
DTU

Generative Latent Class Choice Models: Enriching Smart Card Data with Travel Surveys using Variational Auto-Encoders

Georges Sfeir, Filipe Rodrigues, Francisco Camara Pereira

Technical University of Denmark DTU | MLSM Machine Learning for Smart Mobility group http://mlsm.man.dtu.dk

This project has received funding from the Horizon Europe Framework Programme (HORIZON) under the Marie Skłodowska-Curie grant agreement No. 101063801

ICMC 2024, Puerto Varas, Chile

Outline

- Introduction
- VAE-LCCM
- Case Study and Results
- On-Going Work

Introduction

Motivation

- Researchers rely on travel surveys, whether RP or SP, to develop travel choice models
- Travel surveys provide detailed data on individuals' choices, preferences, socio-economic profiles, and attitudes. However…
- They might suffer from a considerable amount of missing data

- Passive data collection methods (e.g., smart card data) offer a larger volume of observations but lack information on travel preferences, socio-economic characteristics, and attitudes

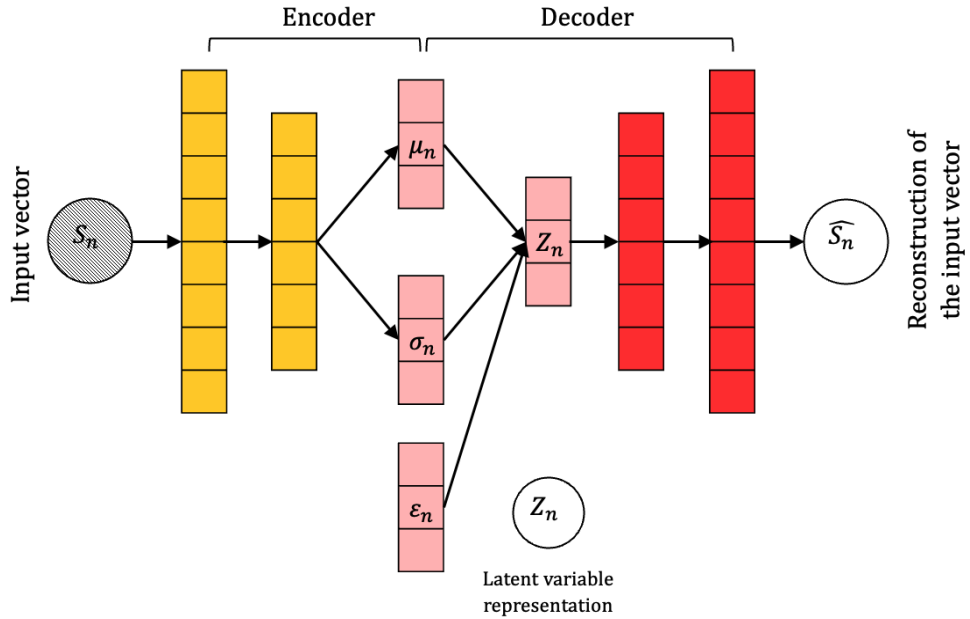
Goals

We will try to integrate a generative model (VAE) with a Latent Class Choice Model (LCCM) to have a behavioral model that can:

- Impute missing data (e.g., socio-economics, attitudinal indicators, etc.)
- Generate synthetic data
- Improve LCCM's goodness-of-fit and out-of-sample generalization without total loss of interpretability
- Improve representation of heterogeneity in tastes and preferences

Variational Auto-Encoders (VAEs)

- A generative model that consists of two neural networks called **encoder** and **decoder**
  1. The **encoder** transforms an input vector S_n into a distribution over a **latent space** Z_n
  2. The **decoder** takes a sample from a **latent space distribution** and reconstructs the input vector



