Maximum likelihood estimation in undirected graphical models

Kaie Kubjas, 2.12.2020

Agenda

- Maximum likelihood estimation for undirected graphical models
  - Gaussian setting
  - Discrete setting
- Bachelor's and Master's thesis topics
- Today's lecture based on lecture notes by Caroline Uhler from the MIT course "Algebraic techniques and semidefinite optimization" (lecture 17)

Graphical models

In the graphical model associated to a graph G:

- an edge (u, v) of the graph G expresses some sort of dependence between the vertices u and v;

- a non-edge (u, v) of the graph G expresses some sort of conditional independence between the vertices u and v.

Examples

- Gene association network
- Stock exchange
- Markov chains
- Hidden Markov models: DNA sequence alignment
- Ising model

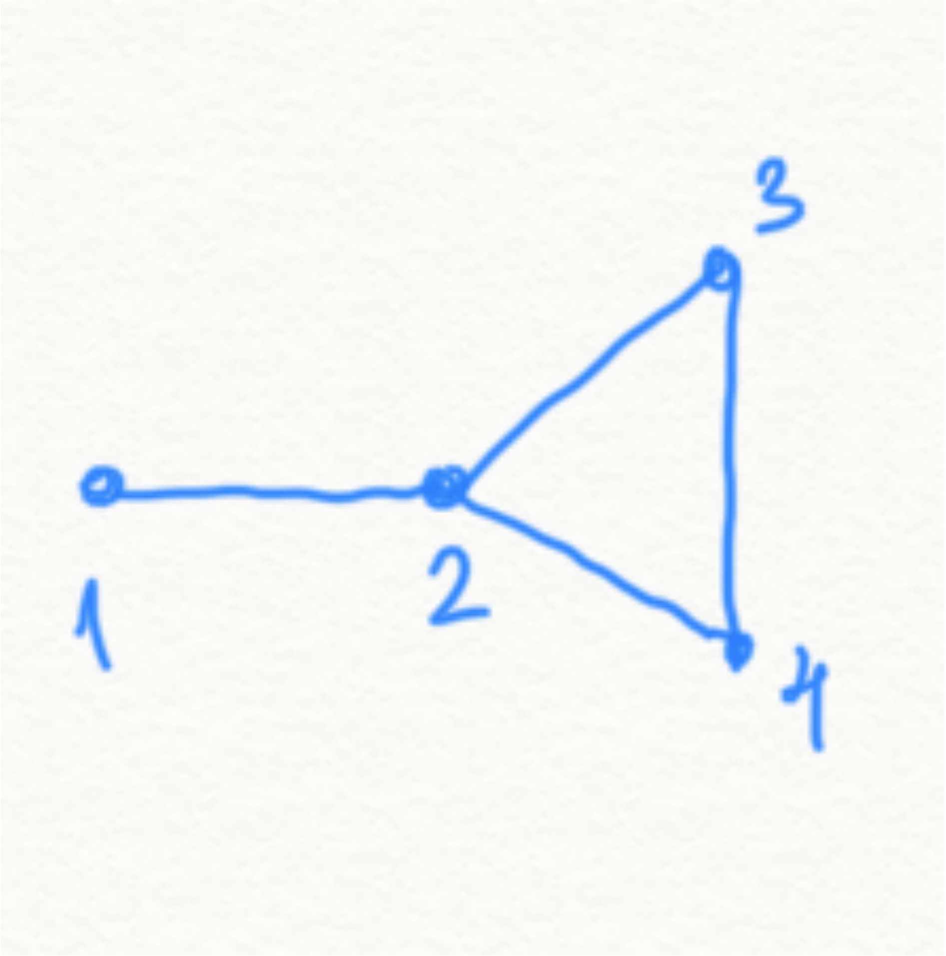
Markov properties

Let $G = (V, E)$ be an undirected graph.

Def: The pairwise Markov property associated to G consists of all conditional independence statements $X_u \perp\!\!\!\perp X_v \mid X_{V \setminus \{u,v\}}$, where (u, v) is not an edge of G.

Def: The global Markov property associated to G consists of all conditional independence statements $X_A \perp\!\!\!\perp X_B \mid X_C$ for all disjoint sets A, B, and C such that C separates A and B in G.

