# Likelihood Inference for Discrete and Gaussian Models

*Lectures on Algebraic Statistics* §2.1 by Drton, Sturmfels & Sullivant

**Speaker:** Felix Rydell

KTH Stockholm

March 8, 2021

# Overview

Goals for today:

1. Maximum Likelihood Estimation and Connections to Algebraic Geometry,
2. Discrete Models with Examples,
3. Gaussian Models.

# Overview

Goals for today:

1. Maximum Likelihood Estimation and Connections to Algebraic Geometry,
2. Discrete Models with Examples,
3. Gaussian Models.

# Overview

Goals for today:

1. Maximum Likelihood Estimation and Connections to Algebraic Geometry,
2. Discrete Models with Examples,
3. Gaussian Models.

# Overview

Goals for today:

1. Maximum Likelihood Estimation and Connections to Algebraic Geometry,
2. Discrete Models with Examples,
3. Gaussian Models.

# Maximum Likelihood Estimation

- Consider a statistical model $\mathcal{P}_\Theta = \{P_\theta : \theta \in \Theta\}$. Given identically distributed independent random variables $X^{(i)} \sim P_\theta$ for $i = 1, \ldots, n$, the *likelihood* function is given by:

$$L(\theta) := \prod_{i=1}^{n} p_\theta(X^{(i)}).$$

We want to find the $\theta$ that maximizes $L$; we may equivalently consider the *log-likelihood* function $\ell := \log L$. (Sometimes a multinomial coefficient is included in the definition of $L(\theta)$.)

- The ML-estimator is the random variable defined as

$$\hat{\theta} := \arg\max_{\theta \in \Theta} \ell(\theta) = \arg\max_{\theta \in \Theta} L(\theta)$$

Note that it is a random variable since it depends on $X^{(i)}$. The *maximum likelihood estimate* of $\theta$ given data $x^{(i)}$ is obtained as $\hat{\theta}$ when substituting $X^{(i)} = x^{(i)}$. Note that this $\theta$ is a parameter that maximizes the likelihood of observing $x^{(1)}, \ldots, x^{(n)}$.

# Maximum Likelihood Estimation

- Consider a statistical model $\mathcal{P}_\Theta = \{P_\theta : \theta \in \Theta\}$. Given identically distributed independent random variables $X^{(i)} \sim P_\theta$ for $i = 1, \dots, n$, the *likelihood* function is given by:

$$L(\theta) := \prod_{i=1}^{n} p_\theta(X^{(i)}).$$

We want to find the $\theta$ that maximizes $L$; we may equivalently consider the *log-likelihood* function $\ell := \log L$. (Sometimes a multinomial coefficient is included in the definition of $L(\theta)$.)

- The ML-estimator is the random variable defined as

$$\hat{\theta} := \arg\max_{\theta \in \Theta} \ell(\theta) = \arg\max_{\theta \in \Theta} L(\theta)$$

Note that it is a random variable since it depends on $X^{(i)}$. The *maximum likelihood estimate* of $\theta$ given data $x^{(i)}$ is obtained as $\hat{\theta}$ when substituting $X^{(i)} = x^{(i)}$. Note that this $\theta$ is a parameter that maximizes the likelihood of observing $x^{(1)}, \dots, x^{(n)}$.

# Algebraic Insights – Part 1

- Sullivant (*Algebraic Statistics*, page 2) gives the following dictionary:

| *Probability/Statistics* | *Algebra/Geometry* |
|---|---|
| Probability distribution | Point in $\Delta$ |
| Statistical model | Semi-algebraic set |
| Exponential Family | Toric Variety |
| Conditional Inference | Lattice points in polytopes |
| **Maximum likelihood estimation** | **Polynomial optimization** |
| Model selection | Geometry of singularities |
| Multivariate Gaussian distribution | Spectrahedral geometry |
| Phylogenetic model | Tensor networks |
| MAP estimates | Tropical geometry |

# Algebraic Insights – Part 2

- The *score* equations (also known as *critical* equations) are given by setting $\partial\ell/\partial\theta_i = 0$ for each $i$. A point satisfying the equation is a *critical* point. Any local maximum is a critical point under the assumption that the parameter space $\Theta$ is open.

- For example, let

$$\log p_\theta(X) = \log q_1(\theta) + q_2(\theta) \tag{1}$$

be a univariate quotient of polynomials of rational coefficients (meaning $q_i \in \mathbb{Q}(\theta)$.) The score equation is given by $q_1'(\theta)/q_1(\theta) + q_2'(\theta) = 0$. This is an *algebraic* expression.

- Recall that solutions to algebraic equations is the main subject of algebraic geometry.

- We define the *ML-degree* of a (possibly multivariate) statistical model of the form (1) to be the number of complex solutions to the score equations for generic data. Note that this bounds the number of real solutions.

# Algebraic Insights – Part 2

- The *score* equations (also known as *critical* equations) are given by setting $\partial\ell/\partial\theta_i = 0$ for each $i$. A point satisfying the equation is a *critical* point. Any local maximum is a critical point under the assumption that the parameter space $\Theta$ is open.

- For example, let

$$\log p_\theta(X) = \log q_1(\theta) + q_2(\theta) \tag{1}$$

be a univariate quotient of polynomials of rational coefficients (meaning $q_i \in \mathbb{Q}(\theta)$.) The score equation is given by $q_1'(\theta)/q_1(\theta) + q_2'(\theta) = 0$. This is an *algebraic* expression.

- Recall that solutions to algebraic equations is the main subject of algebraic geometry.

- We define the *ML-degree* of a (possibly multivariate) statistical model of the form (1) to be the number of complex solutions to the score equations for generic data. Note that this bounds the number of real solutions.

## Algebraic Insights – Part 2

- The *score* equations (also known as *critical* equations) are given by setting $\partial\ell/\partial\theta_i = 0$ for each $i$. A point satisfying the equation is a *critical* point. Any local maximum is a critical point under the assumption that the parameter space $\Theta$ is open.

- For example, let

$$\log p_\theta(X) = \log q_1(\theta) + q_2(\theta) \tag{1}$$

be a univariate quotient of polynomials of rational coefficients (meaning $q_i \in \mathbb{Q}(\theta)$.) The score equation is given by $q_1'(\theta)/q_1(\theta) + q_2'(\theta) = 0$. This is an *algebraic* expression.

- Recall that solutions to algebraic equations is the main subject of algebraic geometry.

- We define the *ML-degree* of a (possibly multivariate) statistical model of the form (1) to be the number of complex solutions to the score equations for generic data. Note that this bounds the number of real solutions.

# Algebraic Insights – Part 2

- The *score* equations (also known as *critical* equations) are given by setting $\partial \ell / \partial \theta_i = 0$ for each $i$. A point satisfying the equation is a *critical* point. Any local maximum is a critical point under the assumption that the parameter space $\Theta$ is open.

- For example, let

$$\log p_\theta(X) = \log q_1(\theta) + q_2(\theta) \tag{1}$$

be a univariate quotient of polynomials of rational coefficients (meaning $q_i \in \mathbb{Q}(\theta)$.) The score equation is given by $q_1'(\theta)/q_1(\theta) + q_2'(\theta) = 0$. This is an *algebraic* expression.

- Recall that solutions to algebraic equations is the main subject of algebraic geometry.

- We define the *ML-degree* of a (possibly multivariate) statistical model of the form (1) to be the number of complex solutions to the score equations for generic data. Note that this bounds the number of real solutions.

# Generic Data

- What is *generic* data? In algebraic geometry, a point in $\mathbb{K}^n$ (for some field $\mathbb{K}$) is generic if it lies in some fixed non-empty Zariski open set (this is a set on the form $\mathbb{K}^n \setminus \mathcal{V}(I)$ for an ideal $\langle 0 \rangle \neq I \subseteq \mathbb{K}[x_1, \dots, x_n]$.) The dimension of an open non-empty set is $n$ and the dimension of the set of non-generic points is at most $n-1$.

- We can think of generic data as "random" data; randomly chosen data has probability 1 of being generic.

- A statement that is true for generic points in $\mathbb{C}^n$ is also true for generic points in $\mathbb{R}^n$. This is essentially because the Zariski closure of $\mathbb{R}^n$ in $\mathbb{C}^n$ is $\mathbb{C}^n$ (the Zariski closure of $X$ is the smallest set of the form $\mathcal{V}(I)$ that contains $X$.)

# Generic Data

- What is *generic* data? In algebraic geometry, a point in $\mathbb{K}^n$ (for some field $\mathbb{K}$) is generic if it lies in some fixed non-empty Zariski open set (this is a set on the form $\mathbb{K}^n \setminus \mathcal{V}(I)$ for an ideal $\langle 0 \rangle \neq I \subseteq \mathbb{K}[x_1, \dots, x_n]$.) The dimension of an open non-empty set is $n$ and the dimension of the set of non-generic points is at most $n-1$.

- We can think of generic data as “random” data; randomly chosen data has probability $1$ of being generic.

- A statement that is true for generic points in $\mathbb{C}^n$ is also true for generic points in $\mathbb{R}^n$. This is essentially because the Zariski closure of $\mathbb{R}^n$ in $\mathbb{C}^n$ is $\mathbb{C}^n$ (the Zariski closure of $X$ is the smallest set of the form $\mathcal{V}(I)$ that contains $X$.)

# Generic Data

- What is *generic* data? In algebraic geometry, a point in $\mathbb{K}^n$ (for some field $\mathbb{K}$) is generic if it lies in some fixed non-empty Zariski open set (this is a set on the form $\mathbb{K}^n \setminus \mathcal{V}(I)$ for an ideal $\langle 0 \rangle \neq I \subseteq \mathbb{K}[x_1, \ldots, x_n]$.) The dimension of an open non-empty set is $n$ and the dimension of the set of non-generic points is at most $n-1$.

- We can think of generic data as "random" data; randomly chosen data has probability $1$ of being generic.

- A statement that is true for generic points in $\mathbb{C}^n$ is also true for generic points in $\mathbb{R}^n$. This is essentially because the Zariski closure of $\mathbb{R}^n$ in $\mathbb{C}^n$ is $\mathbb{C}^n$ (the Zariski closure of $X$ is the smallest set of the form $\mathcal{V}(I)$ that contains $X$.)

