Lecture 6

Combinatorial and Algebraic Statistics

Nils Hemmingsson
2021-02-26

Structure

- First, which will be the bulk of the lecture, I will define a few key concepts and provide some statements, one of which I will show the proof of
- Second, I will talk a little bit about how those concepts can be used in statistics.

Preliminaries

- All sublattices $\mathcal{L}$ of $\mathbb{Z}^k$ we will consider are such that the only non-negative member is the origin, that is $\mathcal{L} \cap \mathbb{N}^k = \{0\}$. For instance, the kernel of non-negative integer matrices are such.

Preliminaries

- All sublattices $\mathcal{L}$ of $\mathbb{Z}^k$ we will consider are such that the only non-negative member is the origin, that is $\mathcal{L} \cap \mathbb{N}^k = \{0\}$. For instance, the kernel of non-negative integer matrices are such.
- We define the fiber of a lattice in the following way:

$$\mathcal{F}(u) := (u + \mathcal{L}) \cap \mathbb{N}^k = \{v \in \mathbb{N}^k : u - v \in \mathcal{L}\}$$

for all $u \in \mathbb{N}^k$. This is the the set that of all non-negative vectors in the same residue class mod $\mathcal{L}$.

Preliminaries

- All sublattices $\mathcal{L}$ of $\mathbb{Z}^k$ we will consider are such that the only non-negative member is the origin, that is $\mathcal{L} \cap \mathbb{N}^k = \{0\}$. For instance, the kernel of non-negative integer matrices are such.
- We define the fiber of a lattice in the following way:

$$\mathcal{F}(u) := (u + \mathcal{L}) \cap \mathbb{N}^k = \{v \in \mathbb{N}^k : u - v \in \mathcal{L}\}$$

  for all $u \in \mathbb{N}^k$. This is the the set that of all non-negative vectors in the same residue class mod $\mathcal{L}$.
- The condition above on the lattice $\mathcal{L}$ ensures that the fiber of any point is finite. Indeed, viewing $\mathcal{L}$ as a hyperplane of rank at most $k - 1$ passing through the origin, a finite vector u may only push $\mathcal{L}$ finitely far into $\mathbb{N}^k \subset \mathbb{Z}^k$

Preliminaries

- All sublattices $\mathcal{L}$ of $\mathbb{Z}^k$ we will consider are such that the only non-negative member is the origin, that is $\mathcal{L} \cap \mathbb{N}^k = \{0\}$. For instance, the kernel of non-negative integer matrices are such.
- We define the fiber of a lattice in the following way:

$$\mathcal{F}(u) := (u + \mathcal{L}) \cap \mathbb{N}^k = \{v \in \mathbb{N}^k : u - v \in \mathcal{L}\}$$

  for all $u \in \mathbb{N}^k$. This is the the set that of all non-negative vectors in the same residue class mod $\mathcal{L}$.
- The condition above on the lattice $\mathcal{L}$ ensures that the fiber of any point is finite. Indeed, viewing $\mathcal{L}$ as a hyperplane of rank at most $k-1$ passing through the origin, a finite vector u may only push $\mathcal{L}$ finitely far into $\mathbb{N}^k \subset \mathbb{Z}^k$
- NB: The sizes of the fibers are not uniformly bounded.

Uses for the bases we will define

- Counting $\mathcal{F}(u)$
- Enumerating $\mathcal{F}(u)$
- Optimizing $\mathcal{F}(u)$, that is, for vectors w minimize $v \cdot w$ for $v \in \mathcal{F}(u)$. A minimum always exists due to finiteness mentioned above.
- Sampling from $\mathcal{F}(u)$

Different Bases of a Lattice

- The bases that we will introduce are the following: Lattice, Markov, Gröbner, Universal Gröbner and Graver Basis. There is the following sequence of inclusions:

$$\text{Lattice basis} \subset \text{ Markov basis} \subset \text{ Gröbner basis}$$

$$\subset \text{Universal Gröbner basis} \subset \text{ Graver Basis basis}$$

which may or may not be strict.

Lattice Basis

- A Lattice basis for a lattice $\mathcal{L} \subset \mathbb{Z}^k$ is a set $\{b_1, ..., b_n\} \subset \mathcal{L}$ such that for each $v \in \mathcal{L}$ there is a unique vector $(a_1, a_2, ..., a_n) \in \mathbb{Z}^n$ such that

$$v = \sum_{i=1}^{n} b_i a_i.$$

  Note that $b_i = (b_{i_1}, b_{i_2}, ..., b_{i_k})$
- This basis is of course not unique, but its cardinality is. The cardinality of the basis is the rank of $\mathcal{L}$.

Running Example

- We will consider the lattice given by the kernel of the linear map

$$\mathbb{Z}^4 \to \mathbb{Z}^4$$

$$(u_1, u_2, u_3, u_4) \to 3u_1 + 3u_2 + 4u_3 + 5u_4$$

Running Example

- We will consider the lattice given by the kernel of the linear map

$$\mathbb{Z}^4 \to \mathbb{Z}^4$$

$$(u_1, u_2, u_3, u_4) \to 3u_1 + 3u_2 + 4u_3 + 5u_4$$

- A lattice basis is given by

$$(1, -1, 0, 0), (0, 1, -2, 1), (0, 3, -1, -1)$$

Markov Basis

- In order to define the Markov Basis of a lattice, we will need to define a graph $\mathcal{F}(u)_{\mathcal{B}}$ for each $u \in \mathbb{N}^k$ and all $\mathcal{B} \subset \mathcal{L}$

Markov Basis

- In order to define the Markov Basis of a lattice, we will need to define a graph $\mathcal{F}(u)_{\mathcal{B}}$ for each $u \in \mathbb{N}^k$ and all $\mathcal{B} \subset \mathcal{L}$
- For each $\mathcal{B} \subset \mathcal{L}$ we define $\mathcal{F}(u)_{\mathcal{B}}$ as the graph which nodes are the elements of $\mathcal{F}(u)$ and two nodes v, v' are connected by an undirected edge iff either $v - v' \in \mathcal{B}$ or $v' - v \in \mathcal{B}$.