Markov properties



- $\mathscr{C}_{\text{pairwise}} = \{1 \perp\!\!\!\perp 3 \,|\, (2,4), 1 \perp\!\!\!\perp 4 \,|\, (2,3)\}$
- $\mathscr{C}_{\text{global}} = \mathscr{C}_{\text{pairwise}} \cup \{1 \perp\!\!\!\perp (3,4) \,|\, 2\}$

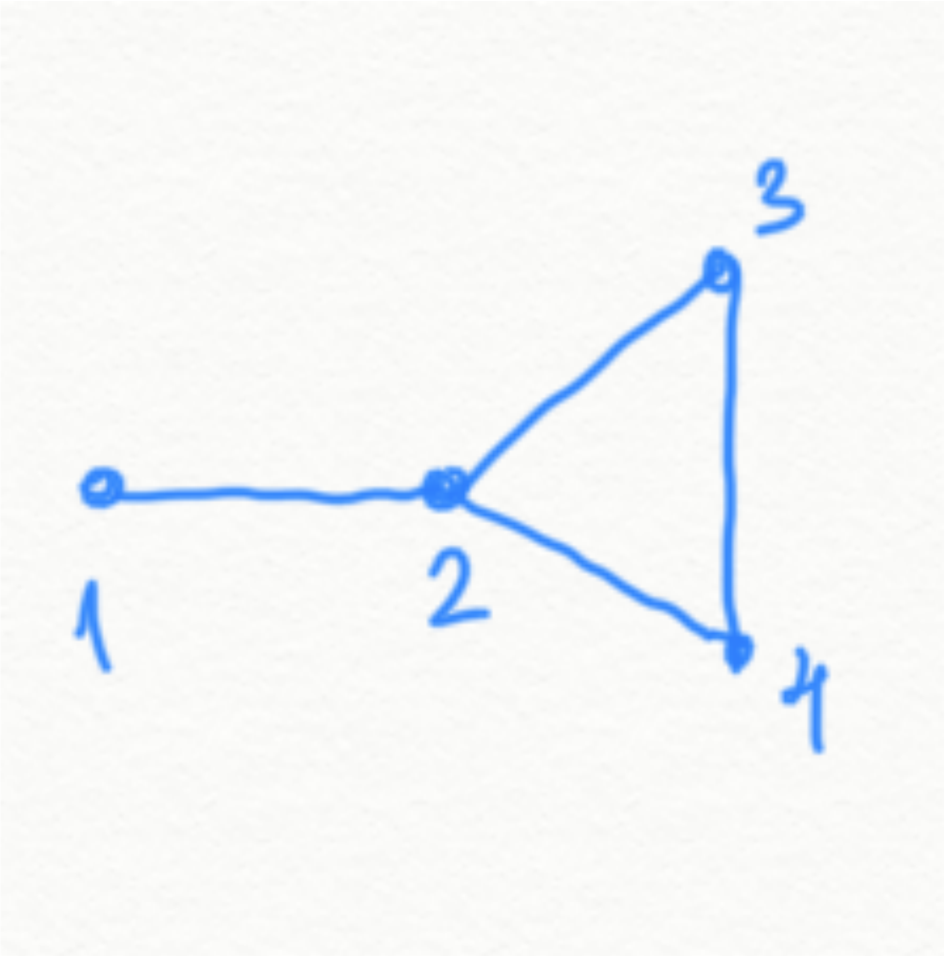
Factorization property

Def: The distribution of X factorizes according to the graph G if its probability density function $f(x)$ can be written as

$$f(x) = \frac{1}{Z} \prod_{C \in \mathscr{C}(G)} \phi_C(x_C),$$

where ϕ_C are some potential functions and $Z < \infty$ is the normalizing constant.

Factorization property



- Factorization property: $p_{ijkl} = \dfrac{1}{Z}\theta^{(12)}_{ij}\theta^{(234)}_{jkl}$ for $(i, j, k, l) \in [0,1]^4$

Comparison of ideals

In this example:

- $I_{\text{pairwise}(G)} \subsetneq I_{\text{global}(G)}$
- $I_G := \langle p_{ijkl} - \theta_{ij}^{(12)} \theta_{jkl}^{(234)} : (i, j, k, l) \in \{0,1\}^4 \rangle \cap \mathbb{R}[p] = I_{\text{global}(G)}$
- The last equality holds since G is a chordal graph

Gaussian setting

- The pairwise Markov property holds for a Gaussian distribution if and only if $K_{u,v} = 0$ for all $(u, v) \notin E$. [Poll]
- Since a Gaussian distribution is positive, it satisfies the pairwise Markov property for a graph G if and only if it factorizes according to graph G by the Hammersley-Clifford theorem.
- Since a Gaussian distribution also satisfies the intersection axiom, it satisfies the pairwise Markov property for a graph G if and only if it satisfies the global Markov property for a graph G.
- NB! This does not mean that the three ideals are equal. In Homework 5, compute the vanishing ideal of I_G.