# Saturation – Part 1

• Let us return to the equation $q_1'(\theta)/q_1(\theta) + q_2'(\theta) = 0$. If we want to solve it we might start by writing $q_1 = f_1/g_1$ and $q_2 = f_2/g_2$ so that we get a polynomial equation:

$$\frac{\frac{f_1'}{g_1} - \frac{f_1 g_1'}{g_1^2}}{f_1/g_1} - \left(\frac{f_2'}{g_2} - \frac{f_2 g_2'}{g_2^2}\right) = 0 \Rightarrow \tag{2}$$

$$\Rightarrow g_2^2(f_1' g_1 - f_1 g_1') - f_1 g_1 (f_2' g_2 - f_2 g_2') = 0. \tag{3}$$

• We now have a nice polynomial equation, but we have gained solutions that we did not have before. For example, points $x$ such that $g_2(x) = f_1(x) = 0$ solves the polynomial equation (3), but clearly not the rational equation (2).

• Problems of this sort are solved using *saturation* of ideals. In this case we study the ideal $I = \langle g_2^2(f_1' g_1 - f_1 g_1') - f_1 g_1 (f_2' g_2 - f_2 g_2') \rangle$ and the ideal of "bad" solutions $J = \langle g_1 f_1 g_2 \rangle$.

# Saturation – Part 1

• Let us return to the equation $q_1'(\theta)/q_1(\theta) + q_2'(\theta) = 0$. If we want to solve it we might start by writing $q_1 = f_1/g_1$ and $q_2 = f_2/g_2$ so that we get a polynomial equation:

$$\frac{\frac{f_1'}{g_1} - \frac{f_1 g_1'}{g_1^2}}{f_1/g_1} - \left(\frac{f_2'}{g_2} - \frac{f_2 g_2'}{g_2^2}\right) = 0 \Rightarrow \tag{2}$$

$$\Rightarrow g_2^2(f_1'g_1 - f_1g_1') - f_1g_1(f_2'g_2 - f_2g_2') = 0. \tag{3}$$

• We now have a nice polynomial equation, but we have gained solutions that we did not have before. For example, points $x$ such that $g_2(x) = f_1(x) = 0$ solves the polynomial equation (3), but clearly not the rational equation (2).

• Problems of this sort are solved using *saturation* of ideals. In this case we study the ideal $I = \langle g_2^2(f_1'g_1 - f_1g_1') - f_1g_1(f_2'g_2 - f_2g_2')\rangle$ and the ideal of "bad" solutions $J = \langle g_1 f_1 g_2 \rangle$.

# Saturation – Part 1

- Let us return to the equation $q_1'(\theta)/q_1(\theta) + q_2'(\theta) = 0$. If we want to solve it we might start by writing $q_1 = f_1/g_1$ and $q_2 = f_2/g_2$ so that we get a polynomial equation:

$$\frac{\frac{f_1'}{g_1} - \frac{f_1 g_1'}{g_1^2}}{f_1/g_1} - \left(\frac{f_2'}{g_2} - \frac{f_2 g_2'}{g_2^2}\right) = 0 \Rightarrow \tag{2}$$

$$\Rightarrow g_2^2(f_1' g_1 - f_1 g_1') - f_1 g_1 (f_2' g_2 - f_2 g_2') = 0. \tag{3}$$

- We now have a nice polynomial equation, but we have gained solutions that we did not have before. For example, points $x$ such that $g_2(x) = f_1(x) = 0$ solves the polynomial equation (3), but clearly not the rational equation (2).

- Problems of this sort are solved using *saturation* of ideals. In this case we study the ideal $I = \langle g_2^2(f_1' g_1 - f_1 g_1') - f_1 g_1(f_2' g_2 - f_2 g_2')\rangle$ and the ideal of "bad" solutions $J = \langle g_1 f_1 g_2 \rangle$.

# Saturation – Part 2

- Let $I$ and $J$ be two ideals of a ring $R$. We define the ideal quotient as follows

$$(I : J) := \{r \in R : rJ \subseteq I\}.$$

- For a noetherian ring $R$ (the polynomial rings $\mathbb{Q}[x], \mathbb{R}[x], \mathbb{C}[x]$, are noetherian,) consider the inclusion of ideals

$$(I : J) \subseteq (I : J^2) \subseteq (I : J^3) \subseteq \cdots .$$

The chain stabilizes at some $(I : J^N)$ and we call this the saturation of $I$ with respect to $J$ and write $(I : J^\infty)$ or $\mathrm{sat}(I, J)$.

**Proposition (Ideals, Varieties and Algorithms, Theorem 7 p. 195)**

Let $\mathcal{V}(J), \mathcal{V}(I)$ be two algebraic sets defined by ideals. Then

$$\overline{\mathcal{V}(I) \setminus \mathcal{V}(J)}^{\mathrm{Zar}} = \mathcal{V}(\mathrm{sat}(I, J)).$$

# Saturation – Part 2

- Let $I$ and $J$ be two ideals of a ring $R$. We define the ideal quotient as follows
$$(I:J) := \{r \in R : rJ \subseteq I\}.$$
- For a noetherian ring $R$ (the polynomial rings $\mathbb{Q}[x], \mathbb{R}[x], \mathbb{C}[x]$, are noetherian,) consider the inclusion of ideals
$$(I:J) \subseteq (I:J^2) \subseteq (I:J^3) \subseteq \cdots .$$

The chain stabilizes at some $(I:J^N)$ and we call this the saturation of $I$ with respect to $J$ and write $(I:J^\infty)$ or $\mathrm{sat}(I,J)$.

**Proposition (Ideals, Varieties and Algorithms, Theorem 7 p. 195)**

Let $\mathcal{V}(J), \mathcal{V}(I)$ be two algebraic sets defined by ideals. Then
$$\overline{\mathcal{V}(I) \setminus \mathcal{V}(J)}^{\mathrm{Zar}} = \mathcal{V}(\mathrm{sat}(I,J)).$$

# Saturation – Part 2

- Let $I$ and $J$ be two ideals of a ring $R$. We define the ideal quotient as follows

$$(I : J) := \{r \in R : rJ \subseteq I\}.$$

- For a noetherian ring $R$ (the polynomial rings $\mathbb{Q}[x], \mathbb{R}[x], \mathbb{C}[x]$, are noetherian,) consider the inclusion of ideals

$$(I : J) \subseteq (I : J^2) \subseteq (I : J^3) \subseteq \cdots.$$

The chain stabilizes at some $(I : J^N)$ and we call this the saturation of $I$ with respect to $J$ and write $(I : J^\infty)$ or $\mathrm{sat}(I, J)$.

**Proposition (Ideals, Varieties and Algorithms, Theorem 7 p. 195)**

Let $\mathcal{V}(J), \mathcal{V}(I)$ be two algebraic sets defined by ideals. Then

$$\overline{\mathcal{V}(I) \setminus \mathcal{V}(J)}^{\mathrm{Zar}} = \mathcal{V}(\mathrm{sat}(I, J)).$$

# Saturation – Part 3

## Example (2.1.3)

- Consider the score equations with table of counts $u$ and parameters $\lambda_i$

$$\frac{u_1 + u_{12}}{\lambda_1} + \frac{u_{12}}{\lambda_1 + \lambda_2 + 2} - \frac{u_2 + u_{12}}{\lambda_1 + 1} - \frac{u_0 + u_1 + u_2 + u_{12}}{\lambda_1 + \lambda_2 + 1} = 0$$

$$\frac{u_1 + u_{12}}{\lambda_2} + \frac{u_{12}}{\lambda_1 + \lambda_2 + 2} - \frac{u_1 + u_{12}}{\lambda_2 + 1} - \frac{u_0 + u_1 + u_2 + u_{12}}{\lambda_1 + \lambda_2 + 1} = 0$$

- In a software like `Macaulay2` or `Singular`, we let $I$ be the ideal generated by the two equations above after clearing denominators. Let $J$ be the ideal generated by all the denominators:

$$J := \langle \lambda_1\lambda_2(\lambda_1 + 1)(\lambda_2 + 1)(\lambda_1 + \lambda_2 + 1)(\lambda_1 + \lambda_2 + 2) \rangle$$

- As explained previously, in $\mathcal{V}(I)$, we have to many points. The ideal that corresponds to the system of Example 2.1.3 is given by $\mathrm{sat}(I, J) = (I, J^{\infty})$; this ideal describes the set of solutions that we are interested in.

# Saturation – Part 3

## Example (2.1.3)

- Consider the score equations with table of counts $u$ and parameters $\lambda_i$

$$\frac{u_1+u_{12}}{\lambda_1}+\frac{u_{12}}{\lambda_1+\lambda_2+2}-\frac{u_2+u_{12}}{\lambda_1+1}-\frac{u_0+u_1+u_2+u_{12}}{\lambda_1+\lambda_2+1}=0$$

$$\frac{u_1+u_{12}}{\lambda_2}+\frac{u_{12}}{\lambda_1+\lambda_2+2}-\frac{u_1+u_{12}}{\lambda_2+1}-\frac{u_0+u_1+u_2+u_{12}}{\lambda_1+\lambda_2+1}=0$$

- In a software like `Macaulay2` or `Singular`, we let $I$ be the ideal generated by the two equations above after clearing denominators. Let $J$ be the ideal generated by all the denominators:

$$J := \langle \lambda_1\lambda_2(\lambda_1+1)(\lambda_2+1)(\lambda_1+\lambda_2+1)(\lambda_1+\lambda_2+2)\rangle$$

- As explained previously, in $\mathcal{V}(I)$, we have to many points. The ideal that corresponds to the system of Example 2.1.3 is given by $\mathrm{sat}(I,J) = (I, J^{\infty})$; this ideal describes the set of solutions that we are interested in.