Markov Basis

- In order to define the Markov Basis of a lattice, we will need to define a graph $\mathcal{F}(u)_{\mathcal{B}}$ for each $u \in \mathbb{N}^k$ and all $\mathcal{B} \subset \mathcal{L}$
- For each $\mathcal{B} \subset \mathcal{L}$ we define $\mathcal{F}(u)_{\mathcal{B}}$ as the graph which nodes are the elements of $\mathcal{F}(u)$ and two nodes v, v' are connected by an undirected edge iff either $v - v' \in \mathcal{B}$ or $v' - v \in \mathcal{B}$.
- A Markov basis for a lattice $\mathcal{L} \subset \mathbb{Z}^k$ is a set $\mathcal{B} \subset \mathcal{L}$ such that for each $u \in \mathbb{N}^k$ $\mathcal{F}(u)_{\mathcal{B}}$ is connected.

Running Example

- Recall that the lattice is given by

$$3u_1 + 3u_2 + 4u_3 + 5u_4 = 0$$

Running Example

- Recall that the lattice is given by

$$3u_1 + 3u_2 + 4u_3 + 5u_4 = 0$$

- A Markov Basis is given by

$$\mathcal{B} = \{(1, -1, 0, 0), (0, 1, -2, 1), (0, 3, -1, -1), (0, 2, 1, -2)\}$$

Running Example

- Recall that the lattice is given by

$$3u_1 + 3u_2 + 4u_3 + 5u_4 = 0$$

- A Markov Basis is given by

$$\mathcal{B} = \{(1, -1, 0, 0), (0, 1, -2, 1), (0, 3, -1, -1), (0, 2, 1, -2)\}$$

- The fact that the last vector is needed may be seen from considering the fiber of $u = (1, 1, 1, 0)$.

Running Example

- Recall that the lattice is given by

$$3u_1 + 3u_2 + 4u_3 + 5u_4 = 0$$

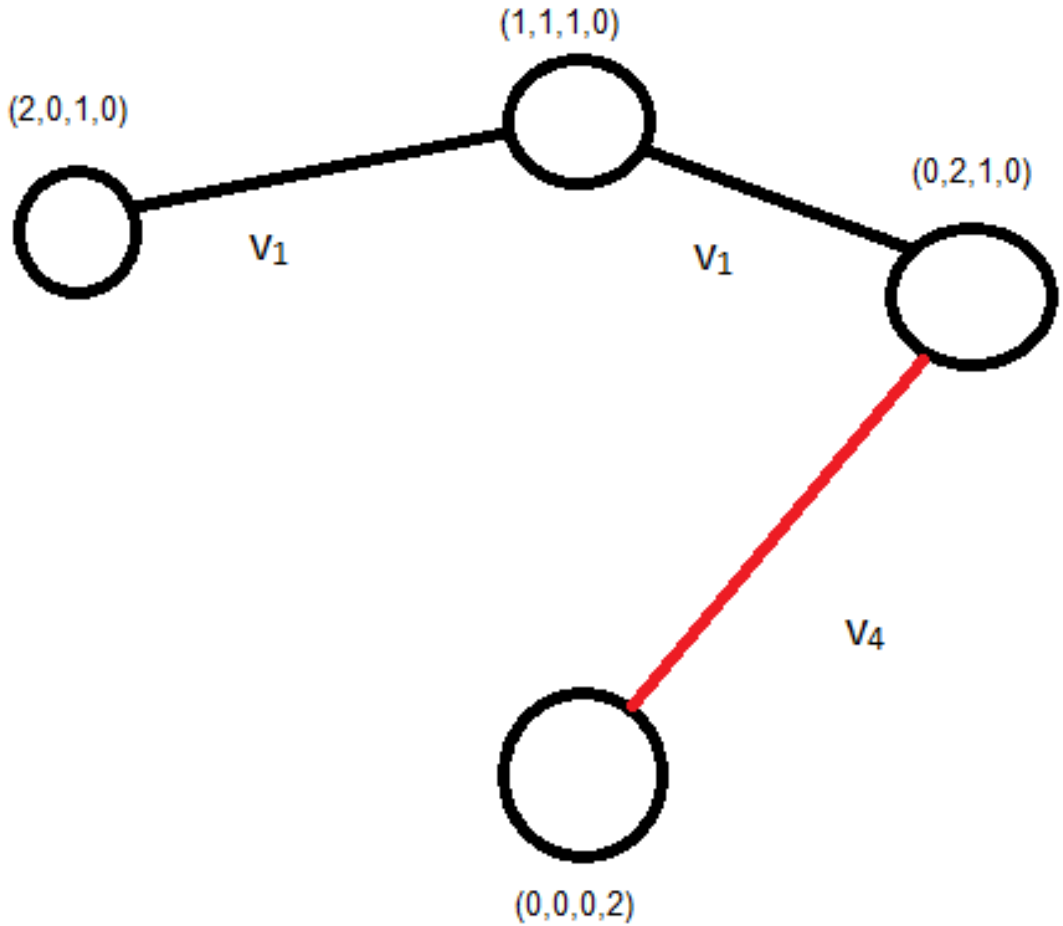
- A Markov Basis is given by

$$\mathcal{B} = \{(1, -1, 0, 0), (0, 1, -2, 1), (0, 3, -1, -1), (0, 2, 1, -2)\}$$