Maximum likelihood estimation in Gaussian graphical models

MLE in Gaussian graphical models

- $G = (V, E)$ undirected graph
- D data, $\bar{X}$ sample mean, S sample covariance matrix
- The log-likelihood function is

$$log(\mu, \Sigma \,|\, D) = -\frac{1}{2} \sum_{i=1}^{n} \left(m \log(2\pi) + \log \det(\Sigma) + \left(X^{(i)} - \mu\right)^T \Sigma^{-1} \left(X^{(i)} - \mu\right) \right)$$

- Using the trace trick gives

$$log(\mu, \Sigma \,|\, D) = -\frac{1}{2} \left(nm \log(2\pi) + n \log \det(\Sigma) + \mathrm{tr} \left(\sum_{i=1}^{n} \left(\left(X^{(i)} - \mu\right) \left(X^{(i)} - \mu\right)^T \right) \Sigma^{-1} \right) \right)$$

MLE in Gaussian graphical models

- MLE in a Gaussian graphical model gives: $\hat{\mu} = \bar{X}$

- The log-likelihood function is

$$log(\mu, \Sigma \mid D) = -\frac{1}{2}\left(nm\log(2\pi) + n\log|\Sigma| + \mathrm{tr}\left(\sum_{i=1}^{n}\left(\left(X^{(i)} - \mu\right)\left(X^{(i)} - \mu\right)^{T}\right)\Sigma^{-1}\right)\right)$$

- After some more simplifications, the maximum likelihood estimation problem becomes:

$$\max_{\Sigma \succcurlyeq 0} \quad \log\det(\Sigma^{-1}) - \mathrm{trace}(\Sigma^{-1}S)$$

$$\text{subject to} \quad \Sigma \in V(I_{\mathrm{pairwise}(G)})$$

MLE in Gaussian graphical models

- This optimization problem becomes convex, if we write it using K instead of Σ:

$$\max_{K \succcurlyeq 0} \quad \log\det(K) - \mathrm{trace}(KS)$$

$$\text{subject to} \quad K \in V(I_G)$$

- I_G gives linear constraints on K [Poll]

- This becomes an unconstrained optimization problem

Likelihood equations

We get the likelihood equations by taking the partial derivatives of the objective function:

$$\frac{1}{\det(K)} \frac{\partial}{\partial K_{ij}} \det(K) - (2 - \delta_{ij}) S_{ij} = 0,$$

where δ_{ij} is the Kronecker delta.

Code

```
R = QQ[k11,k12,k22,k23,k24,k33,k34,k44]
K = matrix {{k11,k12,0,0},{k12,k22,k23,k24},{0,k23,k33,k34},{0,k24,k34,k44}}
X = matrix for i to 3 list for j to 3 list random(30)
S = X*transpose(X)
M1 = jacobian(ideal(det(K)));
M2 = det(K)*jacobian(ideal(trace(K*S)));
I = ideal (M1-M2);
J = saturate(I,det(K))
```

```
ideal (15621672k44 - 255515, 15621672k34 + 46159, 15621672k33 - 39947, 15621672k24 + 134201, 15621672k23 + 22955, 1069537773480k22
- 17602462843, 68465k12 + 312, 136930k11 - 517)
```

Solutions

- For this graph, there is always one solution and it lies in the positive definite cone.
- For 4-cycle, there are five solutions out of which precisely one lies in the positive definite cone.
- Is there always one solution in the positive definite cone?
- Yes, this follows from a result for exponential families.

Exponential families

Prop: Let $\mathscr{M}$ be an exponential family with minimal sufficient statistics $T(x)$ and natural parameter $\eta \in N$, with density $f_\eta(x) = h(x)e^{\eta^t T(x) - A(\eta)}$. Then the likelihood function is strictly concave on N. Furthermore, the maximum likelihood estimate, if it exists, is the unique $\eta \in N$ satisfying

$$T(x) = \mathbb{E}_\eta[T(X)],$$

where x denotes the data vector.

MLE in Gaussian graphical models

Corollary: Assuming that the MLE exists, it is the unique positive definite matrix Σ satisfying

- $\Sigma \in V(I_G)$, and
- $\Sigma_{ij} = S_{ij}$ for all $(i, j) \in E$ or $i = j$.

MLE in Gaussian graphical models

The MLE is a point in the variety of

$$I = \langle \Sigma K - \mathrm{Id} \rangle + \langle K_{ij} : (i,j) \notin E \rangle + \langle \Sigma_{ij} - S_{ij} : (i,j) \in E \text{ or } i = j \rangle$$

Code

```
R = QQ[k11,k12,k22,k23,k24,k33,k34,k44,s11,s12,s13,s14,s22,s23,s24,s33,s34,s44]
K = matrix {{k11,k12,0,0},{k12,k22,k23,k24},{0,k23,k33,k34},{0,k24,k34,k44}}
Sigma = matrix {{s11,s12,s13,s14},{s12,s22,s23,s24},{s13,s23,s33,s34},{s14,s24,s34,s44}}
I1 = ideal (K*Sigma - identity(1))
X = matrix for i to 3 list for j to 3 list random(30)
S = X*transpose(X)
I2 = ideal(Sigma_(0,0)-S_(0,0),Sigma_(0,1)-S_(0,1),Sigma_(1,1)-S_(1,1),Sigma_(1,2)-S_(1,2),Sigma_(1,3)-S_(1,3),Sigma_(2,2)-S_(2,2),Sigma_(2,3)-S_(2,3),Sigma_(3,3)-S_(3,3))
I = I1 + I2
J = eliminate(I,{k11,k12,k22,k23,k24,k33,k34,k44})
```

```
ideal (s44 - 1110, s34 - 669, s33 - 430, s24 - 566, s23 - 394, s22 - 504, 9s14 - 3962, 9s13 - 2758, s12 - 392, s11 - 591)
```

