
Y. Park, C. Carvalho and J. Ghosh
AISTATS 2014

Review by Esther Salazar

October 9, 2014
Overview

- **Data**: Analysis of categorical time series with application from marketing research to healthcare analytics. Examples in healthcare domain:
  - Patient’s diagnosis history such as cardiac conditions (MIMIC-II dataset with publicly 25,328 intensive care unit patient stays)
  - Drug purchase from pharmaceutical companies (*Novartis*, *Teva*, others) for patients (Medicare part-D claim records, 2007-2010, 5% US population)

- **Model**: *Latent vector autoregressive models categorical time series* (LAVA-Cat) model

- Inference procedures for this model, such as *brute-force particle filter* of the EM algorithm, fails to estimate the maximum likelihood parameters due to multiple local optima of the log-likelihood

- The paper proposed two techniques: *asymptotic mean regularization* and *low-resolution augmentation*, do not require any additional parameter tuning and could be implemented by modifying the brute-force EM
Statistical models for categorical time series

These models fall into two classes:

1. Fully observation-based models
   - Mixture transition distribution model (Raftery, 1985)
   - Markovian regression model (Kaufman, 1987)
   - Discrete autoregressive moving average (DARMA), (Jacobs and Lewis, 1983)

2. Latent variable models
   - Hidden Markov models (HMM) with discrete latent variables (Zucchini and MacDonald, 2009)
   - State-space model (SSM) with continuous latent variables (Zhen and Basawa, 2009) using a latent regression model

This paper focuses on a latent vector autoregressive model for categorical time series with a VAR process for the continuous latent variables

This approach has been less popular than HMM and SSM because the continuous latent variables are difficult to reconstruct from categorical observations
Latent Vector-Autoregressive model for Categorical TS

LAVA-Cat

(Latent VAR) \[ x_t = c + \Phi x_{t-1} + \epsilon_t \] (1)

(Observation) \[ p(y_t = k | x_t) = f_k(A_k x_t) \] (2)

(Noise model) \[ \epsilon_t \sim \mathcal{N}(0, \Sigma) \] (3)

where \( y = [y_t]_{t=1}^T \) with \( y_t \in \{1, 2, \ldots, K\} \), \( p(y_t = k | x_t) \) is the probability that the \( k \)th category is observed at time \( t \)

Dimensions: \( c, x_t, \epsilon_t \in \mathbb{R}^{(K-1)}; \quad \Phi, \Sigma \in \mathbb{R}^{(K-1) \times (K-1)} \)

The link function \( f_k(A_k x_t) \) connects a real vector to a categorical value

The goal of this paper is to estimate the maximum likelihood parameters

\[ \theta = \{c, \Phi\} \]
Challenges to estimate the parameters

- Unlike continuous time series, categorical time series contain only finite bits of information.
- The reconstruction of continuous latent variables suffers from a low signal-to-noise ratio.
- The noise reconstruction increases the uncertainty of the estimated parameters.
- EM algorithm becomes susceptible to various factors such as noisy reconstruction and multiple local optima.
The authors focus on a multinomial logistic (softmax) link function

\[ f_k(x_t) = \begin{cases} 
\frac{\exp(x_{tk})}{h(x_t)} & k \in 1, 2, \ldots, (K - 1) \\
1/h(x_t) & k = K 
\end{cases} \]

where \( h(x_t) = \sum_{l=1}^{K-1} \exp(x_{tl}) + 1 \). Also, \( p(y_t = K) \) is the reference probability for the other categorical outcomes.