- The fact that the last vector is needed may be seen from considering the fiber of $u = (1, 1, 1, 0)$.
- Without $(0, 2, 1, -2)$, $\mathcal{F}(u)_{\mathcal{B}}$ would consist of the two connected components
  $C_1 = \{(2, 0, 1, 0), (1, 1, 1, 0), (0, 2, 1, 0)\}$ and
  $C_2 = \{(0, 0, 0, 2)\}$

Running Example

$$\{v_1, v_2, v_3, v_4\} = \{(1, -1, 0, 0), (0, 1, -2, 1), (0, 3, -1, -1), (0, 2, 1, -2)\}$$



A few definitions to state and prove a theorem

- For any vector b we define b^+, b^- as the vectors that are in $\mathbb{N}^k$ and are such that $b^+ - b^- = b$. We then define $\text{fiber}(b)$ as

$$\text{fiber}(b) := \mathcal{F}(b^+) = \mathcal{F}(b^-).$$

A few definitions to state and prove a theorem

- For any vector b we define b^+, b^- as the vectors that are in $\mathbb{N}^k$ and are such that $b^+ - b^- = b$. We then define $\text{fiber}(b)$ as

$$\text{fiber}(b) := \mathcal{F}(b^+) = \mathcal{F}(b^-).$$

- Note that indeed, $\mathcal{F}(b^+) = \mathcal{F}(b^-)$ since $b \in \mathcal{L}$.

A few definitions to state and prove a theorem

- For any vector b we define b^+, b^- as the vectors that are in $\mathbb{N}^k$ and are such that $b^+ - b^- = b$. We then define $\text{fiber}(b)$ as

$$\text{fiber}(b) := \mathcal{F}(b^+) = \mathcal{F}(b^-).$$

- Note that indeed, $\mathcal{F}(b^+) = \mathcal{F}(b^-)$ since $b \in \mathcal{L}$.
- A multiset is a set where we allow multiplity. For instance,

$$\{a\} \cup \{a\} = \{a, a\} \neq \{a\}.$$

A few definitions to state and prove a theorem

- For any vector b we define b^+, b^- as the vectors that are in $\mathbb{N}^k$ and are such that $b^+ - b^- = b$. We then define $\text{fiber}(b)$ as

$$\text{fiber}(b) := \mathcal{F}(b^+) = \mathcal{F}(b^-).$$

- Note that indeed, $\mathcal{F}(b^+) = \mathcal{F}(b^-)$ since $b \in \mathcal{L}$.
- A multiset is a set where we allow multiplity. For instance,

$$\{a\} \cup \{a\} = \{a, a\} \neq \{a\}.$$

- We will utilize the poset structure of the fibers of a vector $u \in \mathbb{N}^k$. That is, $f' \leq f$ iff there is u, u' such that

$$f = \mathcal{F}(u), f' = \mathcal{F}(u'), u \leq u' \text{ coordinate wise}$$

Invariance property of the Markov Basis

Theorem Suppose that $\mathcal{B}$ is a Markov Basis for the lattice $\mathcal{L}$. Then the multiset $\{\text{fiber}(b) : b \in \mathcal{B}\}$ is an invariant of $\mathcal{L}$

Proof Let $f \in \mathbb{N}^k/\mathcal{L}$ be a fiber, we will construct a graph G_f.

Invariance property of the Markov Basis

Theorem Suppose that $\mathcal{B}$ is a Markov Basis for the lattice $\mathcal{L}$. Then the multiset $\{\text{fiber}(b) : b \in \mathcal{B}\}$ is an invariant of $\mathcal{L}$

Proof Let $f \in \mathbb{N}^k/\mathcal{L}$ be a fiber, we will construct a graph G_f. The nodes are the elements of the fiber f, i.e. elements of the same congruence class mod $\mathcal{L}$.

Invariance property of the Markov Basis

Theorem Suppose that $\mathcal{B}$ is a Markov Basis for the lattice $\mathcal{L}$. Then the multiset $\{\text{fiber}(b) : b \in \mathcal{B}\}$ is an invariant of $\mathcal{L}$

Proof Let $f \in \mathbb{N}^k/\mathcal{L}$ be a fiber, we will construct a graph G_f. The nodes are the elements of the fiber f, i.e. elements of the same congruence class mod $\mathcal{L}$. Two nodes v, v' are connected iff there is i such that $v_i \neq 0$, $v'_i \neq 0$.

Invariance property of the Markov Basis

Theorem Suppose that $\mathcal{B}$ is a Markov Basis for the lattice $\mathcal{L}$. Then the multiset $\{\text{fiber}(b) : b \in \mathcal{B}\}$ is an invariant of $\mathcal{L}$

Proof Let $f \in \mathbb{N}^k/\mathcal{L}$ be a fiber, we will construct a graph G_f. The nodes are the elements of the fiber f, i.e. elements of the same congruence class mod $\mathcal{L}$. Two nodes v, v' are connected iff there is i such that $v_i \neq 0$, $v'_i \neq 0$. Note here that this implies that $\{v, v'\}$ is an edge iff $\text{fiber}(v - v') \neq f$.

Invariance property of the Markov Basis

Theorem Suppose that $\mathcal{B}$ is a Markov Basis for the lattice $\mathcal{L}$. Then the multiset $\{\text{fiber}(b) : b \in \mathcal{B}\}$ is an invariant of $\mathcal{L}$

Proof Let $f \in \mathbb{N}^k/\mathcal{L}$ be a fiber, we will construct a graph G_f. The nodes are the elements of the fiber f, i.e. elements of the same congruence class mod $\mathcal{L}$. Two nodes v, v' are connected iff there is i such that $v_i \neq 0$, $v'_i \neq 0$. Note here that this implies that $\{v, v'\}$ is an edge iff $\text{fiber}(v - v') \neq f$. Note also, and this is key, that in this case, $\text{fiber}(v - v') = f'$ for some $f' < f$.