Discrete graphical models

- Let X be a discrete random vector with state space $\mathscr{R} = \prod_{j=1}^{m} [r_j]$.
- Let $P = (p_{i_1 \dots i_m})$ denote the joint probabilities and $U = (u_{i_1 \dots i_m})$ the contingency table.
- The maximum likelihood estimation problem is

$$\max_{p \geq 0} \sum_{(i_1, \dots, i_m) \in \mathscr{R}} u_{i_1 \dots i_m} \log p_{i_1 \dots i_m}$$

$$\text{subject to} \quad p \in V\left(I_G + \left\langle \sum_{(i_1, \dots, i_m) \in \mathscr{R}} p_{i_1 \dots i_m} - 1 \right\rangle\right)$$

- [Poll]

Lagrange multipliers

- Recall that the method of Lagrange multipliers is used to solve the following constrained optimization problem:

$$\max f(x)$$

$$\text{subject to } g_i(x) = 0 \text{ for } i = 1, \ldots, k$$

- The Lagrangian of this optimization problem is

$$L(x, \lambda) = f(x) - \sum_{i=1}^{k} \lambda_i g_i(x).$$

Discrete graphical models

- Let $f_1, \ldots, f_r$ be generators of I_G.
- The Lagrangian for our optimization problem is:

$$L(x, \lambda) = \sum_{(i_1,\ldots,i_m)\in\mathcal{R}} u_{i_1\ldots i_m} \log p_{i_1\ldots i_m} - \lambda_0 \left(\sum_{(i_1,\ldots,i_m)\in\mathcal{R}} p_{i_1\ldots i_m} - 1 \right) - \sum_{j=1}^{r} \lambda_j f_j(x)$$

Lagrange multipliers

The constrained critical points of f are among the unconstrained critical points of L. Hence one has to solve

$$g_1 = 0, \ldots, g_k = 0,$$

$$\frac{\partial f}{\partial x_1} - \sum_{i=1}^{k} \lambda_i \frac{\partial g_i}{\partial x_1} = 0, \ldots, \frac{\partial f}{\partial x_m} - \sum_{i=1}^{k} \lambda_i \frac{\partial g_i}{\partial x_r} = 0$$

Discrete graphical models

$$\sum_{(i_1,\ldots,i_m)\in\mathscr{R}} p_{i_1\ldots i_m} - 1 = 0,$$

$$f_1 = 0, \ldots, f_r = 0,$$

$$\frac{u_{i_1\ldots i_r}}{p_{i_1\ldots i_r}} - \lambda_0 - \sum_{j=1}^{r} \lambda_j \frac{\partial f_j}{\partial p_{i_1\ldots i_r}} = 0 \text{ for all } (i_1, \ldots, i_m) \in \mathscr{R}$$

- One option is to use solve the above system.
- Another option is to use the following result for discrete exponential families.

Discrete exponential families

Cor: Let $A \subseteq \mathbb{Z}^{k\times r}$ such that $\mathbf{1} \in \text{rowspan}(\mathbf{A})$, let $h \in \mathbb{R}^r_{>0}$, and let u be the vector of counts from n i.i.d. samples. Then the maximum likelihood estimate in the log-linear model $\mathscr{M}_{A,h}$ given the data u is the unique solution, if it exists, to the equations

$$Au = nAp \text{ and } p \in \mathscr{M}_{A,h}.$$

Code

```
R1 = QQ[p_(0,0,0,0)..p_(1,1,1,1)]
R2 = QQ[p_(0,0,0,0)..p_(1,1,1,1),a_(0,0)..a_(1,1),b_(0,0,0)..b_(1,1,1)]
IF = ideal flatten flatten flatten for i to 1 list for j to 1 list for k to 1 list for l to 1 list p_(i,j,k,l)-a_(i,j)*b_(j,k,l)
JF = eliminate(IF,join(toList(a_(0,0)..a_(1,1)), toList(b_(0,0,0)..b_(1,1,1))))
JF = sub(JF,R1)
use R1
A = matrix{{ 1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0},
{0,0,0,0,1,1,1,1,0,0,0,0,0,0,0,0},
{0,0,0,0,0,0,0,0,1,1,1,1,0,0,0,0},
{0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1},
{1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0},
{0,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0},
{0,0,1,0,0,0,0,0,0,0,1,0,0,0,0,0},
{0,0,0,1,0,0,0,0,0,0,0,1,0,0,0,0},
{0,0,0,0,1,0,0,0,0,0,0,0,1,0,0,0},
{0,0,0,0,0,1,0,0,0,0,0,0,0,1,0,0},
{0,0,0,0,0,0,1,0,0,0,0,0,0,0,1,0},
{0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,1}}
U = transpose matrix {for i to 15 list random(30)}
P = transpose matrix {toList(p_(0,0,0,0)..p_(1,1,1,1))}
I = JF + ideal (A*U-A*P)
```