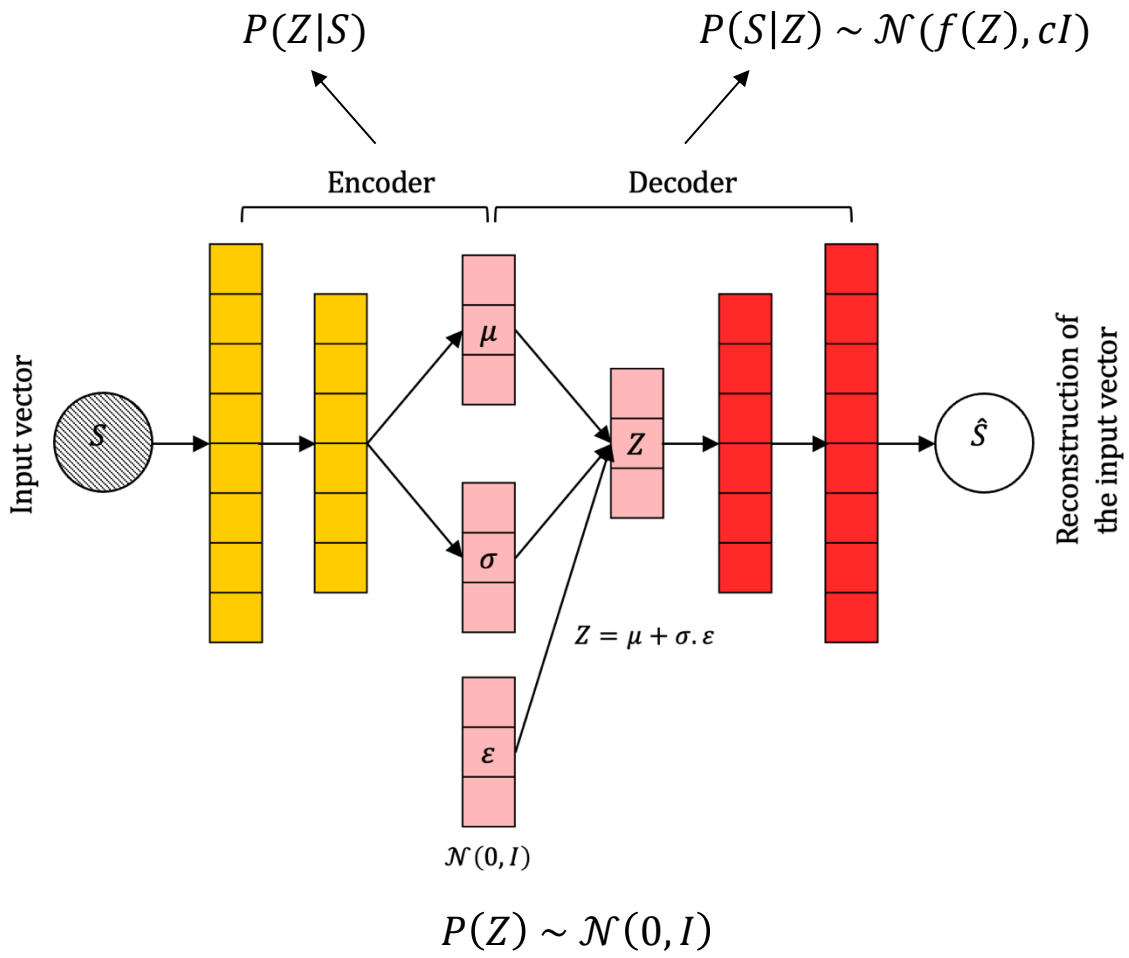
Loss function:

- Reconstruction loss
- Regularization term

- The “variational” refers to the use of variational inference for training the model

Probabilistic Framework

- The goal is to learn a latent space distribution that could have generated the data



$$P(Z|S) = \frac{P(S|Z)P(Z)}{P(S)}$$

- $P(Z|S)$ is intractable due to the denominator P(S)
- Approximation techniques such as variational inference (VI) are required

Variational Inference

- The true posterior distribution $P(Z|S)$ is approximated by considering a family of tractable distributions $q_S(Z) \sim \mathcal{N}(g(S), h(S))$
- Then the parameters of g and h functions are optimized by minimizing the KL-divergence between the approximation $q_S(Z)$ and the true posterior $P(Z|S)$
- KL-divergence:

$$KL\big(q_S(Z)||P(Z|S)\big) = \int q_S(Z) \log \frac{q_S(Z)}{P(Z|S)}\, dZ$$

- KL- divergence cannot be directly minimized, instead we maximize the ELBO:

$$ELBO = E[\log P(S|Z)] - KL(q_S(Z)||P(Z))$$

Applications of VAEs

- Data generation/imputation (e.g., text, image, speech generation)
- Dimensionality reduction (similar to PCA, can learn a lower-dimensional representation of data)
- Semi-supervised learning (the model is trained on both labeled and unlabeled data – independent and dependent variables)