Continuing the proof

Introduce the multiset

$$A = \{f \in \mathbb{N}^k/\mathcal{L} : G_f \text{ is disconnected}\}.$$

Continuing the proof

Introduce the multiset

$$A = \{f \in \mathbb{N}^k/\mathcal{L} : G_f \text{ is disconnected}\}.$$

f has multiplicity $s-1$, where s is the number of disconnected components in G_f.

Continuing the proof

Introduce the multiset

$$A = \{f \in \mathbb{N}^k/\mathcal{L} : G_f \text{ is disconnected}\}.$$

f has multiplicity $s-1$, where s is the number of disconnected components in G_f. Let $f = \mathcal{F}(u)$ be a specific fiber and let $C_1, ..., C_s$ be the disconnected components of G_f.

Continuing the proof

Introduce the multiset

$$A = \{f \in \mathbb{N}^k/\mathcal{L} : G_f \text{ is disconnected}\}.$$

f has multiplicity $s-1$, where s is the number of disconnected components in G_f. Let $f = \mathcal{F}(u)$ be a specific fiber and let $C_1, ..., C_s$ be the disconnected components of G_f. Define $\mathcal{B}_f = \{b \in \mathcal{B} : \text{ fiber}(b) = f\}$ for some minimal Markov Basis $\mathcal{B}$.

Continuing the proof

Introduce the multiset

$$A = \{f \in \mathbb{N}^k/\mathcal{L} : G_f \text{ is disconnected}\}.$$

f has multiplicity $s-1$, where s is the number of disconnected components in G_f. Let $f = \mathcal{F}(u)$ be a specific fiber and let $C_1, ..., C_s$ be the disconnected components of G_f. Define $\mathcal{B}_f = \{b \in \mathcal{B} : \text{ fiber}(b) = f\}$ for some minimal Markov Basis $\mathcal{B}$. Our aim is to prove that $\mathcal{B}_f$ has cardinality $s-1$, and in order to accomplish this we will use induction.

Continuing the proof

Suppose that $\mathcal{B}_{f'}$ has already been constructed for all $f' < f$.

Continuing the proof

Suppose that $\mathcal{B}_{f'}$ has already been constructed for all $f' < f$.
Let $\mathcal{B}_{<f}$ be the union of these $\mathcal{B}_{f'}$.

Continuing the proof

Suppose that $\mathcal{B}_{f'}$ has already been constructed for all $f' < f$. Let $\mathcal{B}_{<f}$ be the union of these $\mathcal{B}_{f'}$. The main idea of the proof is that the connected components of $\mathcal{F}(u)_{\mathcal{B}_{<f}}$ are precisely $C_1, ..., C_s$.

Continuing the proof

Suppose that $\mathcal{B}_{f'}$ has already been constructed for all $f' < f$. Let $\mathcal{B}_{<f}$ be the union of these $\mathcal{B}_{f'}$. The main idea of the proof is that the connected components of $\mathcal{F}(u)_{\mathcal{B}_{<f}}$ are precisely $C_1, ..., C_s$. This is due to the key thing we pointed out before: $\{v, v'\}$ is an edge in G_f iff $\text{fiber}(v - v') = f'$ for some $f' < f$.

Continuing the proof

Now, fix a spanning tree of $\{C_1, ..., C_s\}$.

Continuing the proof

Now, fix a spanning tree of $\{C_1, ..., C_s\}$. That is, make it a connected graph with as few edges as possible.

Continuing the proof

Now, fix a spanning tree of $\{C_1, ..., C_s\}$. That is, make it a connected graph with as few edges as possible. For each edge $\{C_i, C_j\}$ we pick $v \in C_i, v' \in C_j$.

Continuing the proof

Now, fix a spanning tree of $\{C_1, ..., C_s\}$. That is, make it a connected graph with as few edges as possible. For each edge $\{C_i, C_j\}$ we pick $v \in C_i, v' \in C_j$. We have that $\text{fiber}(v - v') = f$ and $\mathcal{B}_f$ consists of the $s - 1$ vectors $v - v'$.

Continuing the proof

Now, fix a spanning tree of $\{C_1, ..., C_s\}$. That is, make it a connected graph with as few edges as possible. For each edge $\{C_i, C_j\}$ we pick $v \in C_i, v' \in C_j$. We have that $\text{fiber}(v - v') = f$ and $\mathcal{B}_f$ consists of the $s - 1$ vectors $v - v'$. It is clear that we must choose our vectors precisely as we do above, as adding one more of the same form renders $\mathcal{B}$ not minimal.

Continuing the proof

Now, fix a spanning tree of $\{C_1, ..., C_s\}$. That is, make it a connected graph with as few edges as possible. For each edge $\{C_i, C_j\}$ we pick $v \in C_i, v' \in C_j$. We have that $\text{fiber}(v - v') = f$ and $\mathcal{B}_f$ consists of the $s - 1$ vectors $v - v'$. It is clear that we must choose our vectors precisely as we do above, as adding one more of the same form renders $\mathcal{B}$ not minimal. Moreover, we need all $s - 1$ vectors $v - v'$ to make $F(u)_{\mathcal{B}_f}$ connected. We are finished.

Continuing the proof

Now, fix a spanning tree of $\{C_1, ..., C_s\}$. That is, make it a connected graph with as few edges as possible. For each edge $\{C_i, C_j\}$ we pick $v \in C_i, v' \in C_j$. We have that $\text{fiber}(v - v') = f$ and $\mathcal{B}_f$ consists of the $s - 1$ vectors $v - v'$. It is clear that we must choose our vectors precisely as we do above, as adding one more of the same form renders $\mathcal{B}$ not minimal. Moreover, we need all $s - 1$ vectors $v - v'$ to make $F(u)_{\mathcal{B}_f}$ connected. We are finished. Note that of course, we might be able to choose v, v' in many different ways, hence $\mathcal{B}$ is not unique.