ML degree

Theorem: Let $\mathscr{M}_\Theta \subseteq \Delta_{r-1}$ be a statistical model. For generic data, the number of solutions to the score equations is independent of u.

Generic = data is outside a variety

Def: The number of solutions to the score equations for generic u is called the maximum likelihood degree (ML degree) of the parametric discrete statistical model $\mathscr{M}_\Theta$.

Chordal graphs

- A graph G is chordal if every induced cycle of length 4 or larger has a chord. [Poll]

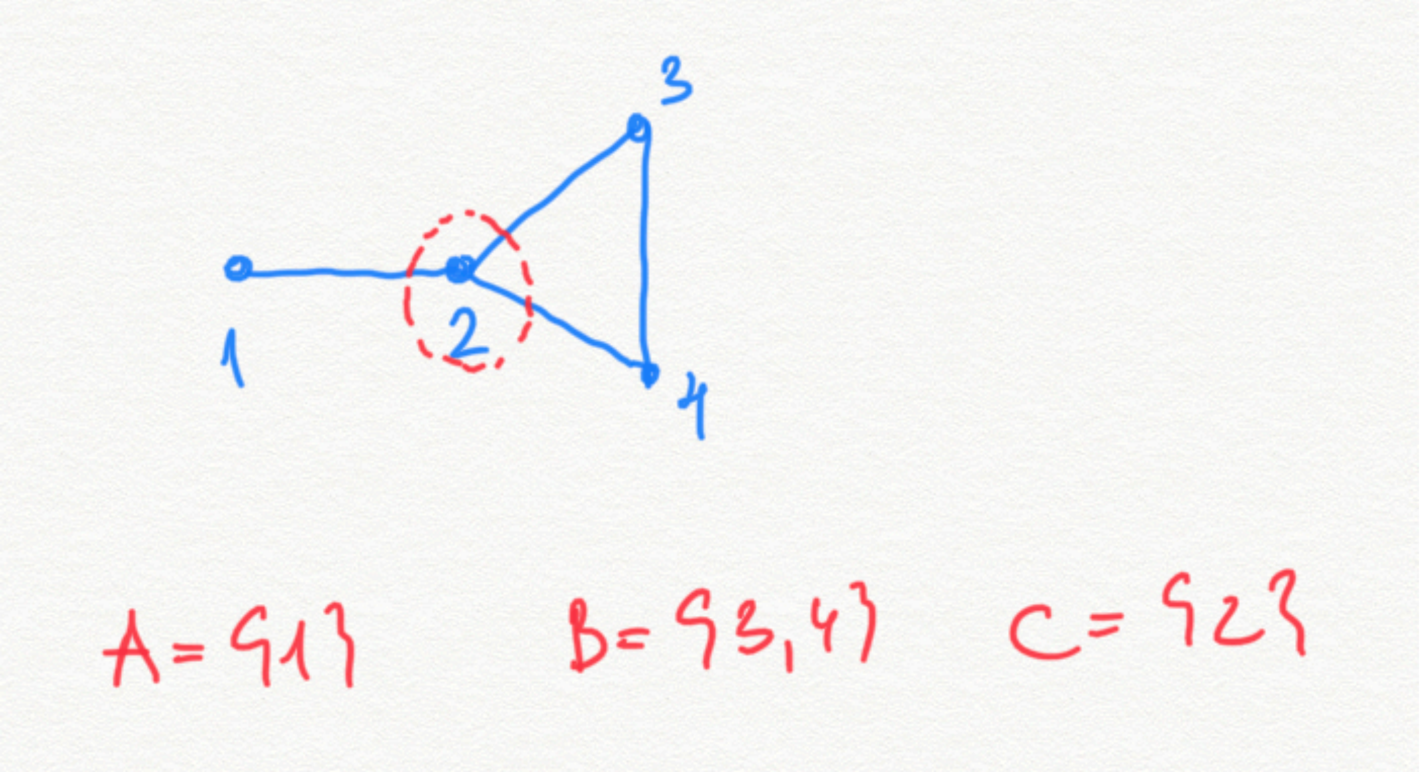
Theorem: The ML degree for a graphical model on G in the discrete or Gaussian setting is equal to one if and only if G is chordal.