VAE-LCCM

Graphical Representation of LCCM

N: number of individuals

K: number of classes

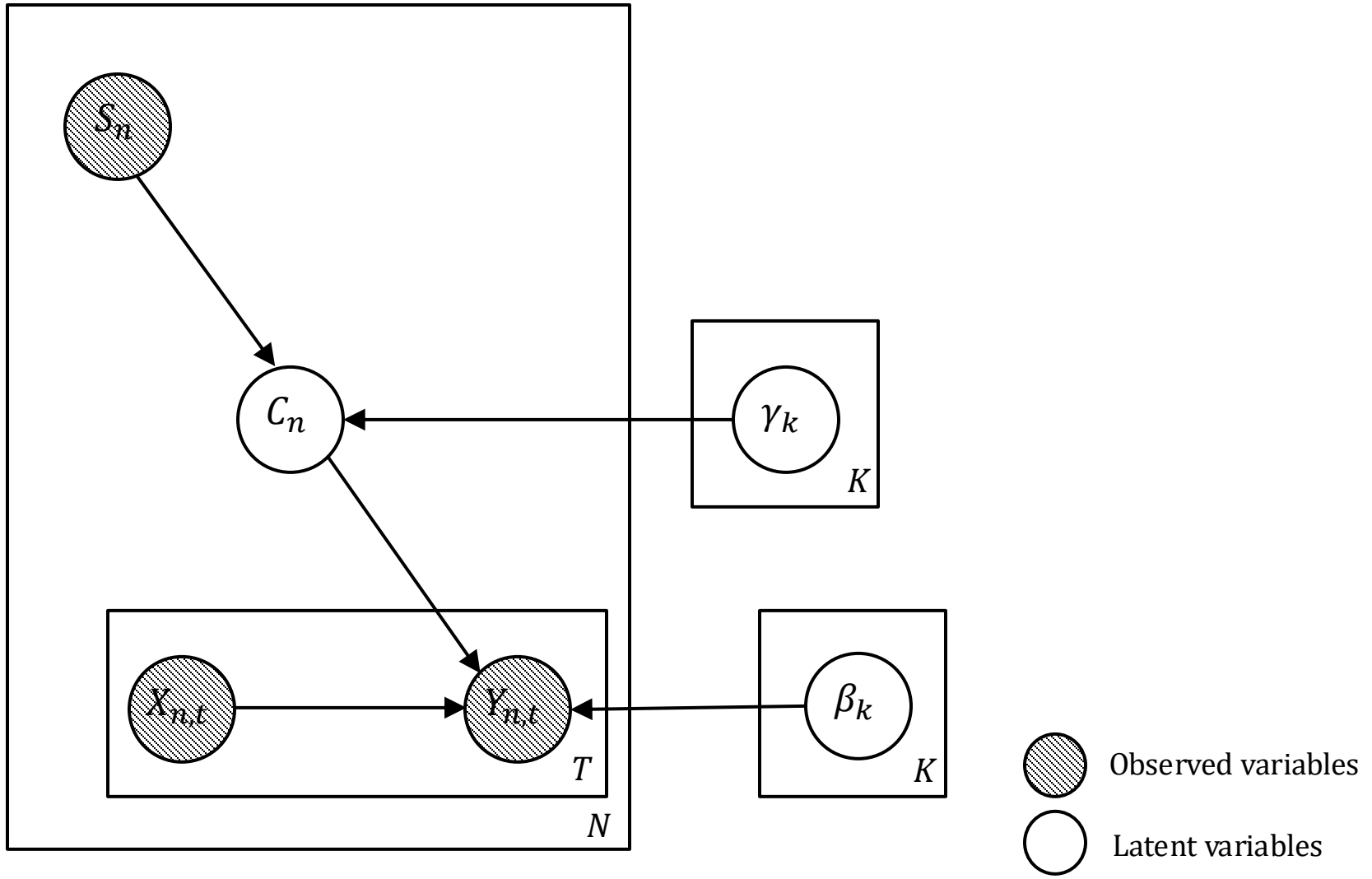
T: number of observations per ind.

