Université du Québec en Outaouais

# A Probabilistic Model for Data Cube Compression and Query Approximation

R. Missaoui, C. Goutte, A.K. Choupo & A. Boujenoui

DOLAP'07 – November 9, 2007

National Research Council Canada

Conseil national de recherches Canada

# Outline

- Introduction and motivation
- Probabilistic Data Modeling
  - Non-negative multi-way array factorization
  - Log-linear modeling
  - Rates of compression and approximation
- Experimental results
  - Data sets
  - Compression and approximation
  - Approximate query answering
- Discussion and conclusion

# Introduction

- Research on data approximation and mining in data cubes
- Some facts
  - Very large data cubes to store and process
  - Data cubes are multi-way tables
  - High dimensional cubes with possibly useless dimensions or associations among dimensions
  - Patterns (e.g., clusters, outliers, correlations) are hidden in large, heterogeneous and sparse data sets
  - Users prefer approximate answers with quick response time rather than exact answers with slow execution time

# Introduction

- Contribution
  - Probabilistic modeling for data approximation, compression and mining in data cubes
  - Focus on non-negative multi-way array factorization (NMF)
  - Potential for approximate query answering
  - Comparison with log-linear modeling (LLM)

# Introduction

- Related work
  - Cube approximation and compression
    - Barbara & Wu, Sarawagi *et al.,* Vitter *et al.*
  - Outlier detection
    - Sarawagi *et al.,* Palpanas *et al*.,
  - Approximate query answering
    - Sampling (Ganti *et al*.), clustering (Yu and Shan), wavelets (Chakrabarti *et al.)*
  - Approximating original multidimensional data from aggregates
    - Iterative proportial fitting (IPF): Palpanas *et al.*

# Probabilistic datacube modeling

- Assume counts in cube $X=[x_{ijk}]$ arise from a probabilistic model $P(i,j,k)$.

$\Rightarrow$ $X$ is a sample from multinomial distribution $P(i,j,k)$.

- Quality of Model $\theta$ is measured by the (log-)likelihood:

$$L(\theta) = \ln P(X \mid \theta) = \sum_{ijk} \ln P(i,j,k)$$

- All models implement a trade-off between fit (high $L(\theta)$) and compression (number of parameters).
- We introduce one such model, NMF, and compare it to the well-known log-linear modeling (LLM).

# Non-negative multi-way array factorization

- Additive sum of **M** non-negative components:

$$P(i,j,k) = \sum_{m=1}^{M} P(m)P(i \mid m)P(j \mid m)P(k \mid m)$$

- Each component is a product of conditionally independent multinomial distributions.

  $\Rightarrow$ Observations behave "the same" in each component

- Equivalent to decomposition of multi-way array $\mathbf{X}$:

$$\frac{1}{N}\mathbf{X} \approx P(i,j,k) = \sum_{m=1}^{M} \mathbf{W}^m \otimes \mathbf{H}^m \otimes \mathbf{A}^m$$

- ...into non-negative factors (probabilities

  $\mathbf{W}=[P(i,m)]$, $\mathbf{H}=[P(j|m)]$, $\mathbf{A}=[P(k|m)]$)

# NMF (cont'd)

- Estimation by maximizing the log-likelihood, or equivalently the deviance: $G^2 = 2\sum_{ijk} x_{ijk} \ln \frac{\hat{x}_{ijk}}{x_{ijk}}$
- Expectation-Maximization(EM) algorithm

  $\Rightarrow$ Iterative algorithm with multiplicative update rules
- More components $\Rightarrow$ better fit, less compression
- Model selection: finding best trade-off
- Use Information Criteria such as AIC or BIC

$$\text{AIC} = \hat{G}^2 - 2df \quad \text{and} \quad \text{BIC} = \hat{G}^2 - df \times \ln N$$

Maximum deviance

Degrees of freedom

# Log-linear modeling

- Decompose the log-probability as an additive sum

$$\ln P(i,j,k) = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}$$

1st order (no interaction) | Interaction between 2 dimensions | Interaction between all dimensions

- Maximum likelihood estimation using Iterative Proportional Fitting.
- Parsimonious model: model that bests fit data
- Backward elimination: start with a large model and use $\chi^2$ to test that removal of interaction yields no significant loss in fit.
- Other variants: forward selection, …

