Latent Factor Discovery in Markov Processes through Optimal Transport

Nhi Pham, Prof. Esteban G. Tabak

New York University

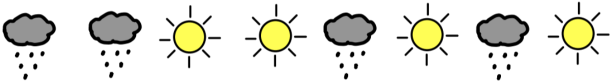
October 23, 2020

Agenda

1. Markov processes, hidden Markov model and its assumptions
2. Hidden Markov model in the context of Optimal Transport
3. The technical specification of the problem
4. Existing factor
5. Latent factor discovery

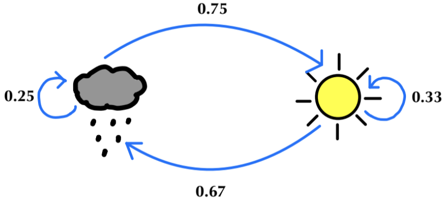
Markov Processes

Markov chains: a model that indicates the probabilities of sequences of events, in which the probability of each event depends only on the state in the previous event.



e.g. Markov chains of weather

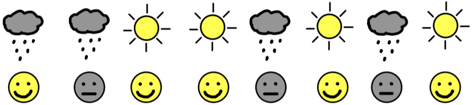
- weather states $S = \{\text{rainy}, \text{sunny}\}$
- sequence of observations over 8 days



Hidden Markov Model (HMM)

- States can not be observed directly but only some probabilistic function of those states.
- HMM allows us to consider **hidden states** z_i as (causal) *factors* of the **observations** x_i.

Question: instead of direct observations of the weather, if we only know about the records of moods of a person in 8 days, what can we say about the weather where the person lives?



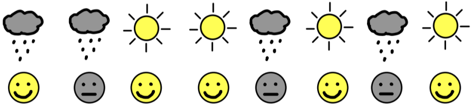
HMM assumptions

Markov assumption: the probability of being in the current state depends only on the previous state.

$$P(z_i | z_1, z_2, ..., z_{i-1}) = P(z_t | z_{i-1})$$

Output independence: the probability of an observation x_i depends only on the immediate hidden state z_i

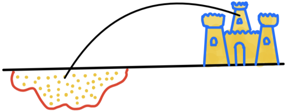
$$P(x_i | z_1, z_2, ..., z_{i-1}, z_i) = P(x_i | z_i)$$



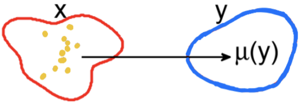
HMM in the context of Optimal Transport

- Propose a different conceptual and computational framework for the HMM problem based on the mathematical theory of optimal transport.
- With two HMM assumptions, seek to understand and quantify dependence between:
  - observations $\vec{x}$ and hidden states $\vec{z}$
  - current hidden state z_i and previous hidden state z_{i-1}

  via optimal transport.



(a) The "laziest" way to build a sand castle

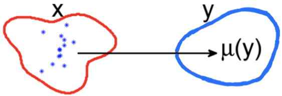


(b) The optimal way to transport $\rho(x|z)$ to $\mu(y)$

A more specific explanation

We treat z_i as a factor of x_i, and z_{i-1} as a factor of z_i:

- For each observation x_i, we seek to remove the variability attributable to the corresponding hidden state z_i, namely seeking optimal maps $Y(x_i, z_i)$ transforming x_i to y_i, where y_i is independent of z_i.
- For each hidden state z_i, we seek to remove the variability attributable to z_{i-1}, namely seeking optimal maps $W(z_i, z_{i-1})$ transforming z_i to w_i, independent of z_{i-1}.

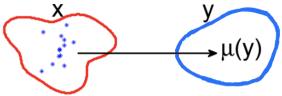


Into technical details

Given a conditional probability density $\rho(x|z)$ (estimated via a set of samples (x_i, z_i)), we seek a z-dependent and minimally distorted map

$$x \rightarrow y, \quad y = Y(x,z)$$

pushing forward $\rho(x|z)$ to z-independent representative distribution $\mu(y)$ – **the barycenter**.