Class Membership:

- S_n: socio-economic characteristics of individual n
- C_n: latent class assignment
- γ_k: class assignment parameters

Choice model:

- $Y_{n,t}$: observed choices of individual n at time t
- $X_{n,t}$: corresponding attributes
- β_k: corresponding parameters



VAE-LCCM

VAE:

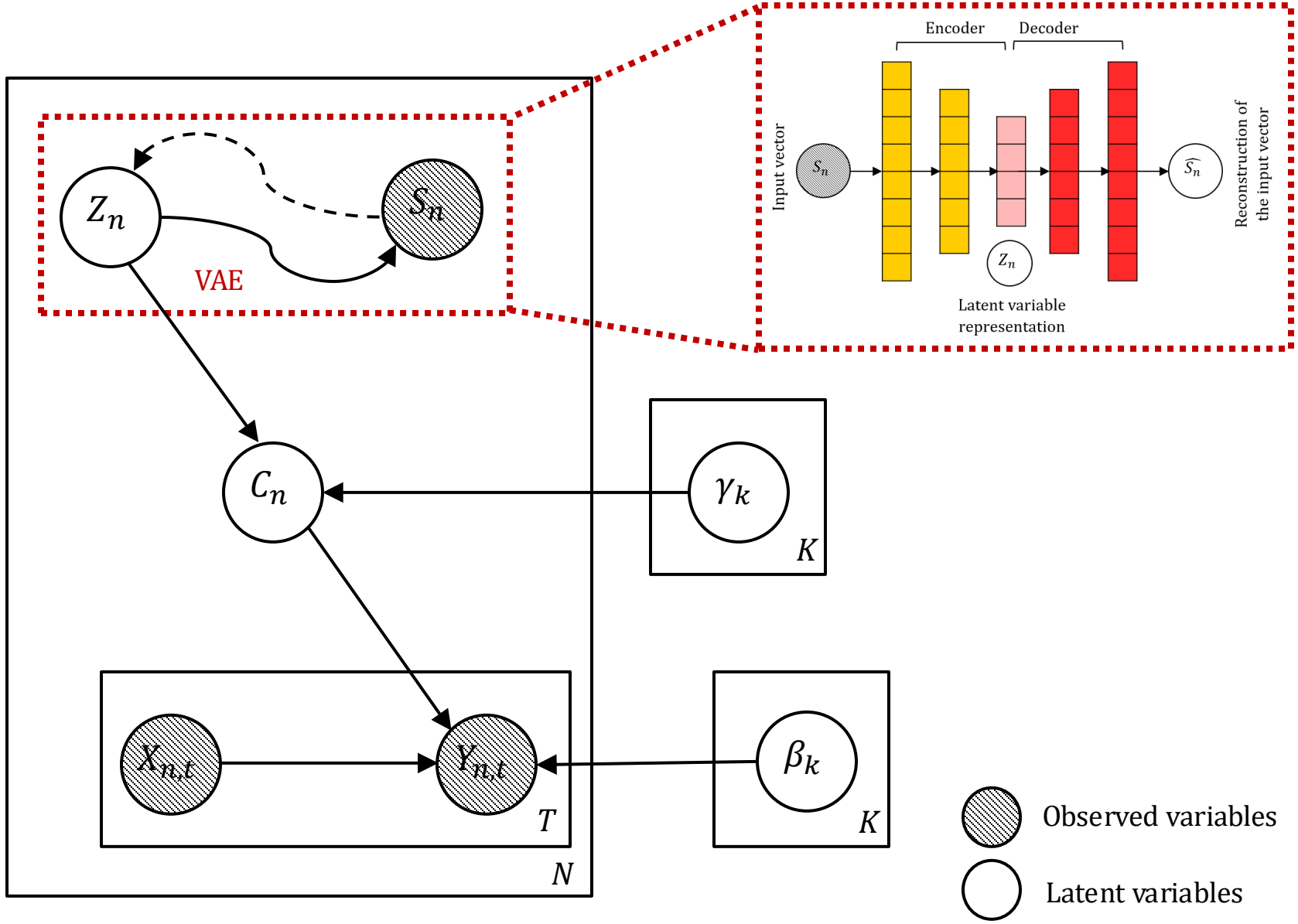
- S_n: socio-economic characteristics of individual n
- Z_n: VAE latent representation of individual n

Class Membership:

- Z_n: VAE latent representation
- C_n: latent class assignment
- γ_k: class assignment parameters

Choice model:

- $Y_{n,t}$: observed choices of individual n at time t
- $X_{n,t}$: corresponding attributes
- β_k: corresponding parameters



VAE-LCCM

The generative process of VAE-LCCM can be summarized as follows:

1. For each latent class $k \in \{1, \dots, K-1\}$
   a. Draw class-assignment parameters $\gamma_k \sim \mathcal{N}(\mu, \sigma)$
2. For each latent class $k \in \{1, \dots, K-1\}$
   a. Draw taste parameters $\beta_k \sim \mathcal{N}(\mu, \sigma)$
3. For each decision-maker $n \in \{1, \dots, N\}$
   a. Draw base distribution sample $\boldsymbol{Z_n \sim \mathcal{N}(0, 1)}$
   b. Draw socio-demographics $\boldsymbol{S_n \sim P(S_n | f_{enc}(Z_n))}$
   c. Draw latent class assignment $\boldsymbol{C_n \sim MNL(\gamma_k, Z_n)}$
   d. For each choice occasion $t \in \{1, \dots, T\}$
      i. Draw observed choice $Y_{nt} \sim MNL(\beta_k, X_{nt})$

Considering the joint distribution of the model and the variational distribution:

$$ELBO = \underbrace{E[\log P(Y|C)] - KL(q_Y(C)||P(C))}_{\text{LCCM}} + \underbrace{E[\log P(S|Z)] - KL(q_S(Z)||P(Z))}_{\text{VAE}}$$

Case Study and Results

TU Data

- TU Data is the national travel survey in Denmark
- Modes: walk, bike, car, public transport
- Data used for estimation (train set):
  - Year: 2022
  - Number of individuals: 5,048
  - Number of trips: 7,915
- Data used for evaluation (test set)
  - Year: 2021
  - Number of individuals: 4,849
  - Number of trips: 7,606

<u>Model Specification</u>

- Attributes: Cost, Time, and Alternative-Specific Constants
- Socio-economics: 12 variables with 57 categories

VAE, LCCM, and VAE-LCCM

- Several VAE architectures were tested:
  - Encoder: ANN architectures consisting of 1 or 2 hidden layers with 25, 50, 100; 50-25; or 100-50 neurons, respectively
  - Decoder: mirrored architecture of the encoder
  - Latent space dimensionality: 5, 10, 25
- The model with the best average reconstruction loss on the test set was selected:
  - 1 hidden layer, 100 neurons, Z= 25

- LCCM: models with 2 and 3 classes were estimated (increasing the number of classes beyond 3 resulted in numerical issues)

- VAE-LCCM: models with 2 to 5 classes were estimated using the best VAE architecture

VAE-LCCM vs. LCCM

	VAE-LCCM (K=2)	VAE-LCCM (K=3)	VAE-LCCM (K=4)	VAE-LCCM (K=5)	LCCM (K=2)	LCCM (K=3)
LL	-3,753.41	-2,959.20	-2,475.13	**-2,460.84**	-4,986.45	-4,900.78
Nb of par.	39	71	103	135	58	109
AIC	7,584.83	6,060.40	**5,156.25**	5,191.67	10,088.89	10,019.56
BIC	7,856.91	6,555.74	**5,874.83**	6,133.50	10,493.53	10,780.01

- Better goodness-of-fit and potentially better representation of heterogeneity
- Possible reasons VAE-LCCM can estimate more classes:
  - Non-linearity introduced by the neural networks of the VAE?
  - Dimensionality reduction of the latent space by the VAE?
  - Parameters initialization?

VAE-LCCM vs. LCCM

	VAE-LCCM (K=2)	VAE-LCCM (K=3)	VAE-LCCM (K=4)	VAE-LCCM (K=5)	VAE
Av. Recon Loss (Test Set)	9.72	**9.71**	9.77	9.76	9.96
μ (distances) (Test Set)	0.61	**0.59**	0.60	0.63	0.70
σ (distances) (Test Set)	0.74	**0.74**	0.74	0.74	0.76