# Rates of compression and approximation

- Approximation: measured by deviance $G^2$:
  - $G^2$=0 means perfect approximation (saturated model)
  - Higher $G^2 \Rightarrow$ worse approximation
- Compression: How much smaller is the model?
  - Compression rate: ratio of parameters over cells:

$$R_c = 1 - \frac{f}{N_c} = \frac{df}{N_c}$$

degrees of freedom
number of cells

  - For NMF:

$$R_c = 1 - \frac{M(I + J + K - 2)}{IJK}$$

number of components

# Experiments: 3 datasets

| | Governance | Customer | Sales |
|---|---|---|---|
| Dimensions | 3 x 4 x 2 x 2 | 2 x 8 x 6 x 5 x 5 | 44 x 4 x 3 |
| Nb. cells | 48 | 2400 | 528 |
| Nb. facts | 214 | 10281 | 5191 |
| Density | 63% | 37% | 50% |

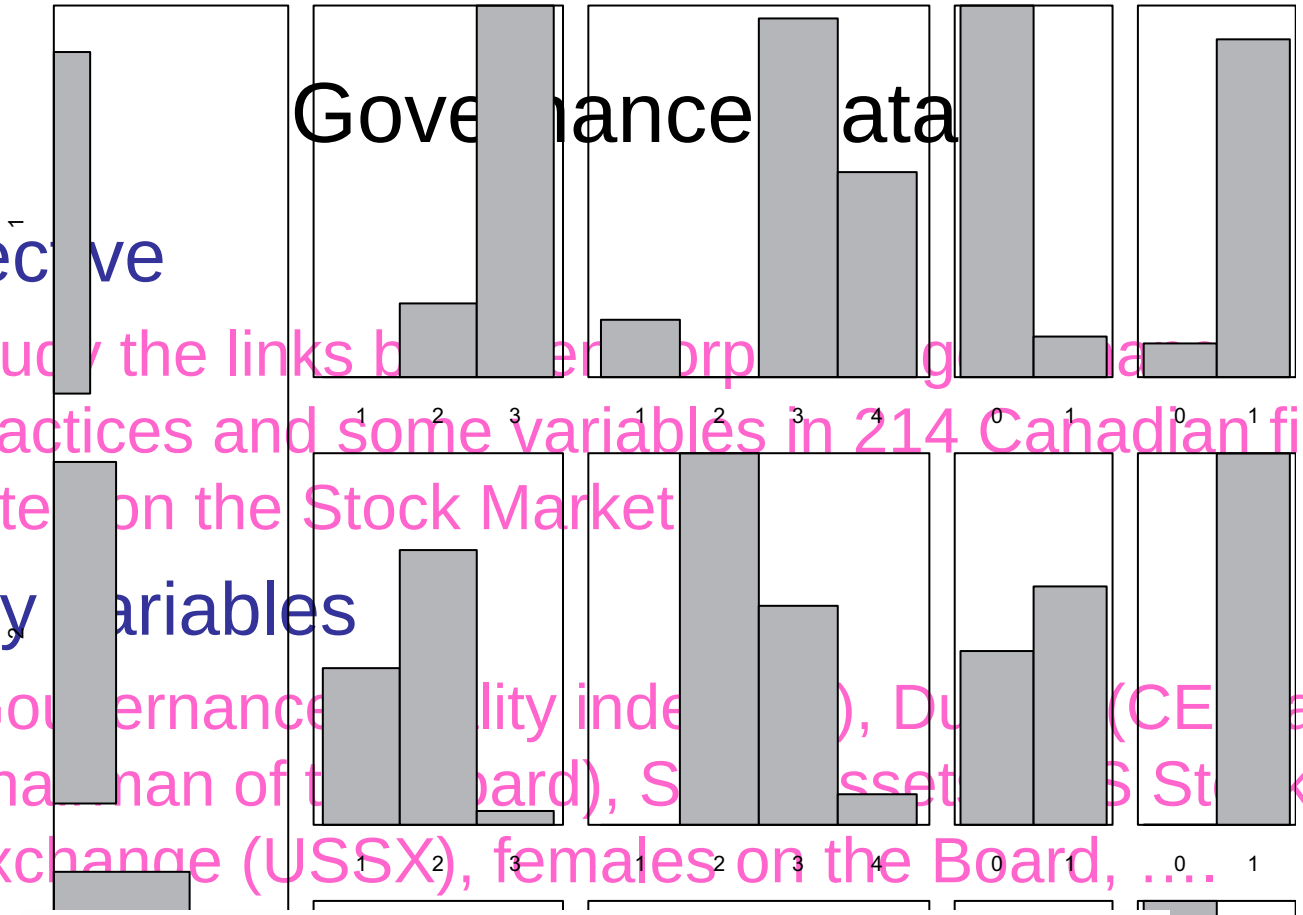
Governance: “Toy” example but real data.