Algorithm obtained from the proof

- Identify the connected components $C_1, ..., C_s$ of the graph G_f .

Algorithm obtained from the proof

- Identify the connected components $C_1, ..., C_s$ of the graph G_f .
- Pick a spanning tree on $C_1, ..., C_s$.

Algorithm obtained from the proof

- Identify the connected components $C_1, ..., C_s$ of the graph G_f .
- Pick a spanning tree on $C_1, ..., C_s$.
- For any edge $\{C_i, C_j\}$ of the spanning tree, pick points $v \in C_i$ and $v' \in C_j$.

Algorithm obtained from the proof

- Identify the connected components $C_1, ..., C_s$ of the graph G_f .
- Pick a spanning tree on $C_1, ..., C_s$.
- For any edge $\{C_i, C_j\}$ of the spanning tree, pick points $v \in C_i$ and $v' \in C_j$.
- Define $\mathcal{B}_f$ as the set of $s-1$ difference vectors $v - v'$.

Algorithm obtained from the proof

- Identify the connected components $C_1, ..., C_s$ of the graph G_f .
- Pick a spanning tree on $C_1, ..., C_s$.
- For any edge $\{C_i, C_j\}$ of the spanning tree, pick points $v \in C_i$ and $v' \in C_j$.
- Define $\mathcal{B}_f$ as the set of $s-1$ difference vectors $v - v'$.
- Move on to the next fiber (unless you are sure to be done).

Notes on the proof above

- There are certain issues with the combinatorial algorithm that the proof above provides
- Firstly, it doesn't give a termination condition - we do not know when we are finished
- Secondly, As we mentioned before, the fibers can become arbitrarily large, which makes the number of calculations needed very large.
- To solve this latter problem, we may use the Graver Basis, which I will not go into more detail here.
- However, there is another way, which uses tools from algebra, and we will go through this now

Quick refresh on ideals

- Recall the definition of ideals: they are subsets I of rings R such that $(I, +)$ is a subgroup of R and for all $r \in R$ and $i \in I, ri = ir \in I$.
- $$\langle a_1, a_2, ..., a_n \rangle = \left\{ \sum_{i=1}^{n} r_i a_i, r_i \in R, n \in \mathbb{N} \right\}$$
  is the ideal generated by $a_1, a_2, ..., a_n$.
- A prime ideal is a proper ideal such that if $ab \in I$ then $a \in I$ or $b \in I$ (or both).
- Recall also that homogeneous polynomials are such that all terms have the same degree.

Lattice Ideal corresponding to our lattice

- Let $\mathcal{L} \subset \mathbb{Z}^k$ be given. The corresponding *Lattice Ideal* is given by

$$I_{\mathcal{L}} := \langle p^u - p^v : u, v \in \mathbb{N}^k \text{ and } u - v \in \mathcal{L} \rangle \subset \mathbb{R}[p_1, p_2, ., , , p_k]$$

- p_i are our indeterminates and $p^u = p_1^{u_1} p_2^{u_2} ... p_k^{u_k}$.
- In our setting, we will let $p_i = P(X = i)$ where X is some random variable with k outcomes.
- We know that every ideal in $\mathbb{R}[p_1, ..., p_k]$ is finitely generated. With this fact, the theorem below proves finiteness of the Markov Basis. Note that we have proved constant cardinality, but not finiteness.
- **Theorem** $\mathcal{B} \subset \mathcal{L}$ is a Markov Basis iff the set of binomials $\{p^{b^+} - p^{b^-} : b \in \mathcal{B}\}$ generates $I_{\mathcal{L}}$.

Gröbner Basis

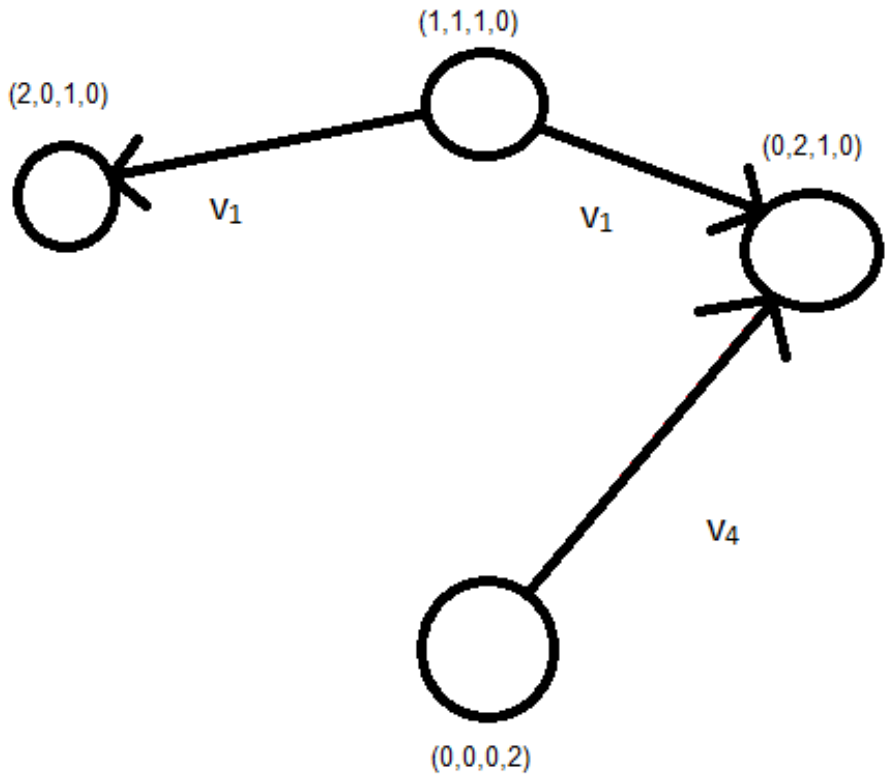
- In order to define a Gröbner Basis, we let a vector $w \in \mathbb{R}^k$ be given. We suppose that w is such that the problem of minimizing $v \cdot w$ for $v \in \mathcal{F}(u)$ has a unique solution for each $u \in \mathbb{N}^k$. The complement of the set of such w have zero measure.
- Next, let $\mathcal{B} \subset \mathcal{L}$ be such that for $b \in \mathcal{B}$, $b \cdot w < 0$.
- We once again define the graph $\mathcal{F}(u)_{\mathcal{B}}$ which nodes are the vectors in $\mathcal{F}(u)$ and for any two vectors $v, v' \in \mathcal{F}(u)$, we introduce a directed edge $v \longrightarrow v'$ if $v' - v \in \mathcal{B}$.