- Data imputation and synthetic data generation capabilities

Conclusion

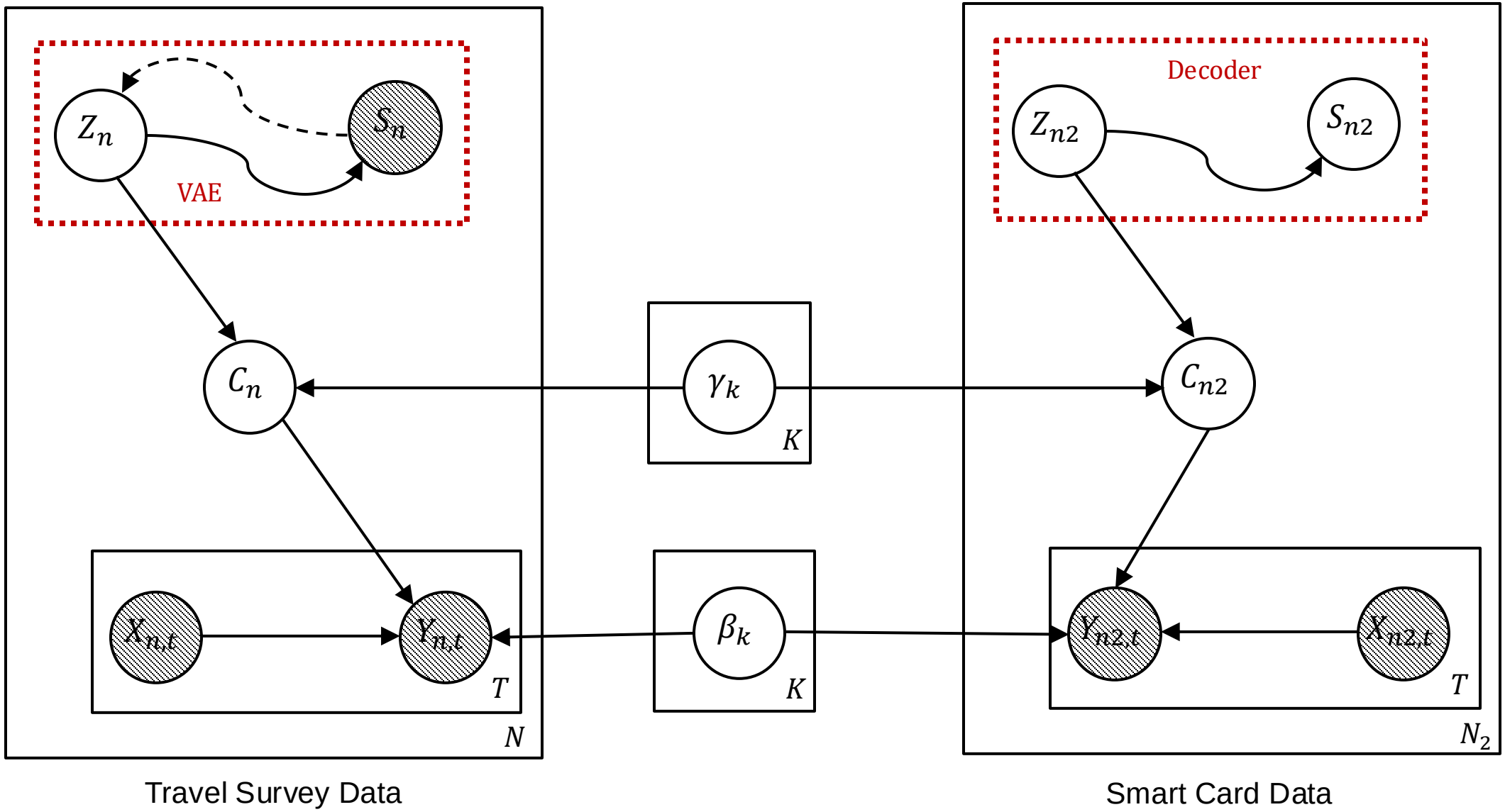
- Proposed a generative choice model (VAE-LCCM):
  - Generate synthetic data and impute missing data
  - Improve goodness-of-fit and out-of-sample generalization without total loss of interpretability
  - Potentially better representation of heterogeneity (more classes are estimated)

On-Going Work

On-Going Work

- Comparing VAE-LCCM with other benchmarks (e.g., PCA, other data imputation techniques)
- Investigating the issue of estimating more classes with VAE-LCCM compared to traditional LCCM (non-linearity, latent space dimensionality, parameters initialization)
- Interpretation of the latent classes

VAE-LCCM (Step 1 & 2)



Generative Latent Class Choice Models: Enriching Smart Card Data with Travel Surveys using Variational Auto-Encoders

Georges Sfeir, Filipe Rodrigues, Francisco Camara Pereira

Technical University of Denmark DTU | MLSM Machine Learning for Smart Mobility group http://mlsm.man.dtu.dk

This project has received funding from the Horizon Europe Framework Programme (HORIZON) under the Marie Skłodowska-Curie grant agreement No. 101063801

Thank you!