Algebraic Statistics

Notes taken by Viet Duc Nguyen

Contents

1 Dimension Theory

Definition 1.1 (Affine Hilbert function). Let $I \subset k[x_1, ..., x_n]$ be an ideal. The **affine Hilbert function** of I is defined to be

$$\mathrm{aHF}_{R/I} : s \mapsto \dim\left(R_{\leq s}/I_{\leq s}\right).$$

Remark 1.2 (Finite dimensional vector space). Note that $R_{\leq s}/I_{\leq s}$ is a subspace of the k-vector space $R_{\leq s}$, the latter is a *finite-dimensional* vector space since there exist $\binom{n+s}{s}$ monomials of degree $\leq s$; these monomials form a *basis*. So both the vector space and the subspace are finite dimensional and we can compute $\dim\left(R_{\leq s}/I_{\leq s}\right) = \dim\left(R_{\leq s}\right) - \dim(I_{\leq s})$.

For a monomial ideal we have an alternative interpretation of the affine Hilbert function: it counts the number of monomials not in the ideal.

Proposition 1.3 (Affine Hilbert function of monomial ideals). Let I be a monomial ideal. Then $\mathrm{aHF}_{R/I}(s)$ is equivalent to the map

$$\mathrm{aHF}_{R/I} : s \mapsto \text{ counts the number of monomial of degree} \leq s \text{ not in } I.$$

Remark 1.4. If I is a monomial ideal, we know for sufficiently large s the above function can be represented by a polynomial, which we call the **Hilbert polynomial** $\mathrm{aHP}_{R/I}$. Moreover, this polynomial is of degree $\dim(V(I))$, where by definition $\dim(V(I))$ is defined as the dimension of the largest coordinate subspace in $V(I)$.

Proposition 1.5 (Reduction to monomial ideals). For any graded order and any ideal I, we have $\mathrm{aHF}_{R/I} = \mathrm{aHF}_{R/(\mathrm{LT}(I))}$.

This allows us to define the Hilbert polynomial for arbitrary ideals. Just pick any graded order and define aHP to be the polynomial representing $\mathrm{aHF}_{R/\mathrm{LT}(I)}$.

$$\mathrm{aHF}_{R/I} \coloneqq \mathrm{aHF}_{R/\mathrm{LT}(I)} = C(\mathrm{LT(I)}) = \text{Hilbert polynomial of } \mathrm{LT}(I)$$

Definition 1.6 (Affine Hilbert polynomial). Let I be an ideal in $k[x_1, ..., x_n]$. For sufficiently large s, the polynomial $\mathrm{aHP}_{R/I}$ that equals $\mathrm{aHF}_{R/I}$ is called the **affine Hilbert polynomial**.

As previously stated, the degree of the affine Hilbert polynomial equals the dimension of $V(I)$ if I is a monomial ideal.

Definition 1.7 (Dimension of a variety). The **dimension of a variety** $V \subset k^n$ is the degree of the affine Hilbert polynomial $\mathrm{aHP}_{R/I(V)}$.

We gave a purely algebraic description of the dimension of a variety:

$$\text{dimension of a variety} \;=\; \text{degree of a polynomial}$$

Remark 1.8 (Warning). Let V be any variety with $V = V(I)$ for some ideal I. Then the degree of the Hilbert polynomial of I need not be equal to the dimension of V. This only holds for algebraically closed fields (if $k = \bar{k}$, then the Nullstellensatz holds and $I(V(I)) = \sqrt{I}$).

$$I \text{ such that } V = V(I) \not\Rightarrow \dim(I) = \dim(V)$$

Proposition 1.9 (Characterization of zero dimensional varieties). Let $V \subset k[x_1, ..., x_n]$ be a nonempty affine variety. Then

$$|V| < \infty \iff \dim(V) = 0.$$

Proof. If V is empty, then the dimension is not defined. So assume $V \neq \emptyset$.

- $\Longrightarrow$: Assume that $V = \{v_1, ..., v_k\} \subset \mathbb{R}^n$. For each $i = 1, ..., n$ we define the polynomial

  $$f_i(x) = (x_i - v_{1i})(x_i - v_{2i}) \cdots (x_i - v_{ki}) \in I(V).$$

  Observe that $\mathrm{LT}(f_i) = x_i^k$ for any graded order. So $(\mathrm{LT}(I(V)))$ contains $x_1^k, \ldots, x_n^k$. By definition,

  $$\dim(V) = \deg(\mathrm{aHP}_{R/I(V)}) = \deg(\mathrm{aHP}_{R/\mathrm{LT}(I(V))}).$$

  The degree of the Hilbert polynomial of a monomial ideal J equals the dimension of $V(J)$ where the dimension of $V(J)$ is defined to be the dimension of the largest coordinate subspace in $V(J)$. Thus, by setting $J = \mathrm{LT}(I(V))$, we obtain

  $$\deg(\mathrm{aHP}_{R/\mathrm{LT}(I(V))}) = \dim(V(\mathrm{LT}(I(V)))).$$

  Since $\mathrm{LT}(I(V))$ contains $x_1^k, \ldots, x_n^k$, its vanishing ideal consists of points with $x_1 = ... = x_n = 0$. Hence, $V(\mathrm{LT}(I(V))) = \{0\}$. Clearly, $\dim(\{0\}) = 0$ (since any coordinate subspace of $\{0\}$ is of dimension 0).