# Saturation – Part 3

## Example (2.1.3)

- Consider the score equations with table of counts $u$ and parameters $\lambda_i$

$$\frac{u_1+u_{12}}{\lambda_1}+\frac{u_{12}}{\lambda_1+\lambda_2+2}-\frac{u_2+u_{12}}{\lambda_1+1}-\frac{u_0+u_1+u_2+u_{12}}{\lambda_1+\lambda_2+1}=0$$

$$\frac{u_1+u_{12}}{\lambda_2}+\frac{u_{12}}{\lambda_1+\lambda_2+2}-\frac{u_1+u_{12}}{\lambda_2+1}-\frac{u_0+u_1+u_2+u_{12}}{\lambda_1+\lambda_2+1}=0$$

- In a software like `Macaulay2` or `Singular`, we let $I$ be the ideal generated by the two equations above after clearing denominators. Let $J$ be the ideal generated by all the denominators:

$$J := \langle \lambda_1\lambda_2(\lambda_1+1)(\lambda_2+1)(\lambda_1+\lambda_2+1)(\lambda_1+\lambda_2+2)\rangle$$

- As explained previously, in $\mathcal{V}(I)$, we have to many points. The ideal that corresponds to the system of Example 2.1.3 is given by $\mathrm{sat}(I,J) = (I, J^\infty)$; this ideal describes the set of solutions that we are interested in.

# Discrete Models – Part 1

- A *parametric discrete model* is given by an open subset $\Theta \subseteq \mathbb{R}^d$ and a rational map $g : \Theta \to \Delta_{k-1}$, meaning each coordinate $g_i$ is a rational function. We consider:

$$\ell(\theta) = \log L(\theta) = \log \prod_{i=1}^{k} g_i(\theta)^{u_i} = \sum u_i \log g_i(\theta),$$

for a table of counts $u_i = \#\{j : X^{(j)} = i\}$.

## Example (2.1.2, (1/3))

- The parametrization of the independence model $\mathcal{M}_{X \perp\!\!\!\perp Y}$ is the map $g : \Delta_{r-1} \times \Delta_{c-1} \to \Delta_{rc-1}, (\alpha, \beta) \mapsto (\alpha_i \beta_j)$.
- For a table of counts $u \in \mathbb{N}^{r \times c}$, we have the log-likelihood

$$\ell(\alpha, \beta) = \sum u_{ij} \log(\alpha_i \beta_j) = \sum_i u_{i+} \log \alpha_i + \sum_j u_{+j} \log \beta_j,$$

where $u_{i+}, u_{+j}$ are the familiar marginal sums.

# Discrete Models – Part 1

- A *parametric discrete model* is given by an open subset $\Theta \subseteq \mathbb{R}^d$ and a rational map $g : \Theta \to \Delta_{k-1}$, meaning each coordinate $g_i$ is a rational function. We consider:

$$\ell(\theta) = \log L(\theta) = \log \prod_{i=1}^{k} g_i(\theta)^{u_i} = \sum u_i \log g_i(\theta),$$

for a table of counts $u_i = \#\{j : X^{(j)} = i\}$.

## Example (2.1.2, (1/3))

- The parametrization of the independence model $\mathcal{M}_{X \perp\!\!\!\perp Y}$ is the map $g : \Delta_{r-1} \times \Delta_{c-1} \to \Delta_{rc-1}, (\alpha, \beta) \mapsto (\alpha_i \beta_j)$.
- For a table of counts $u \in \mathbb{N}^{r \times c}$, we have the log-likelihood

$$\ell(\alpha, \beta) = \sum u_{ij} \log(\alpha_i \beta_j) = \sum_i u_{i+} \log \alpha_i + \sum_j u_{+j} \log \beta_j,$$

where $u_{i+}, u_{+j}$ are the familiar marginal sums.

# Discrete Models – Part 1

- A *parametric discrete model* is given by an open subset $\Theta \subseteq \mathbb{R}^d$ and a rational map $g : \Theta \to \Delta_{k-1}$, meaning each coordinate $g_i$ is a rational function. We consider:

$$\ell(\theta) = \log L(\theta) = \log \prod_{i=1}^{k} g_i(\theta)^{u_i} = \sum u_i \log g_i(\theta),$$

for a table of counts $u_i = \#\{j : X^{(j)} = i\}$.

## Example (2.1.2, (1/3))

- The parametrization of the independence model $\mathcal{M}_{X \perp\!\!\!\perp Y}$ is the map $g : \Delta_{r-1} \times \Delta_{c-1} \to \Delta_{rc-1}, (\alpha, \beta) \mapsto (\alpha_i \beta_j)$.
- For a table of counts $u \in \mathbb{N}^{r \times c}$, we have the log-likelihood

$$\ell(\alpha, \beta) = \sum u_{ij} \log(\alpha_i \beta_j) = \sum_i u_{i+} \log \alpha_i + \sum_j u_{+j} \log \beta_j,$$

where $u_{i+}, u_{+j}$ are the familiar marginal sums.

# Discrete Models – Part 2

## Example (2.1.2, (2/3))

- Observe that $\alpha_i, \beta_j$ are not independent of each other since they need to sum to 1. We resolve this by letting

$$\alpha_r = 1 - \sum_{i=1}^{r-1} \alpha_i, \ \beta_c = 1 - \sum_{j=1}^{c-1} \beta_j$$

- The score equations are

$$\frac{\partial \ell(\alpha, \beta)}{\partial \alpha_i} = \frac{u_{i+}}{\alpha_i} - \frac{u_{r+}}{1 - \sum_{k=1}^{r-1} \alpha_k} = 0, \ \forall i = 1, \ldots, r-1$$

$$\frac{\partial \ell(\alpha, \beta)}{\partial \beta_j} = \frac{u_{+j}}{\beta_j} - \frac{u_{+c}}{1 - \sum_{k=1}^{c-1} \beta_k} = 0, \ \forall j = 1, \ldots, c-1$$

# Discrete Models – Part 2

## Example (2.1.2, (2/3))

- Observe that $\alpha_i, \beta_j$ are not independent of each other since they need to sum to 1. We resolve this by letting

$$\alpha_r = 1 - \sum_{i=1}^{r-1} \alpha_i, \;\; \beta_c = 1 - \sum_{j=1}^{c-1} \beta_j$$

- The score equations are

$$\frac{\partial \ell(\alpha, \beta)}{\partial \alpha_i} = \frac{u_{i+}}{\alpha_i} - \frac{u_{r+}}{1 - \sum_{k=1}^{r-1} \alpha_k} = 0, \;\; \forall i = 1, \dots, r-1$$

$$\frac{\partial \ell(\alpha, \beta)}{\partial \beta_j} = \frac{u_{+j}}{\beta_j} - \frac{u_{+c}}{1 - \sum_{k=1}^{c-1} \beta_k} = 0, \;\; \forall j = 1, \dots, c-1$$

# Discrete Models – Part 3

## Example (2.1.2, (3/3))

- Clearing denominators gives systems of linear equations. For example, if $\alpha' = (\alpha_1, \dots, \alpha_{r-1})$

$$\begin{bmatrix} u_{1+} + u_{r+} & u_{1+} & \cdots & u_{1+} \\ u_{2+} & u_{2+} + u_{r+} & \cdots & u_{2+} \\ \vdots & \vdots & \ddots & \vdots \\ u_{r-1+} & u_{r-1+} & \cdots & u_{r-1+} + u_{r+} \end{bmatrix} \alpha' = \begin{bmatrix} u_{1+} \\ u_{2+} \\ \vdots \\ u_{r-1+} \end{bmatrix}$$

- Under the assumption that $u_{r+} > 0$, the matrix is full-rank and there is a unique solution. One can check that the following is a solution (the MLE):

$$\hat{\alpha}_i = \frac{u_{i+}}{u_{++}}, \; \hat{\beta}_j = \frac{u_{+j}}{u_{++}}.$$

- “Having maximum likelihood degree one can be expressed equivalently by saying that the ML estimate is a rational function of the data.”

# Discrete Models – Part 3

## Example (2.1.2, (3/3))

- Clearing denominators gives systems of linear equations. For example, if $\alpha' = (\alpha_1, \ldots, \alpha_{r-1})$

$$\begin{bmatrix} u_{1+} + u_{r+} & u_{1+} & \cdots & u_{1+} \\ u_{2+} & u_{2+} + u_{r+} & \cdots & u_{2+} \\ \vdots & \vdots & \ddots & \vdots \\ u_{r-1+} & u_{r-1+} & \cdots & u_{r-1+} + u_{r+} \end{bmatrix} \alpha' = \begin{bmatrix} u_{1+} \\ u_{2+} \\ \vdots \\ u_{r-1+} \end{bmatrix}$$

- Under the assumption that $u_{r+} > 0$, the matrix is full-rank and there is a unique solution. One can check that the following is a solution (the MLE):

$$\hat{\alpha}_i = \frac{u_{i+}}{u_{++}}, \ \hat{\beta}_j = \frac{u_{+j}}{u_{++}}.$$

- “Having maximum likelihood degree one can be expressed equivalently by saying that the ML estimate is a rational function of the data.”

# Discrete Models – Part 3

## Example (2.1.2, (3/3))

- Clearing denominators gives systems of linear equations. For example, if $\alpha' = (\alpha_1, \ldots, \alpha_{r-1})$

$$\begin{bmatrix} u_{1+} + u_{r+} & u_{1+} & \cdots & u_{1+} \\ u_{2+} & u_{2+} + u_{r+} & \cdots & u_{2+} \\ \vdots & \vdots & \ddots & \vdots \\ u_{r-1+} & u_{r-1+} & \cdots & u_{r-1+} + u_{r+} \end{bmatrix} \alpha' = \begin{bmatrix} u_{1+} \\ u_{2+} \\ \vdots \\ u_{r-1+} \end{bmatrix}$$

- Under the assumption that $u_{r+} > 0$, the matrix is full-rank and there is a unique solution. One can check that the following is a solution (the MLE):

$$\hat{\alpha}_i = \frac{u_{i+}}{u_{++}}, \ \hat{\beta}_j = \frac{u_{+j}}{u_{++}}.$$

- "Having maximum likelihood degree one can be expressed equivalently by saying that the ML estimate is a rational function of the data."

# Birch's Theorem – Part 1

## Proposition (2.1.5, Birch's Theorem)

Let $A \in \mathbb{N}^{d\times k}$ and $u \in \mathbb{N}^k$. The ML-estimate of the frequencies $\hat{u}$ in $\mathcal{M}_A$ is the unique non-negative solution to $A\hat{u} = Au$ for $\hat{u} \in \mathcal{V}(I_A)$.

