A Weighted Ordered Probit Collaborative Kalman Filter for Hotel Rating Prediction

Myrsini Ntemi, Constantine Kotropoulos, and Emmanouil Gionanidis

Department of Informatics, Aristotle University of Thessaloniki, Greece

IEEE MLSP, Pittsburg, PA, October 14th, 2019

Introduction

Problem statement

- **Goal:** Anticipate hotel ratings that will be given by users with similar taste.

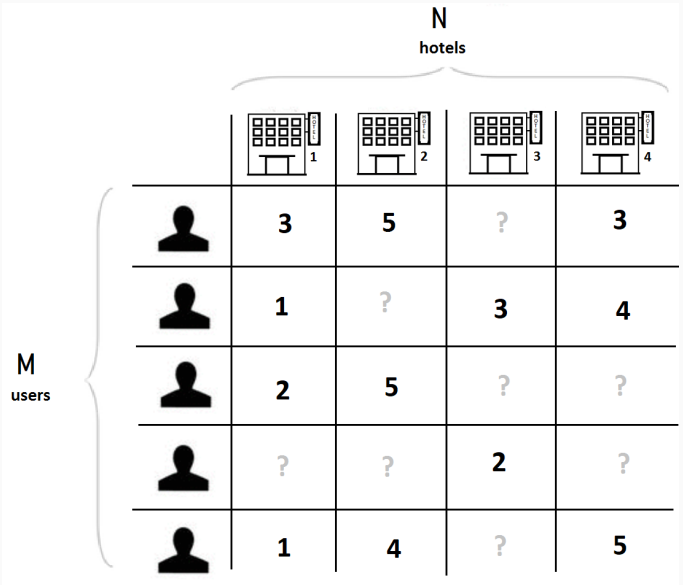

Figure 1: User and hotels organized into a matrix.

- Matrix factorization (MF) techniques model user - item (i.e., hotel accommodation) interactions as inner products in a joint latent space.

Collaborative Kalman filter (CKF) approach

CKF concept

- **Modeling:** Time evolving user - item feature vectors form a dyad.
- **Prediction:** Ordered probit regression.
- The real line is divided into as many regions as the number of rating stars.
- Calculation of the probability each dyad belongs to a specific region.

Figure 2: 5 star rating system.

Challenge

- A hotel popularity may volatile through time, since potential changes may occur in hotel services or accommodation conditions.
- MF techniques and CKF do not take into consideration the latent trend about each hotel popularity through time.

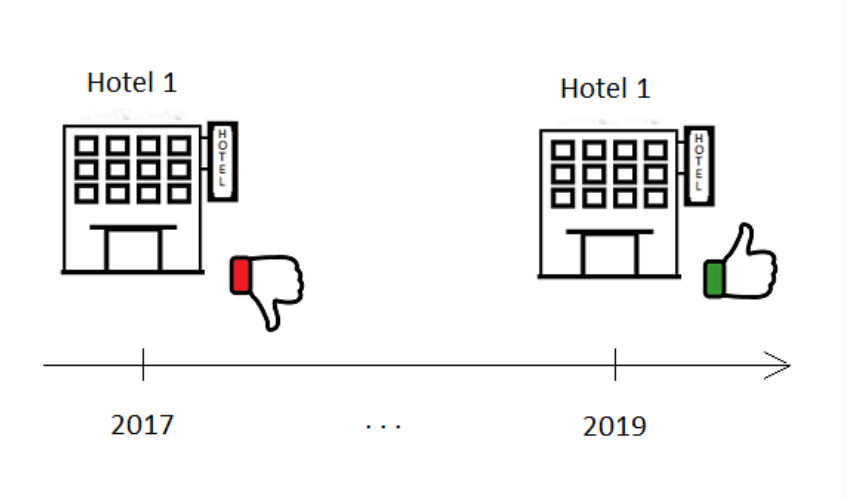

Figure 3: A hotel gaining popularity through time.

Proposed Solution

A weighted ordered probit collaborative Kalman filter

- The magnitude of the regions in the ordered probit model is assumed to evolve through time with respect to each hotel popularity evolution.
- A time window over the past observations is defined and the mean value of the ratings for a specific hotel is computed.
- The region of the ordered probit model, where the mean value falls in, increases.
- The magnitude assigned to each region corresponds to the standard deviation of a random variable following the truncated normal distribution, which is used to model the (approximate) posterior distribution of ratings.