Customer: from FoodMart data in SQL Server analysis Services. Large, high-dimensional table.

Sales: also from FoodMart. One dimension with many modalities (44 product categories)

# Governance data

- Objective
  - Study the links between corporate governance practices and some variables in 214 Canadian firms listed on the Stock Market
- Many variables
  - Governance quality index (GQI), Duality (CEO and Chairman of the board), Size, Assets, US Stock Exchange (USSX), females on the Board, ….




| USSX | Size | Duality: No QI:Low | Med | High | Yes Low | Med | High |
|---|---|---|---|---|---|---|---|
| No | 1 | 0 | 7 | 0 | 4 | 3 | 0 |
| | 2 | 7 | 21 | 12 | 6 | 12 | 4 |
| | 3 | 11 | 13 | 11 | 4 | 4 | 2 |
| | 4 | 0 | 3 | 1 | 0 | 2 | 0 |
| Yes | 1 | 0 | 1 | 2 | 0 | 0 | 0 |
| | 2 | 4 | 12 | 0 | 7 | 10 | 1 |
| | 3 | 4 | 4 | 14 | 5 | 8 | 2 |
| | 4 | 0 | 3 | 7 | 0 | 2 | 1 |

# NMF and LLM in action

- Governance cube
  - 48 cells, four dimensions: QI, Duality, USSX and Size
  - Parsimonious LLM model: {QI*Size*USSX,QI*Duality}

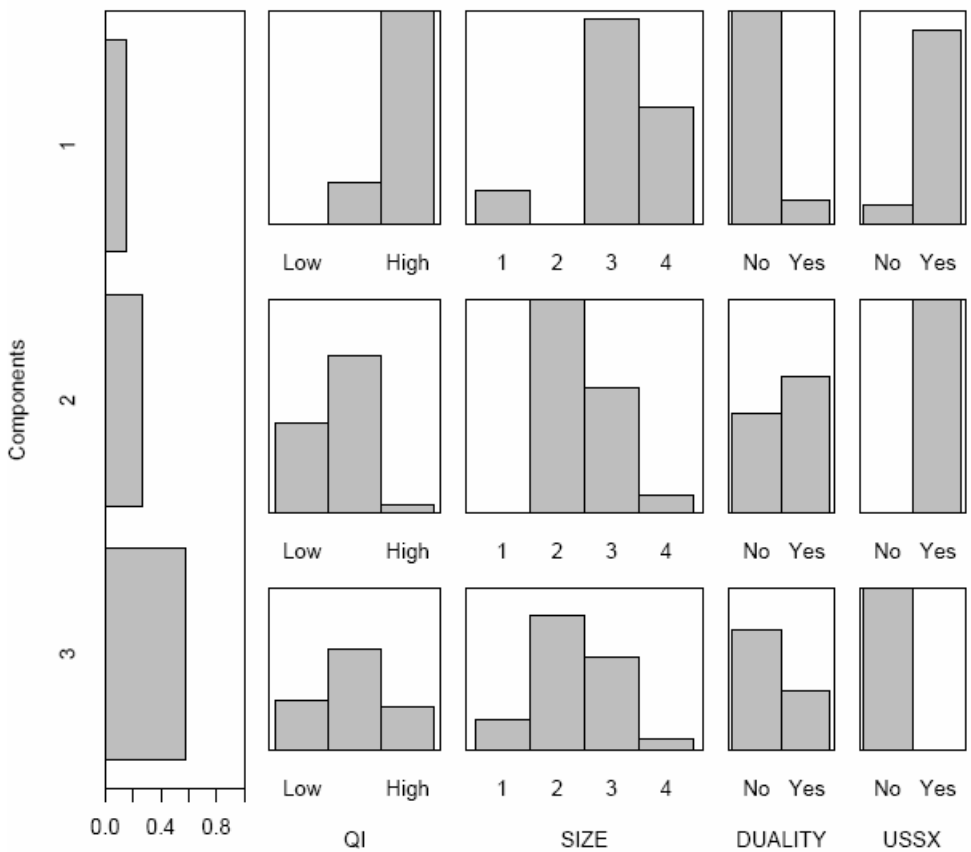
| | | QI | | |
|---|---|---|---|---|
| | SIZE | Lo | Me | Hi |
| USSX =No | 1 | 4 | 10 | 0 |
| | 2 | 13 | 33 | 16 |
| | 3 | 15 | 17 | 13 |
| | 4 | 0 | 5 | 1 |
| =Yes | 1 | 0 | 1 | 2 |
| | 2 | 11 | 22 | 1 |
| | 3 | 9 | 12 | 16 |
| | 4 | 0 | 5 | 8 |

| | DUALITY | |
|---|---|---|
| QI: | No | Yes |
| Low | 26 | 26 |
| Med | 64 | 41 |
| High | 47 | 10 |

Table 6: The two sub-cubes identified by LLM as substitutes for the original data cube: QI*SIZE*USSX, left, and QI*DUALITY, right.