- The MLE $\hat{u}$ differs from the MLE $\hat{p}$ by a constant: $\hat{u} = n\hat{p}$, where $n = \sum u_i$.

- Recall that $I_A$ is the toric ideal $\langle p^z - p^{z'} : z, z' \in \mathbb{N}^k,\ z - z' \in \ker_{\mathbb{Z}} A\rangle$, and $\mathcal{M}_A := \{p \in \Delta : \log p \in \text{rowspan} A\}$. Recall that the Zariski closure of $\mathcal{M}_A$ is $\mathcal{V}(I_A)$.

## Proof (1/2).

- Let $b_1, \ldots, b_l$ be a basis for $\ker_{\mathbb{Z}} A$. Observe that $p \in \mathcal{M}_A$ if and only if $\log p = A^T x$ and $p \in \Delta$ if and only if $b_j^T \log p = 0$ and $\sum p_i = 1$. (For example, $\log p = A^T x$ implies $b_j^T \log p = (Ab_j)^T x = 0$.)
- Let $v$ be the non-negative estimated table of counts. The expression $u^T \log v$ is up to a constant equal to $\log \prod p_i^{u_i}$, if we let $p_i = v_i/n$.
- We wish to maximize is $u^T \log v$, subject to $b_j^T \log v = 0$ for all $j = 1, \ldots, l$ and $\sum v_i = n$.

# Birch's Theorem – Part 1

## Proposition (2.1.5, Birch's Theorem)

Let $A \in \mathbb{N}^{d\times k}$ and $u \in \mathbb{N}^k$. The ML-estimate of the frequencies $\hat{u}$ in $\mathcal{M}_A$ is the unique non-negative solution to $A\hat{u} = Au$ for $\hat{u} \in \mathcal{V}(I_A)$.

- The MLE $\hat{u}$ differs from the MLE $\hat{p}$ by a constant: $\hat{u} = n\hat{p}$, where $n = \sum u_i$.

- Recall that $I_A$ is the toric ideal $\langle p^z - p^{z'} : z, z' \in \mathbb{N}^k,\ z - z' \in \ker_{\mathbb{Z}} A\rangle$, and $\mathcal{M}_A := \{p \in \Delta : \log p \in \text{rowspan} A\}$. Recall that the Zariski closure of $\mathcal{M}_A$ is $\mathcal{V}(I_A)$.

## Proof (1/2).

- Let $b_1, \ldots, b_l$ be a basis for $\ker_{\mathbb{Z}} A$. Observe that $p \in \mathcal{M}_A$ if and only if $\log p = A^T x$ and $p \in \Delta$ if and only if $b_j^T \log p = 0$ and $\sum p_i = 1$. (For example, $\log p = A^T x$ implies $b_j^T \log p = (Ab_j)^T x = 0$.)
- Let $v$ be the non-negative estimated table of counts. The expression $u^T \log v$ is up to a constant equal to $\log \prod p_i^{u_i}$, if we let $p_i = v_i/n$.
- We wish to maximize is $u^T \log v$, subject to $b_j^T \log v = 0$ for all $j = 1, \ldots, l$ and $\sum v_i = n$.

# Birch's Theorem – Part 1

## Proposition (2.1.5, Birch's Theorem)

Let $A \in \mathbb{N}^{d\times k}$ and $u \in \mathbb{N}^k$. The ML-estimate of the frequencies $\hat{u}$ in $\mathcal{M}_A$ is the unique non-negative solution to $A\hat{u} = Au$ for $\hat{u} \in \mathcal{V}(I_A)$.

- The MLE $\hat{u}$ differs from the MLE $\hat{p}$ by a constant: $\hat{u} = n\hat{p}$, where $n = \sum u_i$.

- Recall that $I_A$ is the toric ideal $\langle p^z - p^{z'} : z, z' \in \mathbb{N}^k,\ z - z' \in \ker_{\mathbb{Z}} A\rangle$, and $\mathcal{M}_A := \{p \in \Delta : \log p \in \text{rowspan} A\}$. Recall that the Zariski closure of $\mathcal{M}_A$ is $\mathcal{V}(I_A)$.

## Proof (1/2).

- Let $b_1, \ldots, b_l$ be a basis for $\ker_{\mathbb{Z}} A$. Observe that $p \in \mathcal{M}_A$ if and only if $\log p = A^T x$ and $p \in \Delta$ if and only if $b_j^T \log p = 0$ and $\sum p_i = 1$. (For example, $\log p = A^T x$ implies $b_j^T \log p = (Ab_j)^T x = 0$.)
- Let $v$ be the non-negative estimated table of counts. The expression $u^T \log v$ is up to a constant equal to $\log \prod p_i^{u_i}$, if we let $p_i = v_i/n$.
- We wish to maximize is $u^T \log v$, subject to $b_j^T \log v = 0$ for all $j = 1, \ldots, l$ and $\sum v_i = n$.

# Birch's Theorem – Part 1

## Proposition (2.1.5, Birch's Theorem)

Let $A \in \mathbb{N}^{d\times k}$ and $u \in \mathbb{N}^k$. The ML-estimate of the frequencies $\hat{u}$ in $\mathcal{M}_A$ is the unique non-negative solution to $A\hat{u} = Au$ for $\hat{u} \in \mathcal{V}(I_A)$.

- The MLE $\hat{u}$ differs from the MLE $\hat{p}$ by a constant: $\hat{u} = n\hat{p}$, where $n = \sum u_i$.

- Recall that $I_A$ is the toric ideal $\langle p^z - p^{z'} : z, z' \in \mathbb{N}^k,\ z - z' \in \ker_{\mathbb{Z}} A\rangle$, and $\mathcal{M}_A := \{p \in \Delta : \log p \in \mathrm{rowspan} A\}$. Recall that the Zariski closure of $\mathcal{M}_A$ is $\mathcal{V}(I_A)$.

## Proof (1/2).

• Let $b_1, \ldots, b_l$ be a basis for $\ker_{\mathbb{Z}} A$. Observe that $p \in \mathcal{M}_A$ if and only if $\log p = A^T x$ and $p \in \Delta$ if and only if $b_j^T \log p = 0$ and $\sum p_i = 1$. (For example, $\log p = A^T x$ implies $b_j^T \log p = (Ab_j)^T x = 0$.)

• Let $v$ be the non-negative estimated table of counts. The expression $u^T \log v$ is up to a constant equal to $\log \prod p_i^{u_i}$, if we let $p_i = v_i/n$.

• We wish to maximize is $u^T \log v$, subject to $b_j^T \log v = 0$ for all $j = 1, \ldots, l$ and $\sum v_i = n$.

# Birch's Theorem – Part 1

## Proposition (2.1.5, Birch's Theorem)

Let $A \in \mathbb{N}^{d\times k}$ and $u \in \mathbb{N}^k$. The ML-estimate of the frequencies $\hat{u}$ in $\mathcal{M}_A$ is the unique non-negative solution to $A\hat{u} = Au$ for $\hat{u} \in \mathcal{V}(I_A)$.

- The MLE $\hat{u}$ differs from the MLE $\hat{p}$ by a constant: $\hat{u} = n\hat{p}$, where $n = \sum u_i$.
- Recall that $I_A$ is the toric ideal $\langle p^z - p^{z'} : z, z' \in \mathbb{N}^k,\ z - z' \in \ker_{\mathbb{Z}} A\rangle$, and $\mathcal{M}_A := \{p \in \Delta : \log p \in \text{rowspan} A\}$. Recall that the Zariski closure of $\mathcal{M}_A$ is $\mathcal{V}(I_A)$.

## Proof (1/2).

- Let $b_1, \dots, b_l$ be a basis for $\ker_{\mathbb{Z}} A$. Observe that $p \in \mathcal{M}_A$ if and only if $\log p = A^T x$ and $p \in \Delta$ if and only if $b_j^T \log p = 0$ and $\sum p_i = 1$. (For example, $\log p = A^T x$ implies $b_j^T \log p = (Ab_j)^T x = 0$.)
- Let $v$ be the non-negative estimated table of counts. The expression $u^T \log v$ is up to a constant equal to $\log \prod p_i^{u_i}$, if we let $p_i = v_i/n$.
- We wish to maximize is $u^T \log v$, subject to $b_j^T \log v = 0$ for all $j = 1, \dots, l$ and $\sum v_i = n$.

# Birch's Theorem – Part 1

## Proposition (2.1.5, Birch's Theorem)

Let $A \in \mathbb{N}^{d\times k}$ and $u \in \mathbb{N}^k$. The ML-estimate of the frequencies $\hat{u}$ in $\mathcal{M}_A$ is the unique non-negative solution to $A\hat{u} = Au$ for $\hat{u} \in \mathcal{V}(I_A)$.

- The MLE $\hat{u}$ differs from the MLE $\hat{p}$ by a constant: $\hat{u} = n\hat{p}$, where $n = \sum u_i$.

- Recall that $I_A$ is the toric ideal $\langle p^z - p^{z'} : z, z' \in \mathbb{N}^k,\ z - z' \in \ker_{\mathbb{Z}} A\rangle$, and $\mathcal{M}_A := \{p \in \Delta : \log p \in \operatorname{rowspan} A\}$. Recall that the Zariski closure of $\mathcal{M}_A$ is $\mathcal{V}(I_A)$.

## Proof (1/2).

- Let $b_1, \ldots, b_l$ be a basis for $\ker_{\mathbb{Z}} A$. Observe that $p \in \mathcal{M}_A$ if and only if $\log p = A^T x$ and $p \in \Delta$ if and only if $b_j^T \log p = 0$ and $\sum p_i = 1$. (For example, $\log p = A^T x$ implies $b_j^T \log p = (Ab_j)^T x = 0$.)
- Let $v$ be the non-negative estimated table of counts. The expression $u^T \log v$ is up to a constant equal to $\log \prod p_i^{u_i}$, if we let $p_i = v_i/n$.
- We wish to maximize is $u^T \log v$, subject to $b_j^T \log v = 0$ for all $j = 1, \ldots, l$ and $\sum v_i = n$.