The model at time t

- **Prior distributions:** Input dyad
  $\mathbf{u}_i[t] \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{u}_i}[t], \boldsymbol{\Sigma}_{\mathbf{u}_i}[t])$, $\mathbf{h}_j[t] \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{h}_j}[t], \boldsymbol{\Sigma}_{\mathbf{h}_j}[t])$.
- Prior parameters are the posterior ones of the previous time step
  - $\boldsymbol{\mu}[t] = \boldsymbol{\mu}'[t-1]$
  - $\boldsymbol{\Sigma}[t] = \boldsymbol{\Sigma}'[t-1] + \mathbf{I}\Delta^{[t]}$.
- After having measured the rating $z_{ij}[t]$ the posterior distributions
  - $\mathbf{u}_i'[t] \sim \mathcal{N}(\boldsymbol{\mu}'_{\mathbf{u}_i}[t], \boldsymbol{\Sigma}'_{\mathbf{u}_i}[t])$
  - $\mathbf{h}_j'[t] \sim \mathcal{N}(\boldsymbol{\mu}'_{\mathbf{h}_j}[t], \boldsymbol{\Sigma}'_{\mathbf{h}_j}[t])$.
- **Approximate posterior distributions** q**:** Minimize the KL divergence between q and true posterior distribution p

$$KL(q||p) = \mathbb{E}_q\left[\log \frac{q}{p}\right]. \tag{1}$$

Weighted Ordered Probit Assumption

Dynamically evolving region magnitudes

- kth region $\mathcal{I}_k = (l_k, r_k]$, $l_k < r_k$, $l_k = r_{k-1}$, $r_k = l_{k+1}$, $k = 1, 2, \ldots, 5$.
- Dynamically changing magnitude:
  - Sliding window $\mathcal{W}$ over the time series of star ratings with length $|\mathcal{W}|$.
  - Average rating assigned to hotel j: $\mu_{z_j}[t] = \frac{\sum_{t \in \mathcal{W}} z_j[t]}{|\mathcal{W}|}$
  - If $\mu_{z_j}[t] \in (l_k, r_k], k = 1, 2, \ldots, 5$, the magnitude becomes

$$\sigma^* = \begin{cases} \sigma\ c, & \text{if } \mu_{z_j}[t] \in \mathcal{I}_k \\ \sigma & \text{otherwise} \end{cases} \tag{2}$$

- $\mu_{z_j}[t] \mapsto \mathcal{I}_k$ encloses the evolution of each hotel popularity.

Figure 4: Dynamic magnitude at time t.

Output estimation (1)

- **Goal:** find the optimal approximate distributions $q(\cdot)$ of latent vectors and variables.
- Auxiliary variable $s_{ij}[t] \mid \mathbf{u}_i, \mathbf{h}_j \sim \mathcal{N}(\mathbf{u}_i[t]^T \mathbf{h}_j[t], \sigma^{*2})$.
- Ratings distribution $p(z_{ij}[t] \mid s_{ij}[t]) = \mathbb{I}(s_{ij}[t] \in \mathcal{I}_{z_{ij}[t]})$.
- If the rating $z_{ij}[t] \in \mathcal{I}_{z_{ij}[t]}$, the **approximate distribution** $q(s_{ij}[t])$:

$$q(s_{ij}[t]) = \mathcal{TN}_{\mathcal{I}_{z_{ij}[t]}}(s_{ij}[t] \mid \mathbb{E}_q[\mathbf{u}_i^T[t]]\mathbb{E}_q[\mathbf{h}_j[t]], \sigma^{*2}) \tag{3}$$

- **Prediction:** $\hat{z}_{ij}[t] = \mathbb{E}_q[\mathbf{u}_i^T[t]]\mathbb{E}_q[\mathbf{h}_j[t]]$. Let us assume that $\mathcal{I}_{z_{ij}[t]}$ is constructed by the left boundary $l_{z_{ij}[t]}$ and right one $r_{z_{ij}[t]}$. The following intervals are defined:

$$\zeta_{ij}[t] = \frac{l_{z_{ij}[t]} - \hat{\mu}_{ij}[t]}{\sigma^*} \quad \xi_{ij}[t] = \frac{r_{z_{ij}[t]} - \hat{\mu}_{ij}[t]}{\sigma^*} \tag{4}$$

where $\hat{\mu}_{ij}[t] = \mathbb{E}_q[\mathbf{u}_i'^T[t]]\mathbb{E}_q[\mathbf{h}_j'[t]]$.