The graph $\mathcal{F}(u)_{\mathcal{B}}$

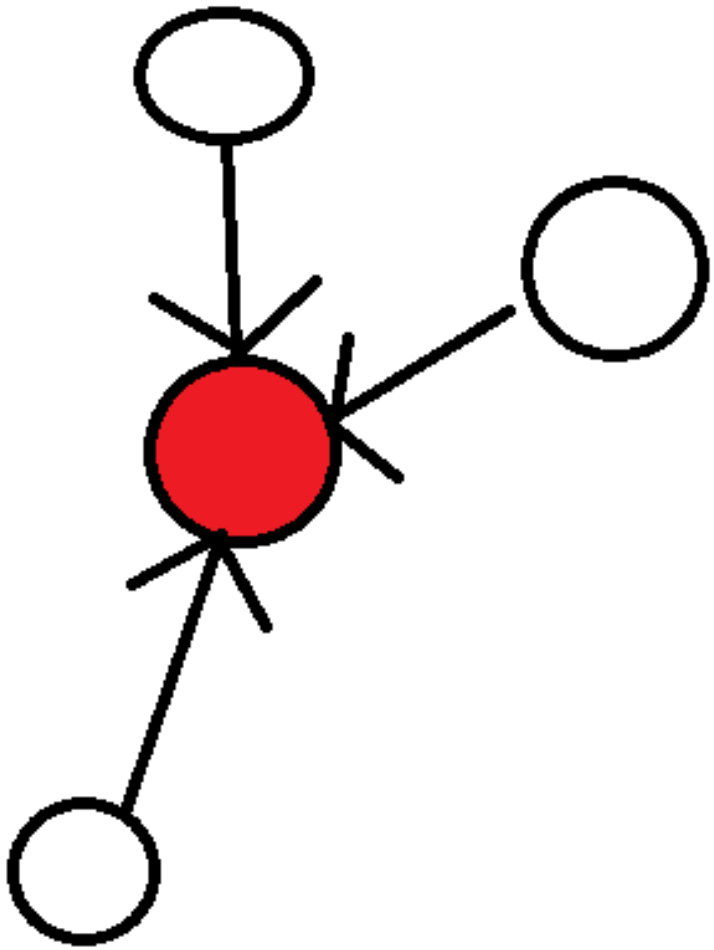
- Directed edge $v \longrightarrow v'$ if $v' - v \in \mathcal{B}$

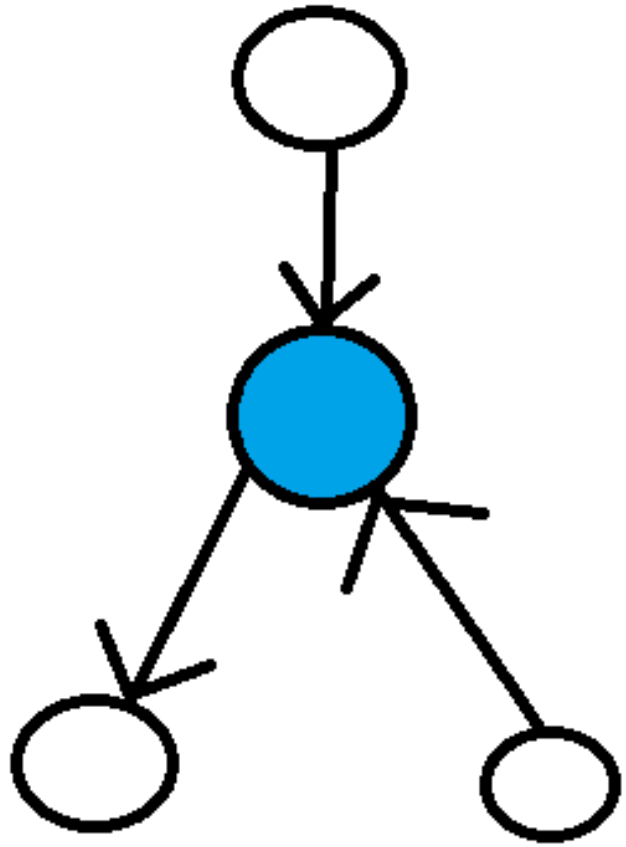
$$\mathcal{B} = \{v_1, v_2, v_3, v_4\}$$
$$= \{(1,-1,0,0), (0,1,-2,1), (0,3,-1,-1), (0,2,1,-2)\}$$

- The basis above is NOT a Gröbner basis!