# Birch's Theorem – Part 2

**Proof (2/2).**

- The first constraint, $b_j^T \log v = 0$, is equivalent to $v \in \mathcal{V}(I_A)$. This is essentially because $v^z = v^{z'}$ if and only if $(z - z')^T \log v = 0$ (let $z = b_j^+$ and $z' = b_j^-$.)
- To solve the optimization problem we use the *method of Lagrange multipliers*. Write $\mathcal{L}(v, \lambda, \gamma) := u^T \log v - \sum \lambda_j b_j^T \log v - \gamma(n - \sum v_i)$.
- Putting the gradient of $\mathcal{L}$ to zero yields that the critical points are the solutions to the $k + l + 1$ equations

$$\frac{u_i}{v_i} + \sum \lambda_j \frac{b_{ij}}{v_i} + \gamma = 0, \ b_j^T \log v = 0, \ \sum v_i = n.$$

- The first conditions after clearing denominators can be written $u + \lambda B = -\gamma v$. We get $Au = (-\gamma)Av$. Since the column sum of $A$ are all equal, we get $\sum (Au)_i = a \sum u_i = -\gamma a \sum v_i$, implying $\gamma = -1$ and $Au = Av$.
- Uniqueness is due to the strict convexity of the likelihood function. $\mathcal{O}$

# Birch's Theorem – Part 2

## Proof (2/2).

- The first constraint, $b_j^T \log v = 0$, is equivalent to $v \in \mathcal{V}(I_A)$. This is essentially because $v^z = v^{z'}$ if and only if $(z - z')^T \log v = 0$ (let $z = b_j^+$ and $z' = b_j^-$.)
- To solve the optimization problem we use the *method of Lagrange multipliers*. Write $\mathcal{L}(v, \lambda, \gamma) := u^T \log v - \sum \lambda_j b_j^T \log v - \gamma(n - \sum v_i)$.
- Putting the gradient of $\mathcal{L}$ to zero yields that the critical points are the solutions to the $k + l + 1$ equations

$$\frac{u_i}{v_i} + \sum \lambda_j \frac{b_{ij}}{v_i} + \gamma = 0, \;\; b_j^T \log v = 0, \;\; \sum v_i = n.$$

- The first conditions after clearing denominators can be written $u + \lambda B = -\gamma v$. We get $Au = (-\gamma)Av$. Since the column sum of $A$ are all equal, we get $\sum (Au)_i = a \sum u_i = -\gamma a \sum v_i$, implying $\gamma = -1$ and $Au = Av$.
- Uniqueness is due to the strict convexity of the likelihood function. $\square$

# Birch's Theorem – Part 2

## Proof (2/2).

- The first constraint, $b_j^T \log v = 0$, is equivalent to $v \in \mathcal{V}(I_A)$. This is essentially because $v^z = v^{z'}$ if and only if $(z - z')^T \log v = 0$ (let $z = b_j^+$ and $z' = b_j^-$.)
- To solve the optimization problem we use the *method of Lagrange multipliers*. Write $\mathcal{L}(v, \lambda, \gamma) := u^T \log v - \sum \lambda_j b_j^T \log v - \gamma(n - \sum v_i)$.
- Putting the gradient of $\mathcal{L}$ to zero yields that the critical points are the solutions to the $k + l + 1$ equations

$$\frac{u_i}{v_i} + \sum \lambda_j \frac{b_{ij}}{v_i} + \gamma = 0, \; b_j^T \log v = 0, \; \sum v_i = n.$$

- The first conditions after clearing denominators can be written $u + \lambda B = -\gamma v$. We get $Au = (-\gamma)Av$. Since the column sum of $A$ are all equal, we get $\sum (Au)_i = a \sum u_i = -\gamma a \sum v_i$, implying $\gamma = -1$ and $Au = Av$.
- Uniqueness is due to the strict convexity of the likelihood function. $\square$

# Birch's Theorem – Part 2

## Proof (2/2).

- The first constraint, $b_j^T \log v = 0$, is equivalent to $v \in \mathcal{V}(I_A)$. This is essentially because $v^z = v^{z'}$ if and only if $(z - z')^T \log v = 0$ (let $z = b_j^+$ and $z' = b_j^-$.)
- To solve the optimization problem we use the *method of Lagrange multipliers*. Write $\mathcal{L}(v, \lambda, \gamma) := u^T \log v - \sum \lambda_j b_j^T \log v - \gamma(n - \sum v_i)$.
- Putting the gradient of $\mathcal{L}$ to zero yields that the critical points are the solutions to the $k + l + 1$ equations

$$\frac{u_i}{v_i} + \sum \lambda_j \frac{b_{ij}}{v_i} + \gamma = 0, \; b_j^T \log v = 0, \; \sum v_i = n.$$

- The first conditions after clearing denominators can be written $u + \lambda B = -\gamma v$. We get $Au = (-\gamma)Av$. Since the column sum of $A$ are all equal, we get $\sum (Au)_i = a \sum u_i = -\gamma a \sum v_i$, implying $\gamma = -1$ and $Au = Av$.
- Uniqueness is due to the strict convexity of the likelihood function.

# Birch's Theorem – Part 2

## Proof (2/2).

- The first constraint, $b_j^T \log v = 0$, is equivalent to $v \in \mathcal{V}(I_A)$. This is essentially because $v^z = v^{z'}$ if and only if $(z - z')^T \log v = 0$ (let $z = b_j^+$ and $z' = b_j^-$.)
- To solve the optimization problem we use the *method of Lagrange multipliers*. Write $\mathcal{L}(v, \lambda, \gamma) := u^T \log v - \sum \lambda_j b_j^T \log v - \gamma(n - \sum v_i)$.
- Putting the gradient of $\mathcal{L}$ to zero yields that the critical points are the solutions to the $k + l + 1$ equations

$$\frac{u_i}{v_i} + \sum \lambda_j \frac{b_{ij}}{v_i} + \gamma = 0, \; b_j^T \log v = 0, \; \sum v_i = n.$$

- The first conditions after clearing denominators can be written $u + \lambda B = -\gamma v$. We get $Au = (-\gamma)Av$. Since the column sum of $A$ are all equal, we get $\sum (Au)_i = a \sum u_i = -\gamma a \sum v_i$, implying $\gamma = -1$ and $Au = Av$.
- Uniqueness is due to the strict convexity of the likelihood function. $\mathcal{O}$

# Junction Trees – Part 1

- Let $\Gamma$ be a decomposable simplicial complex. A *junction tree* is a tree whose vertices are the facets of $\Gamma$, whose edges are labeled by separators in $\Gamma$, and such that each edge splits the set of facets of $\Gamma$ into two subcomplexes $\Gamma_1, \Gamma_2$ in $(\Gamma_1, S, \Gamma_2)$.

- If $\Gamma = [123][134]$, then $[123] - [134]$ represents the unique junction tree. For $\Gamma = [12][13][14]$, there are a few different junction trees, for example $[12] - [13] - [14]$. A junction tree can be obtained by breaking down a decomposable complex down to its constituent simplices.

# Junction Trees – Part 1

• Let $\Gamma$ be a decomposable simplicial complex. A *junction tree* is a tree whose vertices are the facets of $\Gamma$, whose edges are labeled by separators in $\Gamma$, and such that each edge splits the set of facets of $\Gamma$ into two subcomplexes $\Gamma_1, \Gamma_2$ in $(\Gamma_1, S, \Gamma_2)$.

• If $\Gamma = [123][134]$, then $[123] - [134]$ represents the unique junction tree. For $\Gamma = [12][13][14]$, there are a few different junction trees, for example $[12] - [13] - [14]$. A junction tree can be obtained by breaking down a decomposable complex down to its constituent simplices.

# Junction Trees – Part 2

- A *clique* in a graph $G$ is a subgraph that is complete, meaning each vertex is connected to all other vertices by an edge in this subgraph.

**Proposition (2.1.7)**

Let $\Gamma$ be a decomposable simplicial complex. Let $u$ be data such that all marginals along cliques are positive. Let $J(\Gamma)$ be a junction tree for $\Gamma$. Then the maximum likelihood estimates of the table of frequencies is given by

$$\hat{u}_i = \frac{\prod_{F \in V(J(\Gamma))} (u|_F)_{i_F}}{\prod_{S \in E(J(\Gamma))} (u|_S)_{i_S}}$$

In particular, decomposable models have ML degree one.

- The condition on the cliques makes sure that the denominator is non-zero.
- Recall the underlying hierarchical log-linear model

$$\mathcal{M}_\Gamma = \left\{ p \in \Delta : p_i = \frac{1}{Z(\theta)} \prod_F \theta_{i_F}^{(F)} \right\}.$$

# Junction Trees – Part 2

- A *clique* in a graph $G$ is a subgraph that is complete, meaning each vertex is connected to all other vertices by an edge in this subgraph.

## Proposition (2.1.7)

Let $\Gamma$ be a decomposable simplicial complex. Let $u$ be data such that all marginals along cliques are positive. Let $J(\Gamma)$ be a junction tree for $\Gamma$. Then the maximum likelihood estimates of the table of frequencies is given by

$$\hat{u}_i = \frac{\prod_{F \in V(J(\Gamma))} (u|_F)_{i_F}}{\prod_{S \in E(J(\Gamma))} (u|_S)_{i_S}}$$

In particular, decomposable models have ML degree one.

- The condition on the cliques makes sure that the denominator is non-zero.
- Recall the underlying hierarchical log-linear model

$$\mathcal{M}_\Gamma = \left\{ p \in \Delta : p_i = \frac{1}{Z(\theta)} \prod_F \theta_{i_F}^{(F)} \right\}.$$

# Junction Trees – Part 2

- A *clique* in a graph $G$ is a subgraph that is complete, meaning each vertex is connected to all other vertices by an edge in this subgraph.

## Proposition (2.1.7)

Let $\Gamma$ be a decomposable simplicial complex. Let $u$ be data such that all marginals along cliques are positive. Let $J(\Gamma)$ be a junction tree for $\Gamma$. Then the maximum likelihood estimates of the table of frequencies is given by

$$\hat{u}_i = \frac{\prod_{F \in V(J(\Gamma))} (u|_F)_{i_F}}{\prod_{S \in E(J(\Gamma))} (u|_S)_{i_S}}$$

In particular, decomposable models have ML degree one.