Output estimation (2)

- **Expected value:** $\mathbb{E}_q[s_{ij}[t]]$

$$\mathbb{E}_q[s_{ij}[t]] = \hat{\mu}_{ij}[t] + \sigma^* \frac{\phi(\zeta_{ij}[t]) - \phi(\xi_{ij}[t])}{\Phi(\xi_{ij}[t]) - \Phi(\zeta_{ij}[t])} \tag{5}$$

- **Optimal posterior parameters** of $\mathbf{u}_i[t]$ (and similarly for $\mathbf{h}_j[t]$)):

$$\begin{aligned} \boldsymbol{\Sigma}'_{\mathbf{u}_i}[t] &= \left(\boldsymbol{\Sigma}^{-1}_{\mathbf{u}_i}[t] + \frac{\boldsymbol{\mu}'_{\mathbf{h}_j}[t]\, \boldsymbol{\mu}'^{T}_{\mathbf{h}_j}[t] + \boldsymbol{\Sigma}'_{\mathbf{h}_j}[t]}{\sigma^{*2}} \right)^{-1} \\ \boldsymbol{\mu}'_{\mathbf{u}_i}[t] &= \boldsymbol{\Sigma}'_{\mathbf{u}_i}[t] \left(\frac{\mathbb{E}_q[s_{ij}[t]] \boldsymbol{\mu}'_{\mathbf{h}_j}[t]}{\sigma^{*2}} + \boldsymbol{\Sigma}^{-1}_{\mathbf{u}_i}[t]\, \boldsymbol{\mu}_{\mathbf{u}_i}[t] \right) \end{aligned} \tag{6}$$

- The parameters of Eq. (6) are the prior parameters at $t+1$.

Experimental results

Data and parameter setting

- Hotels from 16 Greek destinations scrapped from Tripadvisor and ordered with respect to time.
- Hotel id, user id, rating stars (1-5), and date.
- Dimensionality of $\mathbf{u}_i$ and $\mathbf{h}_j$ was $K = 5$.
- $\sigma = 1.76$, $|\mathcal{W}| = 10$, $c = 10$ through leave-one destination-out validation, 15 experiments for Preveza destination with $c = 1, 2, \ldots, 14$ and $c = 20$.
- **WCKF is an online method:** at every time step t, predictions for the next time step $t + 1$.
- Figure of merit: Root mean square error (RMSE).

Prediction performance

Table 1: Hotel rating prediction performance.

Destination	RMSE WCKF	RMSE CKF	number of users	number of hotels	number of ratings
Anatoliki Thraki	**0.6479**	1.4122	5,593	110	6,762
Athens	**0.6712**	1.4691	5,1476	382	56,056
Argolida	**0.7529**	1.598	24,276	327	26,687
Crete	**0.7447**	1.5782	326,153	3,204	386,634
Ioannina	**0.855**	1.7677	15,819	286	18,541
Kefalonia	**0.7895**	1.7611	44,221	575	50,706
Magnisia	**0.791**	1.637	13,598	439	15,420
Messinia	**0.7341**	1.5638	5,694	59	6,218
Mykonos	**0.8051**	1.6956	80,216	513	89,668
Naxos	**0.8191**	1.7005	31,583	494	34,374
Paros	**0.8004**	1.7007	32,087	462	35,034
Rhodes	**0.6875**	1.5558	9,058	50	9,700
Santorini	**0.8204**	1.7134	170,187	1,097	196,536
Skiathos	**0.7518**	1.5747	6,177	137	6,696
Thessaloniki	**0.6714**	1.4595	26,119	254	30,517

Statistical significance via F-test (Reject H_0 if $\mathcal{F}_{1,2} < F_{1-\beta/2}$ or $\mathcal{F}_{1,2} > F_{\beta/2}$)

Table 2: F-test.