The log-likelihood of the LAVA-Cat model is:

\[
\max_{\theta} \log p_\theta(y) = \max_{\theta} \log \int_X p_\theta(y, X) dX \\
= \max_{\theta} \log \int_X \prod_t f(y_t \mid x_t)p_\theta(x_t \mid x_{t-1}) dX
\]
They derive the lower bound of the log-likelihood and maximize. Using Jensen’s inequality:

$$\log \int_X p_\theta(y, X) dX \geq \int_X q(X) \log \frac{p_\theta(y, X)}{q(X)} dX$$

The lower-bound is maximized by iteratively solving two sub-problems:

$$q = \arg \max_q \int q \log (p_\theta / q)$$

$$\theta = \arg \max_\theta \int q \log (p_\theta / q)$$

- The first maximization problem has a closed-form solution $q(X) = p_\theta(X | y)$
- An approximation of the target distribution is done via particle methods. Then, the second maximization step is

$$\max_\theta \int_X q(X) \log p_\theta(y, X) dX$$

$$= \max_\theta \int_X q(X) \log \prod_t f(y_t | x_t)p_\theta(x_t | x_{t-1}) dX$$

$$\approx \max_\theta \sum_{i,t} w_t^{(i)}(\log f(y_t | x_t^{(i)}) + \log p_\theta(x_t^{(i)} | x_{t-1}^{(i)}))$$
Although the BPEM algorithm is simple, it is not applicable in practice:

1. The algorithm stores at least $2 \times T \times P$ particles and weights
2. It does not fully satisfy the convergence requirements of the EM

These issues are much more noticeable in practice for categorical time series
The authors propose three novel techniques to address the scalability and stability issues of the BPEM:

1. Asymptotic mean regularization
2. Pseudo-Bayesian update
3. Low-resolution augmentation

Then, they combine these three components to obtain the: Low-resolution augmented Asymptotic Moment Regularized EM (LAMORE)
1. Asymptotic mean regularization

The asymptotic mean regularization is a regularization technique that utilizes additional stationarity information. This is a natural extension of the likelihood function for stationary time series

- Joint log-likelihood including a stationarity indicator

\[
\max_{\theta} \log p_\theta(y, I_s = 1) = \max_{\theta} \left( \log p_\theta(I_s = 1 \mid y) + \log p_\theta(y) \right)
\]

where \(\log p_\theta(I_s = 1 \mid y)\) is the AMOR term. This term explains the likelihood of being stationary given \(\theta\)

- The authors show that the AMOR term for the LAVA-Cat model could be written as:

\[
\log p_\theta(I_s = 1 \mid y) \propto -\lambda_{AMOR} \| \mu - \hat{\mu} \|^2 \\
= -\lambda_{AMOR} \| (I - \Phi)^{-1}c - \hat{\mu} \|^2 \\
= -\lambda_{AMOR} (I - \Phi)^{-2} \| c - (I - \Phi)\hat{\mu} \|^2
\]

The M-step of the BPEM algorithm is now modified with this AMOR term:

\[
\min_{c, \Phi} \sum_{i,t} -w^{(i)}_t \log p_\theta(x^{(i)}_t \mid x^{(i)}_{t-1}) + \lambda_{AMOR} \| \mu - \hat{\mu} \|^2
\]
2. Pseudo-Bayesian Update

- The idea of the *Pseudo-Bayesian update* method is to introduce two auxiliary variables: $d_t$ for $c$ and $\Psi_t$ for $\Phi$, respectively.

- At each $t$, $d_t$ and $\Psi_t$ are sequentially updated by solving the following equation:

\[
\min_{d_t, \Psi_t} \sum_i w_t^{(i)} \|x_t^{(i)} - \Psi_t x_{t-1}^{(i)} - d_t\|^2 \\
+ \lambda_{AMOR} (I - \Psi_{t-1})^{-2} \|d_t - (I - \Psi_t)\mu\|^2 \\
+ \lambda_{Bayes} \|d_t - d_{t-1}\|^2 + \lambda_{Bayes} \|\Psi_t - \Psi_{t-1}\|^2
\]