- The condition on the cliques makes sure that the denominator is non-zero.
- Recall the underlying hierarchical log-linear model

$$\mathcal{M}_\Gamma = \left\{ p \in \Delta : p_i = \frac{1}{Z(\theta)} \prod_F \theta_{i_F}^{(F)} \right\}.$$

# Junction Trees – Part 2

- A *clique* in a graph $G$ is a subgraph that is complete, meaning each vertex is connected to all other vertices by an edge in this subgraph.

## Proposition (2.1.7)

Let $\Gamma$ be a decomposable simplicial complex. Let $u$ be data such that all marginals along cliques are positive. Let $J(\Gamma)$ be a junction tree for $\Gamma$. Then the maximum likelihood estimates of the table of frequencies is given by

$$\hat{u}_i = \frac{\prod_{F \in V(J(\Gamma))} (u|_F)_{i_F}}{\prod_{S \in E(J(\Gamma))} (u|_S)_{i_S}}$$

In particular, decomposable models have ML degree one.

- The condition on the cliques makes sure that the denominator is non-zero.
- Recall the underlying hierarchical log-linear model

$$\mathcal{M}_\Gamma = \left\{ p \in \Delta : p_i = \frac{1}{Z(\theta)} \prod_F \theta_{i_F}^{(F)} \right\}.$$

# Iterative Proportional Scaling

- There is no closed-form formula for maximum likelihood estimates for non-decomposable log-linear models. However, the log-likelihood function is convex for these models and therefore computer algorithms are appropriate for computing ML estimates.

- A popular choice for an algorithm is the **Iterative Proportional Scaling Algorithm** (*Lecture Notes on Algebraic Statistics*, page 43.) It inputs $A \in \mathbb{N}^{d\times k}$, a table of counts $u \in \mathbb{N}^k$ and a tolerance $\epsilon > 0$, and outputs expected counts $\hat{u}$.

# Iterative Proportional Scaling

- There is no closed-form formula for maximum likelihood estimates for non-decomposable log-linear models. However, the log-likelihood function is convex for these models and therefore computer algorithms are appropriate for computing ML estimates.

- A popular choice for an algorithm is the **Iterative Proportional Scaling Algorithm** (*Lecture Notes on Algebraic Statistics*, page 43.) It inputs $A \in \mathbb{N}^{d \times k}$, a table of counts $u \in \mathbb{N}^k$ and a tolerance $\epsilon > 0$, and outputs expected counts $\hat{u}$.

# Gaussian Models

- A *Gaussian model* is described by $\mathcal{P}_\Theta = \{\mathcal{N}(\mu,\Sigma) : \theta = (\mu,\Sigma) \in \Theta\}$, where $\Theta \subseteq \mathbb{R}^m \times \mathrm{PD}_m$ ($\mathrm{PD}_m$ is the cone of symmetric positive definite matrices.) We write $X \sim \mathcal{N}(\mu,\Sigma)$ for an $m$-dimensional random vector $X$ if it has the density function

$$f_{\mu,\Sigma}(x) = \frac{1}{(2\pi)^{m/2}(\det\Sigma)^{1/2}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)},$$

where $x, \mu \in \mathbb{R}^m$ and $\Sigma$ is a symmetric positive definite matrix. We call $\mu$ the mean and $\Sigma$ is the covariance matrix.

- The log-likelihood function is up to a constant equal to

$$\ell(\mu,\Sigma) = -\frac{n}{2}\log\det\Sigma - \frac{1}{2}\sum(X^{(i)}-\mu)^T\Sigma^{-1}(X^{(i)}-\mu).$$

It is a simple exercise to show that $v^TAv = \mathrm{trace}(Av^Tv)$, implying

$$\ell(\mu,\Sigma) = -\frac{n}{2}\log\det\Sigma - \frac{1}{2}\,\mathrm{trace}\left(\Sigma^{-1}\sum(X^{(i)}-\mu)(X^{(i)}-\mu)^T\right).$$

# Gaussian Models

- A *Gaussian model* is described by $\mathcal{P}_\Theta = \{\mathcal{N}(\mu, \Sigma) : \theta = (\mu, \Sigma) \in \Theta\}$, where $\Theta \subseteq \mathbb{R}^m \times \mathrm{PD}_m$ ($\mathrm{PD}_m$ is the cone of symmetric positive definite matrices.) We write $X \sim \mathcal{N}(\mu, \Sigma)$ for an $m$-dimensional random vector $X$ if it has the density function

$$f_{\mu,\Sigma}(x) = \frac{1}{(2\pi)^{m/2}(\det \Sigma)^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)},$$

where $x, \mu \in \mathbb{R}^m$ and $\Sigma$ is a symmetric positive definite matrix. We call $\mu$ the mean and $\Sigma$ is the covariance matrix.

- The log-likelihood function is up to a constant equal to

$$\ell(\mu, \Sigma) = -\frac{n}{2} \log\det\Sigma - \frac{1}{2} \sum (X^{(i)} - \mu)^T \Sigma^{-1} (X^{(i)} - \mu).$$

It is a simple exercise to show that $v^T A v = \mathrm{trace}(A v^T v)$, implying

$$\ell(\mu, \Sigma) = -\frac{n}{2} \log\det\Sigma - \frac{1}{2} \mathrm{trace}\left(\Sigma^{-1} \sum (X^{(i)} - \mu)(X^{(i)} - \mu)^T\right).$$

# The Saturated Gaussian Model

- The *saturated* Gaussian model is described by $\Theta = \mathbb{R}^m \times \mathrm{PD}_m$. For this model, we have the ML-estimates

$$\hat{\mu} = \overline{X} = \frac{1}{n}\sum X^{(i)},\ \hat{\Sigma} = S = \frac{1}{n}\sum (X^{(i)} - \mu)(X^{(i)} - \mu)^T.$$

We call $\overline{X}$ the *sample mean* and $S$ the *sample covariance.*

- Let us deduce the formula for $\hat{\mu}$. Observe that

$$\sum (X^{(i)} - \mu)^T \Sigma^{-1} (X^{(i)} - \mu) = \sum X^{(i)T} \Sigma^{-1} X^{(i)} - 2\mu^T \Sigma^{-1} X^{(i)} + \mu^T \Sigma^{-1} \mu.$$

Using the formula

$$x^T A y = \sum_{ij} A_{ij} x_i y_j,$$

we get the score equations

$$0 = \frac{\partial \ell(\mu, \Sigma)}{\partial \mu_k} = -\frac{1}{2} \sum_i \left( 2 \sum_j \Sigma^{-1}_{kj} \mu_j - 2(\Sigma^{-1} X^{(i)})_k \right) \Rightarrow \sum_i \left( \Sigma^{-1}\mu - \Sigma^{-1} X^i \right) = 0$$

# The Saturated Gaussian Model

• The *saturated* Gaussian model is described by $\Theta = \mathbb{R}^m \times \mathrm{PD}_m$. For this model, we have the ML-estimates

$$\hat{\mu} = \overline{X} = \frac{1}{n}\sum X^{(i)},\ \hat{\Sigma} = S = \frac{1}{n}\sum (X^{(i)} - \mu)(X^{(i)} - \mu)^T.$$

We call $\overline{X}$ the *sample mean* and $S$ the *sample covariance.*

• Let us deduce the formula for $\hat{\mu}$. Observe that

$$\sum (X^{(i)} - \mu)^T \Sigma^{-1} (X^{(i)} - \mu) = \sum X^{(i)T} \Sigma^{-1} X^{(i)} - 2\mu^T \Sigma^{-1} X^{(i)} + \mu^T \Sigma^{-1} \mu.$$

Using the formula

$$x^T A y = \sum_{ij} A_{ij} x_i y_j,$$

we get the score equations

$$0 = \frac{\partial \ell(\mu, \Sigma)}{\partial \mu_k} = -\frac{1}{2}\sum_i \left(2\sum_j \Sigma^{-1}_{kj}\mu_j - 2(\Sigma^{-1}X^{(i)})_k\right) \Rightarrow \sum_i \left(\Sigma^{-1}\mu - \Sigma^{-1}X^i\right) = 0$$

# Special Cases – Part 1

- Using $S$, we can rewrite our log-likelihood function

$$\ell(\mu, \Sigma) = -\frac{n}{2} \log \det \Sigma - \frac{n}{2} \operatorname{trace}(S\Sigma^{-1}) - \frac{n}{2}(\overline{X} - \mu)^T (\overline{X} - \mu).$$

## Proposition (2.1.10)

Suppose that $\Theta = \Theta_1 \times \{\mathrm{I}_m\}$ is the parameter space of a Gaussian model. The MLE $\hat{\mu}$ of the mean is the point in $\Theta_1 \subseteq \mathbb{R}^m$ that is the closest to $\overline{X}$ in the $L^2$-norm.

## Proof.

- When $\Sigma$ is the identity matrix $\mathrm{I}_m$, the log-likelihood function reduces to

$$\ell(\mu, \mathrm{I}_m) = -\frac{n}{2} \operatorname{trace} S - \frac{n}{2}(\overline{X} - \mu)^T (\overline{X} - \mu) = -\frac{n}{2} \operatorname{trace} S - \frac{n}{2} \|\overline{X} - \mu\|_2^2.$$

- Therefore, maximizing $\ell$ over $\Theta_1$ is equivalent to minimizing $\|\overline{X} - \mu\|_2$ over $\Theta_1$. $\mathcal{O}$

# Special Cases – Part 1

- Using $S$, we can rewrite our log-likelihood function

$$\ell(\mu, \Sigma) = -\frac{n}{2} \log \det \Sigma - \frac{n}{2} \operatorname{trace}(S\Sigma^{-1}) - \frac{n}{2}(\overline{X} - \mu)^T(\overline{X} - \mu).$$

## Proposition (2.1.10)

Suppose that $\Theta = \Theta_1 \times \{\mathrm{I}_m\}$ is the parameter space of a Gaussian model. The MLE $\hat{\mu}$ of the mean is the point in $\Theta_1 \subseteq \mathbb{R}^m$ that is the closest to $\overline{X}$ in the $L^2$-norm.