Destination	$F_{1-\beta/2}$	$F_{\beta/2}$	$\mathcal{F}_{1,2}$
Anatoliki Thraki	0.9608	1.0108	**1.0534**
Athens	0.9862	1.0140	**1.1122**
Argolida	0.9801	1.0203	**1.0534**
Crete	0.9947	1.0053	**1.0551**
Ioannina	0.9761	1.0245	**1.1683**
Kefalonia	0.9855	1.0147	**1.2176**
Magnisia	0.9739	1.0268	**1.1061**
Messinia	0.9591	1.0426	**1.2024**
Mykonos	0.9891	1.0110	**0.9410**
Naxos	0.9824	1.0179	**1.1313**
Paros	0.9826	1.0177	**1.0184**
Rhodes	0.9671	1.0340	**0.8596**
Santorini	0.9926	1.0074	**1.0083**
Skiathos	0.9606	1.0410	**0.8954**
Thessaloniki	0.9813	1.0190	**0.8692**

Absolute prediction error for Thessaloniki destination

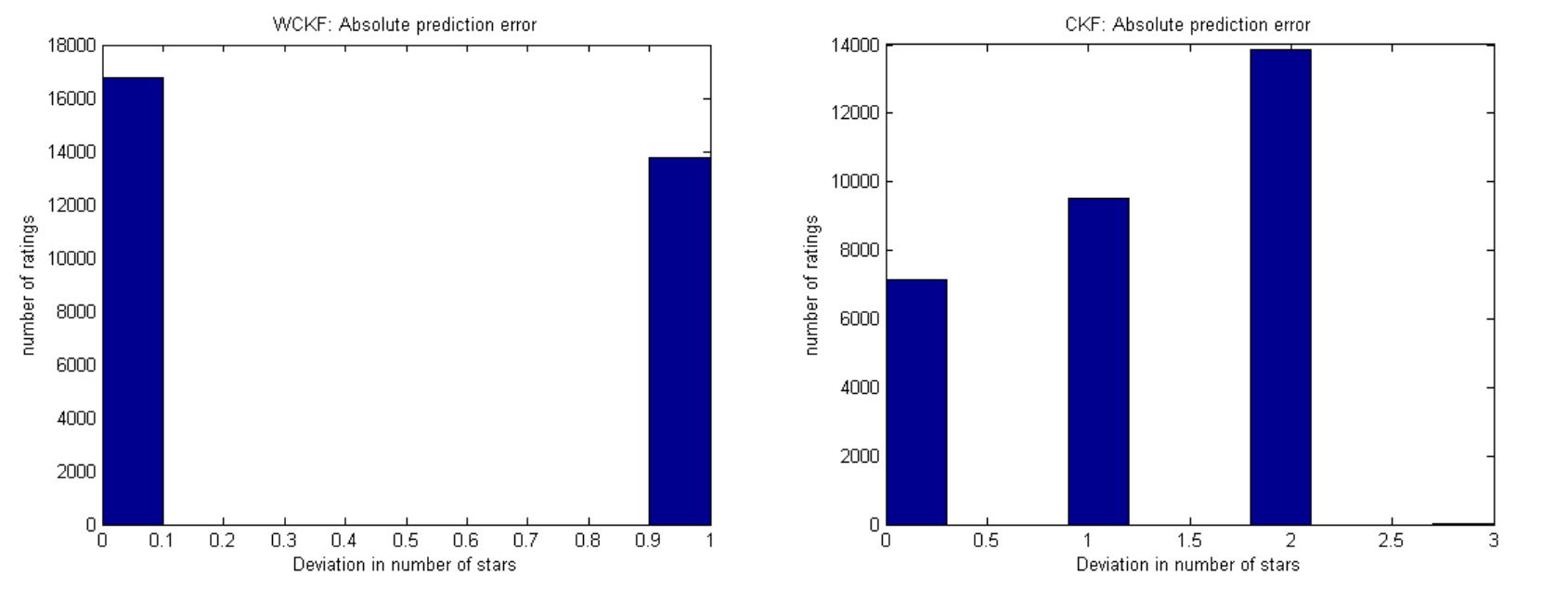

Figure 5: WCKF absolute prediction error for Thessaloniki.

