–	–	–	–	
`PT-`Q_W	0.52	0.91	0.15	0.93	0.95	0.08	–	–	–	–	–	–	–	–	–	
`S-CvM`	0.44	0.98	0.08	0.97	0.02	0.04	–	–	–	–	–	–	–	–	–	
`Stein`	–	–	–	–	–	-	0.51	0.32	0.47	0.51	0.74	–	–	–	–	
`KSD`	–	–	–	–	–	-	0.56	0.66	0.78	0.72	0.08	–	–	–	–	
`MMD`	0.62	0.82	0.15	0.72	0.75	0.82	0.45	0.53	0.61	0.31	0.35	0.43	0.65	0.68	0.10	0.26
`EWD-2`	0.55	0.70	0.10	0.68	0.70	0.84	0.53	0.80	0.69	0.25	0.37	0.44	0.75	0.80	0.07	0.13
`EWD-4`	0.61	0.45	0.65	0.37	0.98	0.87	0.53	0.46	0.33	0.63	0.68	0.48	0.13	0.09	0.93	0.75
`EWD-6`	0.66	0.60	0.52	0.58	0.95	0.91	0.56	0.54	0.63	0.54	0.52	0.38	0.68	0.63	0.07	0.13
`EWD-8`	0.66	0.80	0.15	0.76	0.94	0.84	0.53	0.65	0.76	0.38	0.39	0.48	0.83	0.88	0.12	0.17
`EWD-10`	0.67	0.95	0.15	0.85	0.77	0.65	0.52	0.80	0.80	0.22	0.29	0.46	0.94	0.85	0.05	0.09
`Scott`	0.56	0.97	0.08	0.47	0.72	0.58	0.43	0.99	1	0.03	0.01	0.49	0.98	1	0.005	0.001
`RENAL`	**0.72**	0.71	0.60	0.65	0.92	0.95	**0.61**	0.64	0.62	0.58	0.56	**0.60**	0.70	0.57	0.67	0.44

Table 2. Testing accuracy on real data with $\alpha = 0.05$. For earthquake data, we consider two cases: TPP (time only) and STPP (spatio-temporal).

Model	Weather Time Series					Earthquake TPP					Earthquake STPP				
$P^\star$ / $\widehat{P}$	Average	PIT / PIT	SFO / SFO	SFO / PIT	PIT / SFO	Average	JP / JP	NC / NC	NC / JP	JP / NC	Average	JP / JP	JP / NC	NC / JP	NC / NC
`MMD`	0.52	1	0.98	0.02	0.01	0.37	0.24	0.47	0.42	0.34	0.51	0.15	0.05	0.98	0.97
`RENAL`	**0.66**	0.65	0.67	0.58	0.75	**0.57**	0.57	0.63	0.52	0.54	**0.62**	0.6	0.37	0.72	0.67

Remark: `RENAL` consistently achieves one of the highest Type I accuracies and best Type II accuracies with superior performance since:

(i) It exhibits the highest overall accuracy and balanced performance across all scenarios;

(ii) It requires few model specifications or assumptions and taks time-dependence into account;

(iii) It is applicable across all settings while some baseline methods are limited to certain cases.

References

[1] Patrick Billingsley.
Statistical methods in markov chains.
The annals of mathematical statistics, pages 12–40, 1961.