Sink vs. not a sink





Gröbner Basis

- We say that $\mathcal{B}$ is a Gröbner basis if for each $u \in \mathcal{F}(u)$, $\mathcal{F}(u)_{\mathcal{B}}$ has a unique sink.
- A Gröbner Basis always exists since if $(v' - v) \cdot w < 0$ and $(v'' - v') \cdot w < 0$ then

$$(v'' - v) \cdot w < 0$$

- A Gröbner basis is such that the sink is the minimizer of the optimization problem defined above.
- Gröbner bases are not unique, even when fixing w.

Running Example

- Recall that the lattice is given by

$$3u_1 + 3u_2 + 4u_3 + 5u_4 = 0$$

Running Example

- Recall that the lattice is given by

$$3u_1 + 3u_2 + 4u_3 + 5u_4 = 0$$

- We let $w = (100, 10, 1, 0)$

Running Example

- Recall that the lattice is given by

$$3u_1 + 3u_2 + 4u_3 + 5u_4 = 0$$

- We let $w = (100, 10, 1, 0)$
- To the Markov Basis

$$\{(1, -1, 0, 0), (0, 1, -2, 1), (0, 3, -1, -1), (0, 2, 1, -2)\}$$

we add $(0, 1, 3, -3), (0, 0, 5, -4)$

Universal Gröbner Basis

- A reduced Gröbner Basis is the a set $\mathcal{B}$ such that if $b \in \mathcal{B}$, then b^- is a sink, b^+ is not a sink but $b^+ - e_i$ is a sink for all i such that $b_i^+ \neq 0$.
- The reduced Gröbner basis is unique for fixed w.
- The Universal Gröbner Basis is the union of all reduced Gröbner bases.
- As the name suggests, the Universal Gröbner Basis is unique.

Running Example

- Recall that the lattice is given by

$$3u_1 + 3u_2 + 4u_3 + 5u_4 = 0$$

Running Example

- Recall that the lattice is given by

$$3u_1 + 3u_2 + 4u_3 + 5u_4 = 0$$

- We let $w = (100, 10, 1, 0)$

Running Example

- Recall that the lattice is given by

$$3u_1 + 3u_2 + 4u_3 + 5u_4 = 0$$

- We let $w = (100, 10, 1, 0)$
- To the Gröbner basis

$$\{(1, -1, 0, 0), (0, 1, -2, 1), (0, 3, -1, -1), (0, 2, 1, -2),$$

$$(0, 1, 3, -3), (0, 0, 5, -4)\}$$

we add

$$(1, 0, -2, 1), (3, 0, -1, -1), (2, 0, 1, -2), (1, 0, 3, -3),$$

$$(0, 4, -3, 0), (4, 0, -3, 0), (0, 5, 0, -3), (5, 0, 0, -3)$$

Graver Basis

- We are finally ready to define the last basis. Let a sign vector $\sigma \in \{-1, 1\}^k$ be given and consider

$$\mathcal{L}_\sigma = \{v \in \mathcal{L} : \forall i, \sigma_i v_i \geq 0\}$$

- This set is closed under addition and has a unique, minimal and finite generating set $\mathcal{G}_\sigma$.
- We define the Graver Basis $\mathcal{G}$ of $\mathcal{L}$ as

$$\mathcal{G} = \bigcup_{\sigma \in \{-1,1\}^k} \mathcal{G}_\sigma.$$

- The Graver basis is the unique smallest subset of $\mathcal{L}$ that is such that for each $v \in \mathcal{L}$ there is $\lambda_g \in \mathbb{N}$ such that

$$v = \sum_{g \in \mathcal{G}} \lambda_g g, |v_i| = \sum_{g \in \mathcal{G}} \lambda_g |g_i|$$

Running Example

- Recall that the lattice is given by

$$3u_1 + 3u_2 + 4u_3 + 5u_4 = 0$$

Running Example

- Recall that the lattice is given by

$$3u_1 + 3u_2 + 4u_3 + 5u_4 = 0$$

- In order to form a Graver Basis, we add to the Universal Gröbner Basis above the vectors

$$(1,1,1,-2),(1,2,-1,-1),(2,1,-1,-1),(1,3,-3,0),(2,2,-3,0),$$

$$(3,1,-3,0),(1,4,0,-3),(2,3,0,-3),(3,2,0,-3),(4,1,0,-3).$$

Back to the story of log-linear models

- Let $A = (a_{i,j}) \in \mathbb{N}^{d \times k}$ be a matrix which column sums are all equal.

Back to the story of log-linear models

- Let $A = (a_{i,j}) \in \mathbb{N}^{d \times k}$ be a matrix which column sums are all equal.
- Every column $a_j = (a_{1_j}, ..., a_{d_j})^T$ represents a monomial $\theta^{a_j} = \theta_1^{a_{1_j}} ... \theta_d^{a_{d_j}}$.

Back to the story of log-linear models

- Let $A = (a_{i,j}) \in \mathbb{N}^{d \times k}$ be a matrix which column sums are all equal.
- Every column $a_j = (a_{1_j}, ..., a_{d_j})^T$ represents a monomial $\theta^{a_j} = \theta_1^{a_{1_j}} ... \theta_d^{a_{d_j}}$.
- A defines a map $\phi_A : \mathbb{C}^d \to \mathbb{C}^k$ by

$$\theta \to (\theta^{a_1}, \theta^{a_2}, ..., \theta^{a_k})$$

Back to the story of log-linear models

- Let $A = (a_{i,j}) \in \mathbb{N}^{d \times k}$ be a matrix which column sums are all equal.
- Every column $a_j = (a_{1_j}, ..., a_{d_j})^T$ represents a monomial $\theta^{a_j} = \theta_1^{a_{1_j}} ... \theta_d^{a_{d_j}}$.
- A defines a map $\phi_A : \mathbb{C}^d \to \mathbb{C}^k$ by

$$\theta \to (\theta^{a_1}, \theta^{a_2}, ..., \theta^{a_k})$$

- The closure of the image of this map is called the *Affine Toric Variety*, V_A.