Figure 6: CKF absolute prediction error for Thesaloniki.

Table 3: RMSE of WCKF for each destination and different values of c.

Destination	RMSE c=1	RMSE c=2	RMSE c=3	RMSE c=4	RMSE c=5	RMSE c=6	RMSE c=7	RMSE c=8	RMSE c=9	RMSE c=10	RMSE c=11	RMSE c=12	RMSE c=13	RMSE c=14	RMSE c=20
Anatoliki Thraki	1.341	0.7212	0.7039	0.6937	0.6881	0.6863	0.685	0.6722	0.6559	0.6479	0.6478	0.6459	0.649	0.679	0.7301
Athens	1.4376	0.9694	0.6884	0.6802	0.6741	0.666	0.6744	0.6731	0.6721	0.6712	0.6731	0.6771	0.7099	0.7281	0.8993
Argolida	1.5214	0.8714	0.8663	0.8421	0.7518	0.7509	0.7513	0.7519	0.7522	0.7529	0.7542	0.7718	0.792	0.8102	0.8419
Crete	1.4982	0.743	0.7439	0.7442	0.7446	0.7452	0.7458	0.7451	0.7449	0.7447	0.7459	0.7481	0.7509	0.7531	0.9165
Ioannina	1.5546	0.8571	0.8547	0.8564	0.8578	0.8631	0.8644	0.8601	0.8541	0.855	0.8583	0.8631	0.8685	0.8721	0.9001
Kefalonia	1.5222	0.7937	0.7912	0.7901	0.7851	0.7882	0.7889	0.7891	0.7893	0.7895	0.7937	0.7963	0.798	0.8012	0.8214
Magnisia	1.3421	0.7844	0.7867	0.7908	0.7915	0.7928	0.7931	0.7926	0.7919	0.791	0.7924	0.7931	0.7948	0.7961	0.9211
Messinia	1.428	0.7129	0.7071	0.7361	0.7368	0.7375	0.7381	0.7369	0.735	0.7341	0.7352	0.7361	0.7371	0.7499	0.9082
Mykonos	1.5562	0.8024	0.7991	0.7928	0.7958	0.7964	0.7982	0.7989	0.8032	0.8051	0.8109	0.8123	0.8179	0.8298	1.1099
Naxos	1.3392	0.7641	0.7839	0.7991	0.7999	0.8021	0.8099	0.8199	0.8195	0.8191	0.8201	0.8214	0.8219	0.8241	0.9993
Paros	1.4586	0.7947	0.8137	0.8168	0.8201	0.8214	0.8223	0.8109	0.8013	0.8004	0.8035	0.8145	0.8201	0.8239	1.0081
Rhodes	1.2987	0.8114	0.7032	0.7846	0.7881	0.7901	0.7914	0.7801	0.7611	0.6875	0.7012	0.7069	0.7123	0.7889	0.9044
Santorini	1.4018	0.8097	0.8161	0.8188	0.819	0.8195	0.8208	0.8213	0.8209	0.8204	0.8221	0.8229	0.8235	0.8247	0.9002
Skiathos	1.1793	0.6814	0.6919	0.7416	0.7442	0.7449	0.7492	0.7521	0.7522	0.7518	0.752	0.7529	0.7541	0.7566	0.8002
Thessaloniki	1.2241	0.6246	0.6232	0.6503	0.6541	0.6599	0.6623	0.6682	0.6721	0.6714	0.6708	0.6892	0.6901	0.6945	0.8149

Conclusion

Experimental findings

- A weighted ordered probit collaborative Kalman filter for tripadvisor hotel rating prediction.
- The dynamically changing magnitude of the regions of the weighted ordered probit model with respect to each hotel trend ensured the efficacy of rating predictions.
- WCKF takes into consideration possible changes in hotel services or accommodation conditions and readjusts the rating predictions.

Acknowledgments

This research has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE (project code: T1EDK-02474).

Closing

Thank you for your attention!
Questions?