## Proof.

- When $\Sigma$ is the identity matrix $\mathrm{I}_m$, the log-likelihood function reduces to

$$\ell(\mu, \mathrm{I}_m) = -\frac{n}{2} \operatorname{trace} S - \frac{n}{2}(\overline{X} - \mu)^T(\overline{X} - \mu) = -\frac{n}{2} \operatorname{trace} S - \frac{n}{2}\|\overline{X} - \mu\|_2^2.$$

- Therefore, maximizing $\ell$ over $\Theta_1$ is equivalent to minimizing $\|\overline{X} - \mu\|_2$ over $\Theta_1$.

$\mathcal{O}$

# Special Cases – Part 1

- Using $S$, we can rewrite our log-likelihood function

$$\ell(\mu,\Sigma) = -\frac{n}{2}\log\det\Sigma - \frac{n}{2}\operatorname{trace}(S\Sigma^{-1}) - \frac{n}{2}(\overline{X}-\mu)^T(\overline{X}-\mu).$$

## Proposition (2.1.10)

Suppose that $\Theta = \Theta_1 \times \{\mathrm{I}_m\}$ is the parameter space of a Gaussian model. The MLE $\hat{\mu}$ of the mean is the point in $\Theta_1 \subseteq \mathbb{R}^m$ that is the closest to $\overline{X}$ in the $L^2$-norm.

## Proof.

- When $\Sigma$ is the identity matrix $\mathrm{I}_m$, the log-likelihood function reduces to

$$\ell(\mu,\mathrm{I}_m) = -\frac{n}{2}\operatorname{trace} S - \frac{n}{2}(\overline{X}-\mu)^T(\overline{X}-\mu) = -\frac{n}{2}\operatorname{trace} S - \frac{n}{2}\|\overline{X}-\mu\|_2^2.$$

- Therefore, maximizing $\ell$ over $\Theta_1$ is equivalent to minimizing $\|\overline{X}-\mu\|_2$ over $\Theta_1$.

# Special Cases – Part 1

- Using $S$, we can rewrite our log-likelihood function

$$\ell(\mu, \Sigma) = -\frac{n}{2}\log\det\Sigma - \frac{n}{2}\operatorname{trace}(S\Sigma^{-1}) - \frac{n}{2}(\overline{X}-\mu)^T(\overline{X}-\mu).$$

## Proposition (2.1.10)

Suppose that $\Theta = \Theta_1 \times \{\mathrm{I}_m\}$ is the parameter space of a Gaussian model. The MLE $\hat{\mu}$ of the mean is the point in $\Theta_1 \subseteq \mathbb{R}^m$ that is the closest to $\overline{X}$ in the $L^2$-norm.

## Proof.

- When $\Sigma$ is the identity matrix $\mathrm{I}_m$, the log-likelihood function reduces to

$$\ell(\mu, \mathrm{I}_m) = -\frac{n}{2}\operatorname{trace} S - \frac{n}{2}(\overline{X}-\mu)^T(\overline{X}-\mu) = -\frac{n}{2}\operatorname{trace} S - \frac{n}{2}\|\overline{X}-\mu\|_2^2.$$

- Therefore, maximizing $\ell$ over $\Theta_1$ is equivalent to minimizing $\|\overline{X}-\mu\|_2$ over $\Theta_1$. $\mathcal{O}$

# Special Cases – Part 2

**Proposition (2.1.12)**

Suppose that $\Theta = \mathbb{R}^m \times \Theta_2$. Then $\hat{\mu} = \overline{X}$ and $\hat{\Sigma}$ is the maximizer of

$$\ell(\Sigma) = -\frac{n}{2} \log \det \Sigma - \frac{n}{2} \operatorname{trace} S\Sigma^{-1}$$

in the set $\Theta_2$.

**Proof.**

- The inverse of $\Sigma$ is also positive definite (recall that a matrix is positive definite if and only if all its eigenvalues are positive.) Therefore $(\overline{X} - \mu)^T \Sigma^{-1} (\overline{X} - \mu) \geq 0$ and equality holds if and only if $\mu = \overline{X}$.

# Special Cases – Part 2

## Proposition (2.1.12)

Suppose that $\Theta = \mathbb{R}^m \times \Theta_2$. Then $\hat{\mu} = \overline{X}$ and $\hat{\Sigma}$ is the maximizer of

$$\ell(\Sigma) = -\frac{n}{2} \log \det \Sigma - \frac{n}{2} \operatorname{trace} S\Sigma^{-1}$$

in the set $\Theta_2$.

## Proof.

- The inverse of $\Sigma$ is also positive definite (recall that a matrix is positive definite if and only if all its eigenvalues are positive.) Therefore $(\overline{X} - \mu)^T \Sigma^{-1} (\overline{X} - \mu) \geq 0$ and equality holds if and only if $\mu = \overline{X}$.

$\mathcal{O}$

# Special Cases – Part 3

**Theorem (2.1.14)**

*Let $G = (V, E)$ be an undirected graph and $\Theta = \mathbb{R}^m \times \Theta_2$, where*

$$\Theta_2 = \{\Sigma \in \mathrm{PD}_m : (\Sigma^{-1})_{ij} = 0 \text{ if } ij \notin E\}.$$

*The ML-estimate of $\Sigma$ given a positive definite sample covariance matrix $S$ is, the unique positive definite matrix $\hat{\Sigma}$ such that $\hat{\Sigma}_{ij} = S_{ij}, ij \in E$ and $(\hat{\Sigma}^{-1})_{ij} = 0$ for $ij \notin E$.*

- Models of this form as by definition called *Gaussian graphical models*.
- The inverse $K = \Sigma^{-1}$ of the covariance matrix is known as the *concentration* matrix and in some cases it is more convienient to parametrize the model via concentration matrices.
- Observing $\log\det K = -\log\det\Sigma$ allows us to instead consider the log-likelihood function (up to scaling)

$$K \mapsto \log\det K - \mathrm{trace}(SK).$$

# Special Cases – Part 3

**Theorem (2.1.14)**

*Let* $G = (V, E)$ *be an undirected graph and* $\Theta = \mathbb{R}^m \times \Theta_2$*, where*

$$\Theta_2 = \{\Sigma \in \mathrm{PD}_m : (\Sigma^{-1})_{ij} = 0 \text{ if } ij \notin E\}.$$

*The ML-estimate of* $\Sigma$ *given a positive definite sample covariance matrix* $S$ *is, the unique positive definite matrix* $\hat{\Sigma}$ *such that* $\hat{\Sigma}_{ij} = S_{ij}$*,* $ij \in E$ *and* $(\hat{\Sigma}^{-1})_{ij} = 0$ *for* $ij \notin E$*.*

- Models of this form as by definition called *Gaussian graphical models.*
- The inverse $K = \Sigma^{-1}$ of the covariance matrix is known as the *concentration* matrix and in some cases it is more convienient to parametrize the model via concentration matrices.
- Observing $\log\det K = -\log\det\Sigma$ allows us to instead consider the log-likelihood function (up to scaling)

$$K \mapsto \log\det K - \operatorname{trace}(SK).$$

# Special Cases – Part 3

**Theorem (2.1.14)**

*Let $G = (V, E)$ be an undirected graph and $\Theta = \mathbb{R}^m \times \Theta_2$, where*

$$\Theta_2 = \{\Sigma \in \mathrm{PD}_m : (\Sigma^{-1})_{ij} = 0 \text{ if } ij \notin E\}.$$

*The ML-estimate of $\Sigma$ given a positive definite sample covariance matrix $S$ is, the unique positive definite matrix $\hat{\Sigma}$ such that $\hat{\Sigma}_{ij} = S_{ij}$, $ij \in E$ and $(\hat{\Sigma}^{-1})_{ij} = 0$ for $ij \notin E$.*

- Models of this form as by definition called *Gaussian graphical models.*
- The inverse $K = \Sigma^{-1}$ of the covariance matrix is known as the *concentration* matrix and in some cases it is more convienient to parametrize the model via concentration matrices.
- Observing $\log\det K = -\log\det\Sigma$ allows us to instead consider the log-likelihood function (up to scaling)

$$K \mapsto \log\det K - \mathrm{trace}(SK).$$

# Special Cases – Part 3

**Theorem (2.1.14)**

*Let* $G = (V, E)$ *be an undirected graph and* $\Theta = \mathbb{R}^m \times \Theta_2$*, where*

$$\Theta_2 = \{\Sigma \in \mathrm{PD}_m : (\Sigma^{-1})_{ij} = 0 \text{ if } ij \notin E\}.$$

*The ML-estimate of* $\Sigma$ *given a positive definite sample covariance matrix* $S$ *is, the unique positive definite matrix* $\hat{\Sigma}$ *such that* $\hat{\Sigma}_{ij} = S_{ij}, ij \in E$ *and* $(\hat{\Sigma}^{-1})_{ij} = 0$ *for* $ij \notin E$.

- Models of this form as by definition called *Gaussian graphical models.*
- The inverse $K = \Sigma^{-1}$ of the covariance matrix is known as the *concentration* matrix and in some cases it is more convienient to parametrize the model via concentration matrices.
- Observing $\log\det K = -\log\det\Sigma$ allows us to instead consider the log-likelihood function (up to scaling)

$$K \mapsto \log\det K - \mathrm{trace}(SK).$$

# Linear Spaces of Symmetric Matrices – Part 1

- An LSSM is a linear space (in particular a variety) $\mathcal{L} \subseteq \mathbb{S}^n$, where $\mathbb{S}^n$ is the set of $n$-dimensional symmetric matrices. We assume it contains at least one invertible matrix. Let $\bullet$ denote the trace operator on matrices, $A \bullet B := \mathrm{trace}(AB)$.
- The log-likelihood function for the concentration matrix is

$$\ell(K) = \log\det K - \mathrm{trace}\, SK.$$

Let $A_1, \ldots, A_m$ be matrices that span $\mathcal{L}$, so that $\mathcal{L} = \{\sum \lambda_i A_i : \lambda_i \in \mathbb{C}^n\}$. The score equations are

$$(\ell(M))'_{A_i} = \nabla \ell(M) \bullet A_i = (M^{-1} - S) \bullet A_i = 0.$$

- We define $\mathcal{L}^{-1} := \overline{\{M^{-1} : M \in \mathcal{L} \cap \mathrm{GL}(\mathbb{S}^n)\}}^{\mathrm{Zar}}$ and $\mathcal{L}^{\perp} := \{M \in \mathbb{S}^n : \mathrm{trace}(M\mathcal{L}) = 0\}$.
- If we write $\mathrm{mld}(\mathcal{L})$ for the ML-degree, then for a generic $S$

$$\mathrm{mld}(\mathcal{L}) = \#\big((\mathcal{L}^{-1} - S) \cap \mathcal{L}^{\perp}\big)$$

