Challenges of multidimensional transactional data

Miklós Koren
#istandwithceu

EEA Research Committee Session

Representing transactional data

What is transactional data?

- Many observational datasets are transactional:
  - administrative: customs declarations, VAT/sales tax declarations, wage data
  - private sector: sales, customer service events, website logs

Star schema

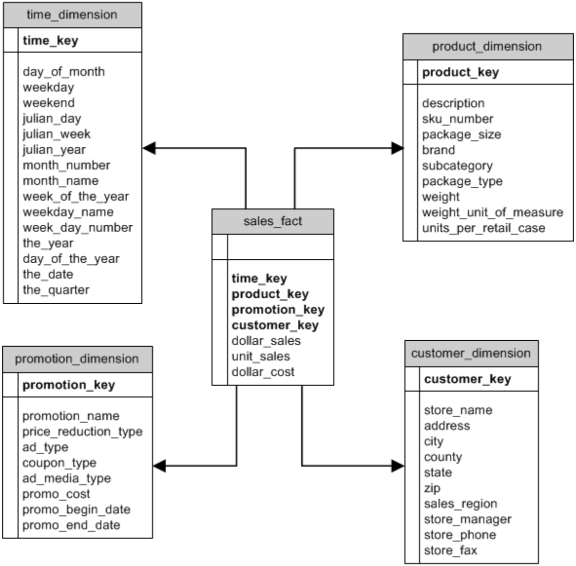
Dimension

- An attribute *identifying* the transaction.
- Typically categorical: salesperson, client, region, product.
- But: time, space.

Fact

- An attribute *characterizing* the transaction.
- Typically numerical: quantity, price, freight charge.

Star schema in a relational database



An econometrician's view

$$X_{ijklmnop}$$

- dimensions: i, j, k, l, m, n, o, p
- fact: X

Real-world examples

Real-world examples

- Product-level export (U.S.): Armenter and Koren (2013)
- VAT (Belgium): Dhyne, Magerman and Rubinova (2015)
- Procurement (Hungary): Koren, Szeidl, Szucs and Vedres (2017)

Product-level export (U.S.)

- Transaction: product line on a customs declaration
- Observations: 22 million/year
- Dimensions:
  - Products: 9,000 Schedule-B codes
  - Exporting firms: 160,000
  - Dates: 365 days
  - Destination countries: 200
- Combinations of dimensions: 100 trillion
- Fraction of zeros: 99.999978%

VAT (Belgium)

- Transaction: B2B sales (partner-specific VAT declaration)
- Observations: 15 million/year
- Dimensions:
  - Buying firms: 2.7 million
  - Selling firms: 2.7 million
- Combinations of dimensions: 7.3 trillion
- Fraction of zeros: 99.999795%

Procurement (Hungary)

- Transaction: Public procurement tender
- Observations: 20,000/year
- Dimensions:
  - Products: 5,900 9-digit CPV codes
  - Buying firms: 7,700
  - Selling firms: 24,000
  - Dates: 365 days
- Combinations of dimensions: 400 trillion
- Fraction of zeros: 99.99999999%

Modeling transactional data

Two approaches to statistical modeling

Dimensions first

$$X_{ijklmnop} \sim F()$$

independently across dimensions

Two approaches to statistical modeling

Dimensions first

$$X_{ijklmnop} \sim F()$$

independently across dimensions

Transactions first

$$\{X, i, j, k, l, m, n, o, p\} \sim F()$$

independently across transactions

Challenges for estimation, inference and prediction

1. Too many dimensions
2. Too many observations
3. Too many zeros
4. Too many fixed effects

Challenges for estimation, inference and prediction

1. Too many dimensions
2. Too many observations
3. Too many zeros
4. Too many fixed effects
5. Continuous dimensions

Too many dimensions

- Challenging to estimate fixed effects.
- (Within transformation can be applied if balanced.)

Too many observations

- Computational constraints: memory, time.
- Common approach: arbitrary sample (e.g., zoom in on positive flows)
  - Unknown statistical properties.

Too many zeros

- In typical transactional data, more than 99.999% of potential categories have $n = 0$.
- Multi-level modeling of zero and non-zero facts.
  - Particularly challenging with fixed effects.
- Endangers numerical accuracy.
- Prediction is hard.

Too many fixed effects

- It is common to include fixed effects for each dimension.
- This becomes prohibitive with 4-5 dimensions and trillions of fixed effects to estimate.
- Particularly with nonlinear estimators.

Continuous dimensions

- Some dimensions are continuous: time, space.
- Common approach: discretize (year, month, city, ZIP-code).
  - Arbitrary interval definitions (see: Modifiable Area Unit Problem).
  - Independence assumption may not be valid.
  - Unnecessary duplication of data (memory, time).

What can we do?

What can we do?

Estimation

Use multinomial or other discrete choice model for transactional data. Nonlinear, but much fewer observations.

What can we do?

Estimation

Use multinomial or other discrete choice model for transactional data. Nonlinear, but much fewer observations.

Inference

For simple null models (e.g., independent dimensions), simulating an empirical joint distribution F is easy. (Armenter and Koren, 2013)

What can we do?

Estimation

Use multinomial or other discrete choice model for transactional data. Nonlinear, but much fewer observations.

Inference

For simple null models (e.g., independent dimensions), simulating an empirical joint distribution F is easy. (Armenter and Koren, 2013)

Prediction

Empirical Bayes may handle large number of zeros well ("missing butterfly problem").

Conclusion

- Transactional data is everywhere and is very useful.
- But also very sparse: with categories far exceeding observations.
- Model transactions rather than dimensions.