Let us define $B_t = (d_t \quad \Psi_t) \in \mathbb{R}^{(K-1) \times K}$:

\[
\min_{B_t} \left\| \begin{pmatrix}
    w_t X_t \\
    \mu^T \\
    \lambda_{Bayes} B_{t-1}^T
  \end{pmatrix}
- \begin{pmatrix}
    w_t^T \\
    \lambda_{AMOR}' \\
    \lambda_{AMOR}' \mu^T
  \end{pmatrix}
\right\|^2
\]

Thus, the solution of this least square problem is given as follows:

\[
B_t^* = (R^T R)^{-1} R^T S^T
\]

where $R \in \mathbb{R}^{(P+K+1) \times K}$ and $S \in \mathbb{R}^{(P+K+1) \times (K-1)}$. 
3. Low-resolution augmentation

**Core idea:** More observations usually help in reducing the variance of the estimated parameters. The AR structure of the data help us to obtain another set of observations at a low resolution:

\[ x_t = c_L + \Phi_L x_{t-2} + \epsilon_t \]

where \( c_L = c(I + \Phi) \) and \( \Phi_L = \phi^2 \)

The transformed parameters \( c_L \) and \( \Phi_L \) can be estimated by maximizing \( \log p_h(\theta)(g(y), I_s) \) using a low-resolution transformed time series. For instance, \( x_t \) for \( t = 2, 4, 6, \ldots, T \)

The authors point out that, if \( \Phi > 0 \), the estimated parameters for the low-resolution time series are:

\[
\Phi \approx (\sqrt{\hat{\Phi}_L + 2 \times \hat{\Phi}})/3 \\
c \approx (\hat{c}_L(I + \hat{\Phi})^{-1} + 2 \times \hat{c})/3
\]
Algorithm 3: LAMORE algorithm

Data: y, \( \theta_{\text{init}} \)
Result: \( \theta \)
\( \hat{\mu} = \text{Empirical-AM}(y); \)
while until converge do
    Initialize \( \{d_0, \zeta_0\} = \theta \) and \( \{e_0, \zeta_0\} = \theta \);
    for \( t \in 1:T \) do
        \( \{x_t^{(i)}, w_t^{(i)}\} = \text{1Step-BPF}(y_t, \theta, \{x_{t-1}^{(i)}, w_{t-1}^{(i)}\}); \)
        \( d_t, \psi_t = \text{BAM}(\{x_t^{(i)}, w_t^{(i)}\}, d_{t-1}, \psi_{t-1}); \)
        if \( t \mod 2 == 0 \) then
            \( e_t, \zeta_t = \text{BAM}(\{x_t^{(i)}, w_t^{(i)}\}, e_{t-2}, \zeta_{t-2}); \)
        end
    end
    Set \( c = (2d_T + e_T)/3 \) and \( \Phi = (2\psi_T + \zeta_{T/2})/3; \)
    Set \( \theta = \{c, \Phi\}; \)
end
Case Study: Medicare Part-D

- They want to measure the inertia of drug re-purchases in the Medicare part-D program for patients with at least one chronic condition (2007-2010, 5% US pop.)

- **Data:** For every beneficiary, the drug purchase sequence is $y_t \in \{\text{Novartis, Teva, others}\}$

  $\Phi$ is restricted to be a diagonal matrix:

  $\Phi = \begin{pmatrix} \phi_{\text{Novartis}} & 0 \\ 0 & \phi_{\text{Teva}} \end{pmatrix}$

- The model parameters are estimated for each individual, independently. So, each individual has a different $\Phi$ (re-purchase inertia)

- The model provides better predictive performance compared with *multinomial logistic regression*

\[
\log \frac{p(y_t = k)}{p(y_t = K)} = \beta_0^k + \beta_1^k y_{t-1} + \cdots + \beta_L^k y_{t-L}
\]

---

Figure 6: Purchase inertia parameters in the Medicare part-D program.

Figure 7: Medicare Drug purchase predictive performance of LAVA-Cat using LAMORE. Each cell shows Area under Receiver Operating Characteristic curves (AUROC) for one-versus-all settings. The blue dotted lines are the performance curves from the glmnet algorithm using lagged features.