# NMF and LLM in action

- Governance cube
  - Parsimonious NMF model (3 components)



# NMF and LLM in action

- Governance cube
  - Parsimonious NMF model (3 components)

| Comp1 | Duality:No | | | Yes | |
|---|---|---|---|---|---|
| | Size | QI:Md | Hi | Md | Hi |
| USSX =No | 1 | 0 | 0 | 0 | 0 |
| | 3 | 0 | 1 | 0 | 0 |
| | 4 | 0 | 1 | 0 | 0 |
| =Yes | 1 | **0** | **2** | **0** | **0** |
| | 3 | 3 | **13** | 0 | **1** |
| | 4 | **1** | **7** | 0 | **1** |

| Comp2 | Duality=No | | | | =Yes | | |
|---|---|---|---|---|---|---|---|
| | Size | QI:Lo | Md | Hi | Lo | Md | Hi |
| USSX =Yes | 2 | **5** | **9** | **0** | **7** | **12** | **1** |
| | 3 | **3** | **5** | 0 | **4** | **7** | 0 |
| | 4 | **0** | 1 | 0 | **1** | **1** | 0 |

| Comp3 | Duality=No | | | | =Yes | | |
|---|---|---|---|---|---|---|---|
| | Size | QI:Lo | Md | Hi | Lo | Md | Hi |
| USSX = No | 1 | **2** | **5** | **2** | **1** | **2** | **1** |
| | 2 | **11** | **22** | **9** | **5** | **11** | **5** |
| | 3 | **7** | **15** | **6** | **4** | **7** | **3** |
| | 4 | **1** | **2** | **1** | **0** | **1** | **0** |

Table 2: Components of the 3-component NMF model. Cells where each component dominates are in bold. Non represented rows and columns are uniformly equal to 0.

# Compression vs. approximation

| *GOVERNANCE* | Sub-cubes | Param | $R_c$(%) | $G^2$ |
|---|---|---|---|---|
| NMF (best BIC) | 2 | 16 | 66.7 | 56 |
| NMF (best AIC) | 3 | 24 | 50.0 | 35 |
| LLM | 2 | 26 | 45.8 | 23 |

*CUSTOMER* $N_c$=2x8x6x5x5, N=10281

| | | | | |
|---|---|---|---|---|
| NMF (best BIC) | 5 | 110 | 95.4 | 1020 |
| NMF (best AIC) | 6 | 132 | 94.5 | 917 |
| LLM | 4 | 567 | 76.4 | 595 |

*SALES* $N_c$=44x4x3, N=5191

| | | | | |
|---|---|---|---|---|
| NMF (best BIC) | 8 | 392 | 25.8 | 715 |
| NMF (best AIC) | - | 528 | 0 | 0 |
| LLM | - | 528 | 0 | 0 |

- Good compression on GOVERNANCE and CUSTOMER cubes
- BIC: more parsimonious NMF than AIC (or LLM)
- LLM *approximates* better
- NMF *compresses* better
- Eg: NMF models 2400 cells in *CUSTOMER* with 110 parameters only!

# Approximate query answering

- Query reformulation on NMF components
- Select a portion of the cube (*Slice* and *Dice* differ on the extent of the selection)
- Probabilistic model cuts the processing time as:
  - Only necessary cells need to be calculated (no need to compute entire cube).
  - Irrelevant (i.e., outside of the query scope) components may be ignored.
- Saving is important if query selects a small part of the cube and components are well distributed.