- In this case, the MLE can be written as a rational function of data.

Chordal graphs

<u>Def:</u> The triple of vertices (A, B, C) forms a decomposition of a graph G if

- A, B, C are disjoint,
- A, B are non-empty,
- $V = A \cup B \cup C$,
- the induced graph G_C is complete, and
- C separates A from B (there are no edges between A and B).



Chordal graphs

Def: A graph is decomposable if it is complete or there exists a decomposition into decomposable subgraphs $G_{A \cup C}$ and $G_{B \cup C}$.

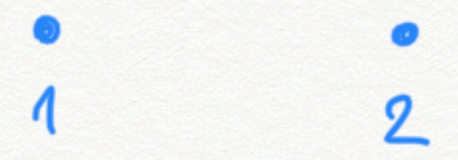
- By first finding decompositions of $G_{A \cup C}$ and $G_{B \cup C}$ and then finding decompositions of decomposed graphs, we end up with a clique decomposition $C_1, \ldots, C_r$ with separators $D_1, \ldots, D_k$.

- A graph is decomposable if and only if it is chordal.

Chordal graphs

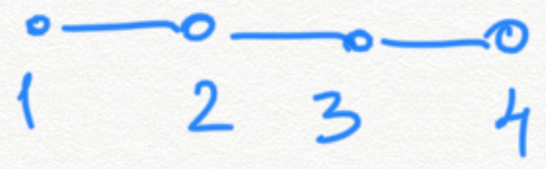
Which of the following graphs are given with their decomposition?

1.



$C_1=\{1\}$ $C_2=\{2\}$

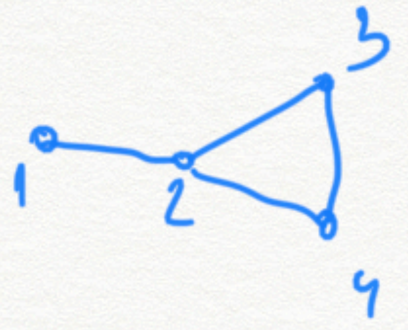
2.



$C_1=\{1,2\}$ $C_2=\{2,3\}$ $C_3=\{3,4\}$

$D_1=\{2\}$ $D_2=\{3\}$

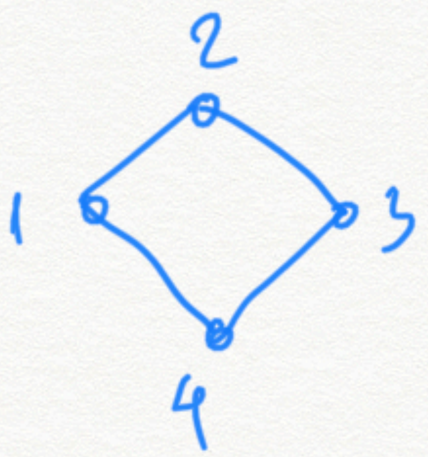
3.



$C_1=\{1,2\}$ $C_2=\{2,3,4\}$

$D_1=\{2\}$

4.



