Génération de Séries Temporelles Synthétiques via un Temporal Structural Causal Model (TemporalMLPSCM) pour le Pré-entraînement d'un Foundation Model Tabulaire (TempTabFM)

Noé Amar — Neuralk AI

1. Introduction

Les modèles fondationnels (Foundation Models) ont récemment révolutionné le NLP et la vision grâce à l'accès à des corpus massifs et diversifiés (ex: CommonCrawl, LAION). En revanche, il n'existe **aucun équivalent pour les données tabulaires temporelles**, principalement à cause de :

- contraintes de confidentialité (finance, santé, industrie),
- fragmentation des sources de données,
- absence de benchmarks standardisés et accessibles.

L'objectif est donc de **construire un générateur synthétique fiable** permettant de créer une grande variété de séries temporelles tabulaires, afin de constituer un corpus de pré-entraînement pour un **Tabular Time Series Foundation Model**, que nous appelons **TempTabFM**.

Les Structural Causal Models (SCM) constituent une méthode rigoureuse de simulation de données complexes. Des travaux récents comme TabICL / TabPFN utilisent déjà un générateur SCM (MLPSCM), mais il est **purement statique**. **Notre contribution :** étendre le MLP-SCM afin de créer des données **temporelles, cohérentes et diversifiées** :

TemporalMLPSCM

Nous introduisons :

- une dynamique temporelle contrôlable (autoregression + périodicité),
- une exploration des régimes temporels via des hyperparamètres,
- un cadre d'évaluation basé sur la signature statistique des datasets.

2. Du MLPSCM au TemporalMLPSCM

2.1 Le générateur MLPSCM original (TabICL)

Le MLPSCM génère des données tabulaires statiques en trois étapes :

1. **Échantillonnage des causes** $C \in \mathbb{R}^{N \times d_{\text{cause}}}$ via `XSampler`, mélange flexible de distributions.
2. **Propagation causale via MLP** Un MLP profond transforme les causes, couche par couche, en un vecteur latent h.
3. **Sélection des features** Des indices idx_X, idx_y sélectionnent les features et la variable cible. Cela joue le rôle d'un DAG implicite.

Ce modèle génère des datasets diversifiés, mais **aucune structure temporelle** n'est conservée :

$$X_t \to X_{t+1} \quad \Rightarrow \quad \text{impossible d'entraîner un modèle de forecasting.}$$

2.2 Extension temporelle : TemporalMLPSCM

Chaque échantillon devient une trajectoire $(X_t, y_t)_{t=1..T}$. Chaque couche k possède désormais un état temporel $h_{k,t}$.

Équation générale (par couche et pas de temps):

$$h_{k,t} = \text{MLP}_k(\text{inp}_{k,t}) + \alpha \cdot h_{k,t-1} + \beta \cdot h_{k,t-p} + \varepsilon_t \quad \text{avec } p = \text{période}$$

Logique :

- α : mémoire à court terme (autoregression)
- β, p : mémoire à long terme / saisonnalité
- ϵ_t : bruit gaussien pour diversité

Enfin, les états sont concaténés :

$$z_t = [h_{1,t}, \ldots, h_{L,t}], \quad X_t = z_t[\text{idx}_X], \quad y_t = z_t[\text{idx}_y].$$

2.3 Deux régimes temporels étudiés

- **Stratégie A : Autoregression pure** ($\beta = 0$) $\to$ ACF décroissante, mémoire de type AR(1)
- **Stratégie B : Autoregression + Périodicité** ($\beta > 0$) $\to$ Pics d'ACF aux multiples de p, fréquences claires dans le spectre FFT

Cette combinaison permet de couvrir de nombreux comportements temporels — similaires à ceux de la finance, de la météo ou de la consommation en grande distribution.

3. Cadre d'Évaluation

3.1 Signature d'un dataset

Pour chaque dataset $X \in \mathbb{R}^{T \times D}$, nous extrayons :

(1) Signature marginale (par feature)

$$\{\text{moyenne}, \text{std}, \text{skew}, \text{kurtosis}, q_{10}, q_{50}, q_{90}\}$$

(2) Signature temporelle

- Autocorrélation ACF(lag)
- Temps de décohérence : $\min(l : |\text{ACF}(l)| < e^{-1})$
- Fréquence dominante via periodogramme

(3) Structure multivariée

$$\Sigma = \text{corr}(X) \quad \Rightarrow \quad \text{structure} = \{\Sigma_{i,j}\}_{i<j}$$

L'ensemble constitue un **vecteur de signature**, clé de l'analyse.

3.2 Distances inter-datasets

Pour un ensemble $\{S_1, \ldots, S_n\}$ de signatures :

$$D_{i,j} = \|S_i - S_j\|_2 \quad \Rightarrow \quad \text{plus } D \text{ est élevé, plus les mondes générés sont différents.}$$

Nous calculons :

$$\{\text{mean}, \text{min}, \text{max}, \text{std}\}$$

sur les matrices de distances :

- Global (signature complète)
- Marginal / Temporel / Structure
- Corrélations uniquement

3.3 Visualisation

Nous utilisons :

- heatmaps de distances,
- histogrammes,

Cela permet de **cartographier l'espace des mondes générés**.

4. Perspectives et Limitations

- Architecture MLP uniquement (pas encore Transformer causal)
- Clipping / normalisation peut nuire à la dynamique
- Pas encore de test downstream (pré-entraînement réel TempTabFM)

5. Conclusion

TemporalMLPSCM propose une extension réaliste et contrôlable du MLPSCM statique. Le modèle introduit une **vraie dynamique temporelle**, tout en conservant l'esprit causal implicite de TabICL.

Contribution majeure : $\rightarrow$ Un moteur de génération **synthetic-but-plausible** pour construire un **corpus de pré-entraînement massif**, première étape vers le development de **TempTabFM**, un foundation model tabulaire pour les données temporelles.