# *Slice* and *Dice* (cont'd)

**CUSTOMER**

| Dimensions | Modalities | | | | | |
|---|---|---|---|---|---|---|
| | Data | C1 | C2 | C3 | C4 | C5 |
| Status | 1,2 | 1,2 | 1,2 | 1,2 | 1,2 | 1,2 |
| Income | 1-8 | 4-8 | 1-3 | 1-3 | 2,3 | 1-4,6,8 |
| Children | 0-5 | 0-5 | 0-5 | 0-5 | 0-5 | 0-5 |
| Occupation | 1-5 | 4,5 | 1-5 | 1,2 | 1,2 | 4,5 |
| Education | 1-5 | 1-5 | 3 | 1,2 | 1-3 | 4,5 |

- *Slice*: (Status,Income,Children,Occupation) for customers with Education=4
  - "Slice" C1 and C5 only; add them to get the answer.
- *Dice*: (Status,Income,Occupation) for customers with Education=4 or 5, and Children>2
  - "Dice" C1 and C5 only, add them to get the answer.

# Approximate query answering: *Roll-up*

- Aggregate values over all (or subset of) modalities of one or several dimensions
- Easily implemented by summing over probabilistic profiles in the model
- For example, roll-up over dimension k:

$$\sum_{k=1}^{K} \underbrace{P(i,j,k)}_{\approx X_{ijk}/N} = \sum_{m=1}^{M} P(m)P(i\mid m)P(j\mid m) \underbrace{\sum_{k=1}^{K} P(k\mid m)}_{=1} = \sum_{m=1}^{M} P(m)P(i\mid m)P(j\mid m)$$

- Get rolled-up model “for free” from original model
- Roll-up on model much faster than on data

# *Roll-up* (cont'd)

| Dimensions | Modalities | | | | | |
|---|---|---|---|---|---|---|
| | Data | C1 | C2 | C3 | C4 | C5 |
| Status | 1,2 | 1,2 | 1,2 | 1,2 | 1,2 | 1,2 |
| Income | 1-8 | 4-8 | 1-3 | 1-3 | 2,3 | 1-4,6,8 |
| Children | 0-5 | 0-5 | 0-5 | 0-5 | 0-5 | 0-5 |
| Occupation | 1-5 | 4,5 | 1-5 | 1,2 | 1,2 | 4,5 |
| Education | 1-5 | 1-5 | 3 | 1,2 | 1-3 | 4,5 |

- *Roll-up1*: Income,Occupation,and Education only
  - Combine 3 probabilistic profiles (instead of 5)
- *Roll-up2*: Climb up the Income hierarchy [1,3],[4,5],[7,8]
  - Component C1 is irrelevant for interval [1,3]
  - Components C2 and C3 are irrelevant for [4,5] and [7,8]

# Conclusion – NMF vs LLM

- Differences
  - Better compression (but less precision) with NMF
  - NMF finds homogeneous dense regions (components) in cubes and relevant members of all dimensions in components
  - LLM identifies important associations between dimensions for all members of selected dimensions
  - LLM imposes more constraints (density and data size)
  - NMF is more precise for selection queries while LLM seems more appropriate for aggregation queries (due to IPF)

# Conclusion – NMF vs LLM

- Similarity
  - Probabilistic modeling
  - Approximation/compression and outlier detection (by comparing estimated values with actual data)
- Complementarity
  - NMF and LLM are therefore complementary techniques

# Conclusion

- Future work
  - Incremental update of a precomputed model when new dimensions or dimension members are added
  - Use NMF to identify dense components that are further modeled with LLM
  - Efficient implementation of model selection procedures for NMF and LLM
  - Experimentation on very large data cubes (e.g., DBLP data)

# References


- Daniel Barbara and Xintao Wu. Using loglinear models to compress datacube. In Proceedings of the First International Conference on Web-Age Information Management, p. 311–322, London, UK, 2000. Springer-Verlag.
- Cyril Goutte, Rokia Missaoui & Ameur Boujenoui. Data Cube Approximation and Mining using Probabilistic Modelling, *Research Report* # 49284, ITI, CNRC, 20 pages, March 2007.

  http://iit-iti.nrc-cnrc.gc.ca/iit-publications-iti/docs/NRC-49284.pdf
- Themis Palpanas, Nick Koudas, and Alberto Mendelzon. Using datacube aggregates for approximate querying and deviation detection. IEEE TKDD,17(11):1465–1477, 2005.
- Sunita Sarawagi, Rakesh Agrawal, and Nimrod Megiddo. Discovery-driven exploration of olap data cubes. In EDBT ’98: Proceedings of the 6th ICDT, p. 168–182, London, UK, 1998. Springer-Verlag.
- J.S.Vitterand and M.Wang. Approximate computation of multidimensional aggregates of sparse data using wavelets. In Proceeding of the SIGMOD’99 Conference, pages193–204,1999.