$C_1=\{1,2\}$ $C_2=\{2,3\}$ $C_3=\{3,4\}$ $C_4=\{4,1\}$

$D_1=\{1\}$ $D_2=\{2\}$ $D_3=\{3\}$ $D_4=\{4\}$

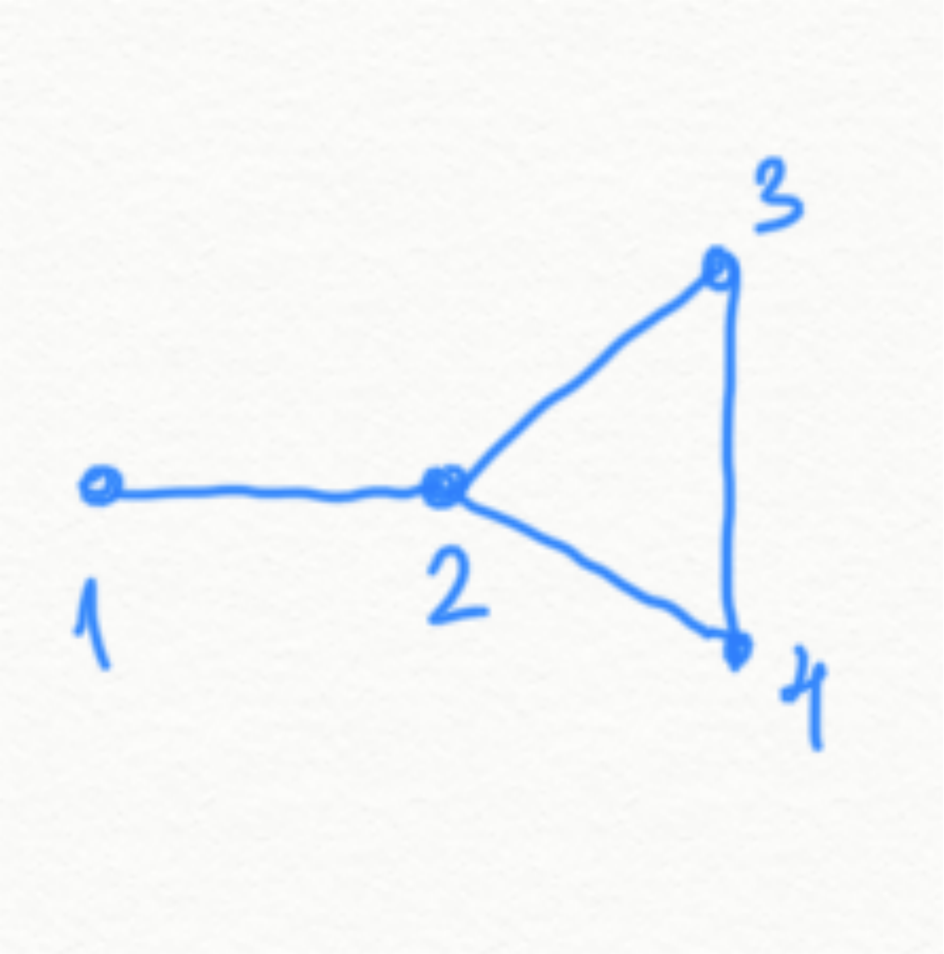
Chordal graphs

<u>Prop:</u> Let G be a chordal graph with clique decomposition $C_1, \ldots, C_r$ and with separators $D_1, \ldots, D_k$. Let $U = (u_{i_1 \ldots i_m})$ be the contingency table. The MLE in the corresponding graphical model is

$$v_{i_1 \ldots i_m} = \frac{\prod_{j=1}^{r} (u|_{C_j})|_{i_{C_j}}}{\prod_{j=1}^{k} (u|_{D_j})|_{i_{D_j}}} \text{ for all } (i_1, \ldots, i_m) \in \mathcal{X},$$

where $u|_F$ denotes the marginals over F.

Chordal graphs



- The clique decomposition of the graph is $C_1 = \{1,2\}$ and $C_2 = \{2,3,4\}$ with the separator $D_1 = \{2\}$.

- The MLE is given by the formula

$$v_{ijkl} = \frac{u_{ij++}u_{+jkl}}{u_{+j++}}.$$

- For non-decomposable models log-linear models, hill-climbing methods are used in practice to compute the MLE.

Learning the graph

- We have assumed that the graph is given
- One option to learn the graph is via constraint-based learning
- Given observed data, one can test which Markov properties hold and construct the graph from these results
- The result of each test is yes or no, which tells whether an edge is present or absent in the graph
- See the book: “Graphical Models with R” by Højsgaard, Edwards, and Lauritzen

Conclusion

- Both in the Gaussian and in the discrete setting, there is only one critical point of the likelihood function in the model and it is the MLE
- Special results for finding the MLE both for Gaussian and discrete graphical models (more generally to exponential families)
- The ML degree of a graphical model is one if and only if the graph is chordal
    - Formula for the MLE in the discrete case

Thank you!

- Thank you for attending and for your hard work!
- Please fill out the course survey
- Period III: Computational Algebraic Geometry (MS-E1142)

Master’s thesis topics

Topic 1: Toric fiber products and graphical models

- Toric fiber product is a construction that allows to construct from two ideals in smaller polynomial rings another ideal in a larger ring.

- In the case of graphical models, this means constructing the ideal of a graph from the ideals of subgraphs.

- Goal: In the Lauritzen's book "Graphical models", identify all results for MLE of graphical models that are special cases of the MLE result for toric fiber products.

- NB! This topic requires strong algebra background.