This yields a **Wasserstein barycenter problem:**

$$\mu, Y = \arg\min \int_z \left(\int c\,(x,y)\,\rho(x|z)\;dx \right) \nu(z)\;dz,$$

$$y = Y(x;z) \sim \mu, \quad c(x,y) \text{ such as } \frac{1}{2}\|x-y\|^2$$

Introduction of the Primal and the Dual problems

We start with Kantorovich relaxation from map $Y(x;z)$ pushing forward $\rho(x|z)$ to μ to joint distributions $\pi(x,y|z)$:

$$\min_{\mu,\pi} \int \left(\int c(x,y)\pi(x,y|z)dxdy \right) \nu(z)dz$$

$$\text{s.t } \forall z, \int \pi(x,y|z)dy = \rho(x|z) \qquad \int \pi(x,y|z)dx = \mu(y)$$

The corresponding Lagrange Dual Problem:

$$\max_{\phi,\psi} \ \phi(x,z)\rho(x|z)dx\, \nu(z)dz$$

$$\text{s.t} \quad \phi(x,z) + \psi(y,z) \leq c(x,y) \qquad \forall y \int \psi(y,z)\nu(z)dz \geq 0$$

More into the Dual - Candidate Map

Define $c(x, y) = \frac{1}{2}\|x - y\|^2$, the first constraint of the dual becomes

$$\psi(y, x) \leq \frac{1}{2}\|x - y\|^2 - \phi(x, z)$$

The solution needs to satisfy

$$y = x - \nabla_x \phi(x, z)$$

In applications, we replace the conditional distributions $\rho(x|z)$, underlying distributions $\rho(x)$ and $\nu(z)$ by the data that consists of a finite set of samples (x_i, z_i).

$$y_i = x_i - \nabla_x \phi(x_i, z_i)$$

Rigid z-dependent translations linear in factors

With restriction to rigid transformation linear in factors, candidate maps for 2 factors in HMM (z_t as a factor of x_t and z_{t-1} as a factor of z_t):

$$Y(x_i, z_i) = x_i - (\alpha z_i + \beta)$$

$$W(z_i, z_{i-1}) = z_i - (\gamma z_{i-1} + \delta)$$

- **Existing factor:** given the sequence of states z, filter the variability attributable to z from x.
- **Factor discovery:** the sequence of states z not given, seek to minimize the variability in terms of unknown factors.

For each case, we need to determine the optimal $\alpha, \beta, \gamma, \delta$.

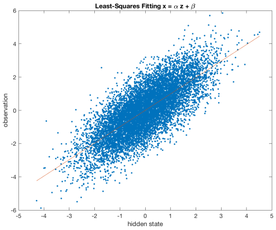
Existing Factor

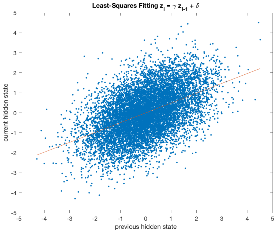
- In real world, data often includes known factors z in addition to x.
- We seek to filter the variability attributable to z from the original data with minimal distortion.
- Applications: Remove confouding factors in datasets, i.e in biostatistics, medical diagnosis, weather forecast,...

Existing Factor - An Example

10000 samples of (x_i, z_i) are generated:

- $z_{i+1} = \gamma_0 z_i + \delta_0$, where $\gamma_0 \leq 1$ is randomly chosen, and δ_0 is the noise drawn from the Gaussian distribution (0, 1).
- x_i is generated from the Gaussian distribution $(z_i, 1)$.

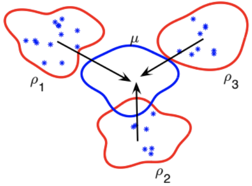




Latent Factor Discovery in theory

If z is not given, or if all known factors are filtered, we only have the unlabeled data $\rho(x)$.

We seek the sequences of hidden states $\vec{z}$ that best explains the variability – minimize the variability of the barycenters over all maps Y and W.



Latent Factor Discovery in applications

In terms of samples, this is equivalent to seeking the assignment z_i that minimizes some combination of the variance of y and of w, namely

$$\min_{\{\vec{z}\}} c_1 Var\left(\{y_i = Y(x_i, z_i\}\right) + c_2 Var\left(\{w_i = W(z_i, z_{i-1}\}\right)$$

with $i \in \{1, ..., m\}$

In this task, we need to find the optimal $\vec{z}, \alpha, \beta, \gamma,$ and δ.

Conclusion and Future Work

- Optimal transport, specifically the Wasserstein barycenter problem, provides a different conceptual and computational framework to look into time series models (hidden Markov model).
- Multiple uses: consolidation of databases, weather forecast, medical diagnosis, etc.
- Future works include developing a methodology to determine the continuous hidden variables $\vec{z}$ and extension to higher order transformation.