Back to the story of log-linear models

- Let $A = (a_{i,j}) \in \mathbb{N}^{d \times k}$ be a matrix which column sums are all equal.
- Every column $a_j = (a_{1_j}, ..., a_{d_j})^T$ represents a monomial $\theta^{a_j} = \theta_1^{a_{1_j}} ... \theta_d^{a_{d_j}}$.
- A defines a map $\phi_A : \mathbb{C}^d \to \mathbb{C}^k$ by

$$\theta \to (\theta^{a_1}, \theta^{a_2}, ..., \theta^{a_k})$$

- The closure of the image of this map is called the *Affine Toric Variety*, V_A.
- If we restrict ϕ_A to $\mathbb{R}^d_{\geq 0}$ and consider its image in the probability simplex Δ_{k-1} we obtain the log-linear model $\mathcal{M}_A$.

Theorem

Theorem If $\mathcal{L} = \ker_Z(A)$ then the lattice ideal $I_\mathcal{L}$ is a prime ideal. The homogeneous polynomials contained in it are exactly the homogeneous polynomials in $\mathbb{R}[p_1, ..., p_k]$ that vanish on probability distributions in the log-linear model specified by the matrix A. In other words, the toric variety $V_A = V(I_\mathcal{L})$ is the Zariski closure of the log-linear model $\mathcal{M}_A$

Application of the Lattice Ideal

Let

$$A = \begin{bmatrix} 3 & 0 & 0 & 2 & 1 & 2 & 1 & 0 & 0 \\ 0 & 3 & 0 & 1 & 2 & 0 & 0 & 2 & 1 \\ 0 & 0 & 3 & 0 & 0 & 1 & 2 & 1 & 2 \end{bmatrix}$$

and consider the the assicoated log-linear model $\mathcal{M}_A$. We may find the Markov Basis for the lattice $\mathcal{L} = \ker_{\mathbb{Z}}(A)$ using the algorithm from the proof above, and after each step (or doing it smarter) checking if the corresponding binomials $\{p^{b^+} - p^{b^-} : b \in \mathcal{B}\}$ generate the lattice ideal $I_{\mathcal{L}}$. The basis has 17 vectors so in practice, we would certainly use a computer.

Rock Paper Scissors

Suppose now Bobby and Sally plays a game of Rock Paper Scissors, each round consists of three games and you may pick at most 2 different choices each round. After 1000 rounds, Sally want to analyze Bobby's choices and suspects he picks the three choices independently, but not necessarily with equal probability. She introduces

$$u = (u_{rrr}, u_{ppp}, u_{sss}, u_{rrp}, u_{rrs}, u_{ppr}, u_{pps}, u_{ssr}, u_{ssp})$$

where u_{rrs} is the number of rounds Bobby played 2 rocks and 1 scissor, and so forth.

Rock Paper Scissors

Let $p_{ppp}, p_{rrr}...$ be the true probabilities of bobbys choices. Sally introduces

$$v = (3p_{rrr}, 3p_{ppp}, 3p_{sss}, p_{rrp}, p_{rrs}, p_{ppr}, p_{pps}, p_{ssr}, p_{ssp})$$

Under the null hypothesis of random independent choices we have $v \in \mathcal{M}_A$ and so we have that $v \in V_A$, i.e. v vanishes for all $f \in I_{\mathcal{L}}$ ($\mathcal{L} = \ker_{\mathbb{Z}}(A)$). That is, if $p = v$ then $p^w - p^{w'} = 0$ for all $w, w' \in \mathbb{N}^k, w - w' \in \mathcal{L}$.

Rock Paper Scissors

Hence, all Sally needs to do to test if Bobby is indeed making independent choices is to use the Markov Basis of $\mathcal{L} = \ker_{\mathbb{Z}}(A)$, the Metropolis-Hastings Algorithm and the hypothesis testing methods from Lecture 1. However, we need to change a few terms in the hypergeometric distribution slightly in order to take the constant 3 in

$$v = (3p_{rrr}, 3p_{ppp}, 3p_{sss}, p_{rrp}, p_{rrs}, p_{ppr}, p_{pps}, p_{ssr}, p_{ssp})$$

into consideration.

Relational data in practice - A few notes on the method above

- The Markov Bases may be complicated to compute and the algebra used to find them may produce moves that inapplicable to observed data.
- For certain models, new algorithms have been found to circumvent this, which also provides a scalable exact conditional test.

Relational data in practice - Heuristic Tests

- Heuristic tests are based on graphical comparisons between observed statistics and random ones obtained from the fitted model.
- Say we want to estimate how well a model $P_\theta(G)$ fits an observed graph g_{obs}
- We compute a maximum likelihood estimator for $\hat{\theta}$ of θ.
- We calculate some network statistics $s(g_{obs})$, say number of edges or the degrees of the nodes, and then compare it to $s(g_1), .., .s(g_n)$ where g_i are simulated graphs from $P_{\hat{\theta}}$.
- One issue that there is no obvious discrepancy measure between the observed graph and the model.
- A second is that the distribution of $s(g)$ is not immediately known under the hypothesis that g_{obs} fits the model $P_\theta(G)$.

Relational data in practice - Asymptotic tests

- Asymptotic tests solves the issues presented above by providing formal testing criteria for evaluating model fit
- However, classical tools such as the log-likelihood ratio test are not directly applicable since the usual asymptotics do not apply to many complex models, since the iid assumption on random edges does not hold.
- There are remedies to these issues, but they are often used on a case-by-case basis.