## Linear Spaces of Symmetric Matrices – Part 1

- An LSSM is a linear space (in particular a variety) $\mathcal{L} \subseteq \mathbb{S}^n$, where $\mathbb{S}^n$ is the set of $n$-dimensional symmetric matrices. We assume it contains at least one invertible matrix. Let $\bullet$ denote the trace operator on matrices, $A \bullet B := \mathrm{trace}(AB)$.
- The log-likelihood function for the concentration matrix is

$$\ell(K) = \log\det K - \mathrm{trace}\, SK.$$

Let $A_1, \ldots, A_m$ be matrices that span $\mathcal{L}$, so that $\mathcal{L} = \{\sum \lambda_i A_i : \lambda_i \in \mathbb{C}^n\}$. The score equations are

$$(\ell(M))'_{A_i} = \nabla \ell(M) \bullet A_i = (M^{-1} - S) \bullet A_i = 0.$$

- We define $\mathcal{L}^{-1} := \overline{\{M^{-1} : M \in \mathcal{L} \cap \mathrm{GL}(\mathbb{S}^n)\}}^{\mathrm{Zar}}$ and $\mathcal{L}^{\perp} := \{M \in \mathbb{S}^n : \mathrm{trace}(M\mathcal{L}) = 0\}$.
- If we write $\mathrm{mld}(\mathcal{L})$ for the ML-degree, then for a generic $S$

$$\mathrm{mld}(\mathcal{L}) = \#\big((\mathcal{L}^{-1} - S) \cap \mathcal{L}^{\perp}\big)$$

# Linear Spaces of Symmetric Matrices – Part 1

- An LSSM is a linear space (in particular a variety) $\mathcal{L} \subseteq \mathbb{S}^n$, where $\mathbb{S}^n$ is the set of $n$-dimensional symmetric matrices. We assume it contains at least one invertible matrix. Let $\bullet$ denote the trace operator on matrices, $A \bullet B := \mathrm{trace}(AB)$.
- The log-likelihood function for the concentration matrix is

$$\ell(K) = \log\det K - \mathrm{trace}\, SK.$$

Let $A_1, \ldots, A_m$ be matrices that span $\mathcal{L}$, so that $\mathcal{L} = \{\sum \lambda_i A_i : \lambda_i \in \mathbb{C}^n\}$. The score equations are

$$(\ell(M))'_{A_i} = \nabla \ell(M) \bullet A_i = (M^{-1} - S) \bullet A_i = 0.$$

- We define $\mathcal{L}^{-1} := \overline{\{M^{-1} : M \in \mathcal{L} \cap \mathrm{GL}(\mathbb{S}^n)\}}^{\mathrm{Zar}}$ and $\mathcal{L}^{\perp} := \{M \in \mathbb{S}^n : \mathrm{trace}(M\mathcal{L}) = 0\}$.

- If we write $\mathrm{mld}(\mathcal{L})$ for the ML-degree, then for a generic $S$

$$\mathrm{mld}(\mathcal{L}) = \#\left((\mathcal{L}^{-1} - S) \cap \mathcal{L}^{\perp}\right)$$

# Linear Spaces of Symmetric Matrices – Part 1

- An LSSM is a linear space (in particular a variety) $\mathcal{L} \subseteq \mathbb{S}^n$, where $\mathbb{S}^n$ is the set of $n$-dimensional symmetric matrices. We assume it contains at least one invertible matrix. Let $\bullet$ denote the trace operator on matrices, $A \bullet B := \mathrm{trace}(AB)$.
- The log-likelihood function for the concentration matrix is

$$\ell(K) = \log\det K - \mathrm{trace}\, SK.$$

Let $A_1, \dots, A_m$ be matrices that span $\mathcal{L}$, so that $\mathcal{L} = \{\sum \lambda_i A_i : \lambda_i \in \mathbb{C}^n\}$. The score equations are

$$(\ell(M))'_{A_i} = \nabla \ell(M) \bullet A_i = (M^{-1} - S) \bullet A_i = 0.$$

- We define $\mathcal{L}^{-1} := \overline{\{M^{-1} : M \in \mathcal{L} \cap \mathrm{GL}(\mathbb{S}^n)\}}^{\mathrm{Zar}}$ and $\mathcal{L}^{\perp} := \{M \in \mathbb{S}^n : \mathrm{trace}(M\mathcal{L}) = 0\}$.
- If we write $\mathrm{mld}(\mathcal{L})$ for the ML-degree, then for a generic $S$

$$\mathrm{mld}(\mathcal{L}) = \#\big((\mathcal{L}^{-1} - S) \cap \mathcal{L}^{\perp}\big)$$

# Linear Spaces of Symmetric Matrices – Part 2

- Observe that the set $(\mathcal{L}^{-1} - S) \cap \mathcal{L}^{\perp}$ is a variety. We can use software to calculate the cardinality. Generically, there is no non-invertible matrix in this intersection.

- We simulate generic points by taking random rational data (and do this a few times to make sure that we did not get a non-generic point.)

- The *degree* of a variety is the *degree* of its defining radical ideal. The definition is a bit technical, but can be calculated with software.

**Proposition (Linear spaces of symmetric matrices with non-maximal maximum likelihood degree, Theorem 1.1)**

The ML-degree of a linear space $\mathcal{L} \subset \mathbb{S}^n$ is at most the degree of the variety $\mathbb{P}\mathcal{L}^{-1}$. This is an equality if and only if the intersection $\mathbb{P}\mathcal{L}^{-1} \cap \mathbb{P}\mathcal{L}^{\perp}$ is empty.

# Linear Spaces of Symmetric Matrices – Part 2

- Observe that the set $(\mathcal{L}^{-1} - S) \cap \mathcal{L}^{\perp}$ is a variety. We can use software to calculate the cardinality. Generically, there is no non-invertible matrix in this intersection.

- We simulate generic points by taking random rational data (and do this a few times to make sure that we did not get a non-generic point.)

- The *degree* of a variety is the *degree* of its defining radical ideal. The definition is a bit technical, but can be calculated with software.

**Proposition (Linear spaces of symmetric matrices with non-maximal maximum likelihood degree, Theorem 1.1)**

The ML-degree of a linear space $\mathcal{L} \subset \mathbb{S}^n$ is at most the degree of the variety $\mathbb{P}\mathcal{L}^{-1}$. This is an equality if and only if the intersection $\mathbb{P}\mathcal{L}^{-1} \cap \mathbb{P}\mathcal{L}^{\perp}$ is empty.

# Linear Spaces of Symmetric Matrices – Part 2

- Observe that the set $(\mathcal{L}^{-1} - S) \cap \mathcal{L}^{\perp}$ is a variety. We can use software to calculate the cardinality. Generically, there is no non-invertible matrix in this intersection.

- We simulate generic points by taking random rational data (and do this a few times to make sure that we did not get a non-generic point.)

- The *degree* of a variety is the *degree* of its defining radical ideal. The definition is a bit technical, but can be calculated with software.

**Proposition (Linear spaces of symmetric matrices with non-maximal maximum likelihood degree, Theorem 1.1)**

The ML-degree of a linear space $\mathcal{L} \subset \mathbb{S}^n$ is at most the degree of the variety $\mathbb{P}\mathcal{L}^{-1}$. This is an equality if and only if the intersection $\mathbb{P}\mathcal{L}^{-1} \cap \mathbb{P}\mathcal{L}^{\perp}$ is empty.

# Linear Spaces of Symmetric Matrices – Part 2

- Observe that the set $(\mathcal{L}^{-1} - S) \cap \mathcal{L}^{\perp}$ is a variety. We can use software to calculate the cardinality. Generically, there is no non-invertible matrix in this intersection.

- We simulate generic points by taking random rational data (and do this a few times to make sure that we did not get a non-generic point.)

- The *degree* of a variety is the *degree* of its defining radical ideal. The definition is a bit technical, but can be calculated with software.

**Proposition (Linear spaces of symmetric matrices with non-maximal maximum likelihood degree, Theorem 1.1)**

The ML-degree of a linear space $\mathcal{L} \subset \mathbb{S}^n$ is at most the degree of the variety $\mathbb{P}\mathcal{L}^{-1}$. This is an equality if and only if the intersection $\mathbb{P}\mathcal{L}^{-1} \cap \mathbb{P}\mathcal{L}^{\perp}$ is empty.

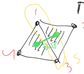
# Outro

- Thank you for listening!

Combinatorial and Algebraic Statistics

Presentation notes: Examples of junction trees.

unique choice



$\Gamma = [123][134]$

$\mathcal{J}(\Gamma) = [123] - [134]$

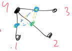
$([123], S, [134])$
$S = [13]$

unique choice


$\Gamma = [123][134][25]$

$\mathcal{J}(\Gamma) = [134] \underset{[13]}{-} [123] \underset{[2]}{-} [25]$

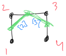
three choices



$\Gamma = [14][24][34]$

$\mathcal{J}(\Gamma) = [34] \underset{[4]}{-} [14] \underset{[4]}{-} [12]$

## Proposition 2.1.7 example:




$\Gamma = [12][23][34]$.

$\mathcal{J}(\Gamma) = [12] \underset{[2]}{-} [23] \underset{[3]}{-} [34]$

$$\hat{u}_{i_1 i_2 i_3 i_4} = \frac{\prod_{F \in V(\mathcal{J}(\Gamma))} (u|_F)_{i_F}}{\prod_{S \in E(\mathcal{J}(\Gamma))} (u|_S)_{i_S}} =$$

$$= \frac{(u|_{[12]})_{i_1 i_2} (u|_{[23]})_{i_2 i_3} (u|_{[34]})_{i_3 i_4}}{(u|_{[2]})_{i_2} (u|_{[3]})_{i_3}}$$

$$= \frac{u_{i_1 i_2 + +}\, u_{+ i_2 i_3 +}\, u_{+ + i_3 i_4}}{u_{+ i_2 + +}\, u_{+ + i_3 +}}$$