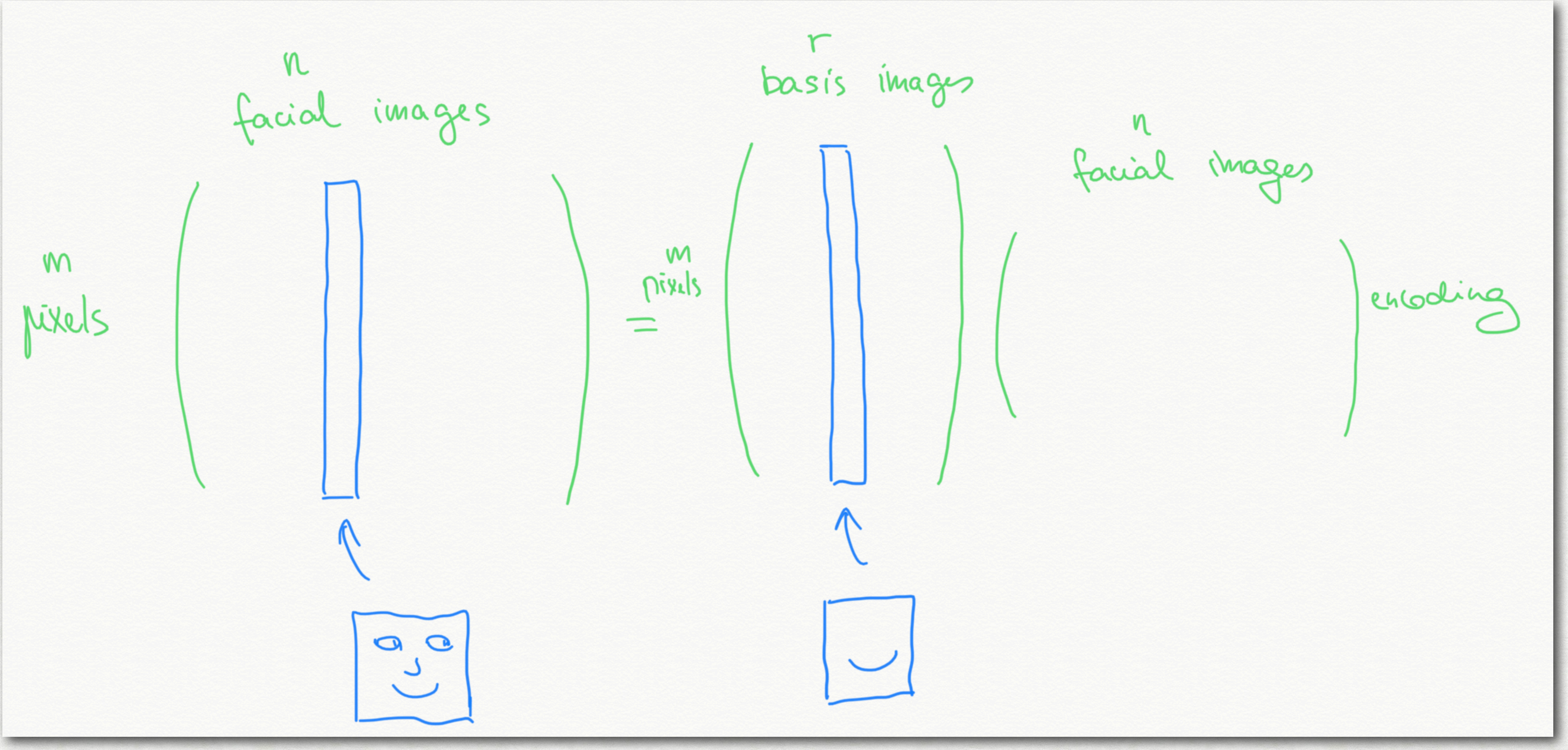
NONNEGATIVE FACTORIZATIONS AND RANK

Def: Given a matrix $M \in \mathbb{R}_{\geq 0}^{m \times n}$, a pair $(A, B) \in \mathbb{R}_{\geq 0}^{m \times r} \times \mathbb{R}_{\geq 0}^{r \times n}$ such that $M = AB$ is called a size-r nonnegative factorization of M.

$$\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

LEARNING THE PARTS OF FACES

- Lee and Seung, 1999



Topic 2: Uniqueness of NMF

- For many of the applications it is desirable that there exists a unique nonnegative matrix factorization (up to scalings and permutations).
- Together with Krone, we recently gave a necessary condition for uniqueness.
- Goal: Compare the necessary condition with two well-known sufficient conditions for uniqueness: separability and sufficiently scattered.

Topic 3: Size-2 nonnegative approximations

- Size-2 nonnegative factorizations are better understood than general case.
- Nevertheless, given a matrix M, it is not know which matrices A, B give the best size-2 nonnegative approximation AB to M.
- Goal: Study the best size-2 nonnegative factorizations for 3×4 and 4×4 matrices and explore whether conjectures in a recent paper with Sodomaco and Tsigaridas hold in these cases.

Topic 4: Deep nonnegative matrix factorizations in biology

- Nonnegative matrix factorizations are used in biology for studying the expression of genes in different tissues (e.g. healthy and cancer tissues)
- More generally one can define deep nonnegative matrix factorizations: $M = A_1A_2\ldots A_nB$, where all factors are nonnegative.
- Goal: Use deep nonnegative matrix factorizations for a biological dataset and study how to choose the sizes of matrices in the factorization.

Bachelor’s thesis topics

Topic 1: Rank-1 tensor completion for small tensors

- Tensors are higher dimensional analogues of matrices
- Whether a partial tensor can be completed to a rank-1 tensor depends generically only on the locations of observed entries
- Goal: For small tensors, study which partial tensors allow completion to a rank-1 tensor
- This topic requires the use of abstract algebra and in particular studying the symmetries of a tensor