- $\Longleftarrow$: Let V be of dimension 0. Hence, the Hilbert polynomial of $I(V)$ is a constant for sufficiently large s. This means

  $$\dim(k[x_1, ..., x_n]_{\leq s}/I(V)_{\leq s}) = C.$$

  Let $s \geq C$. Then for any $i = 1, ..., n$ the set of vectors $x_i^{\{0,\ldots,s\}}$ is linearly dependent in $k[x_1, ..., x_n]_{\leq s}/I(V)_{\leq s}$. So, define the polynomial f_i to be

  $$0 \neq f_i := \sum_{k=0}^{s} \alpha_k x_i^k \in I(V)_{\leq s}.$$

  Since this holds for any $s \geq C$, $f_i \neq 0$ in $I(V)$. Hence, $f_i \in I(V)$ has only finitely many roots (since it is nonzero); also f_i vanishes on V. Thus, V has only finitely elements $y \in V$ with different coordinates y_i. Since i was chosen arbitrarily, V is finite.

$\square$

2 Maximum Likelihood Estimation

Definition 2.1 (Parameter space). An open subset $\Theta \subset \mathbb{R}^d$ is called the **parameter space**. Elements $\theta = (\theta_1, ..., \theta_d) \in \Theta$ are called **parameters**.

Definition 2.2 (Algebraic statistical model). An **algebraic statistical model** is a map $\mathbf{f} = (f_1, ..., f_m) : \mathbb{C}^d \to \mathbb{C}^m$ with $f_i \in \mathbb{Q}[\theta_1, ..., \theta_d]$ such that

- $f_1 + ... + f_m - 1 = 0 \in \mathbb{Q}[\theta_1, ..., \theta_d]$ is the zero polynomial, and
- $\mathbf{f}(\theta) > 0$ for all parameters $\theta \in \Theta$.

For each parameter $\theta \in \Theta$ a statistical model $\mathbf{f}$ defines a **probability distribution** on the state space $\{1, ..., m\}$, that is, $f_i(\theta) = p_i$ means that state $i \in \{1, ..., m\}$ occurs with probability $p_i \in [0, 1]$ for parameter θ.

Assume we are given the number of occurrences of states $1, ..., m$ of an experiment by a vector $\mathbf{u} = (u_1, ..., u_m) \in \mathbb{N}^m$. Fix a parameter $\theta \in \Theta$. The probability that the state $i \in \{1, ..., m\}$ appears u_i times is given by

$$f_i(\theta)^{u_i}.$$

The problem of **maximum likelihood estimation** is to find the best parameter θ that maximizes $\prod_{i=1}^m f_i(\theta)^{p_i}$. Maximizing this function is equivalent to maximizing the so called **log-likelihood function**

$$\ell_u(\theta) = \sum_{i=1}^m u_i \cdot \log f_i(\theta).$$

From calculus, we know that a necessary condition for a local and global maximum $\hat{\theta}$ is that the derivative of ℓ_u must vanish at $\hat{\theta}$ (note that if Θ were not open, then the derivative need not vanish at a global maximum; on the other hand a global maximum need no exist). Thus, we need to find a solution to d-many equations, called the **critical equations**

$$\frac{\partial \ell_u}{\partial \theta_1} = \sum_{i=1}^m \frac{u_i}{f_i} \frac{\partial f_i}{\partial \theta_1} = 0$$

$$\dots$$

$$\frac{\partial \ell_u}{\partial \theta_d} = \sum_{i=1}^m \frac{u_i}{f_i} \frac{\partial f_i}{\partial \theta_d} = 0$$

Our goal is to find all solutions $\theta \in \mathbb{C}^d$ to the critical equations.

Let $\mathcal{H}$ be the locus where all the denominators of the rational functions in the critical equations vanish. The set of solutions $\theta \in \Theta$ outside $\mathcal{H}$ is an *algebraic variety* in $\mathbb{C}^d$ called the **likelihood variety**.

Proposition 2.3. For generic data u, the number of solutions to the critical equations is independent of u.

Proof.

$$\frac{\partial}{\partial \theta_i} \log \frac{f_j}{g_j} = \frac{g_j}{f_j} \cdot \left(\frac{\partial f_j g_j - \partial g_j f_j}{g_j^2} \right) = \frac{\partial f_j g_j - \partial g_j f_j}{f_j g_j} = \frac{\partial f_j}{f_j} - \frac{\partial g_j}{g_j}$$

$\square$

2.1 Computing the likelihood variety

The ideal $(\frac{\partial \ell_u}{\partial \theta_1}, \dots, \frac{\partial \ell_u}{\partial \theta_d})$ is generated by *rational* functions. Let's find another set of generators that consists of only polynomials. We introduce unknowns $z = z_1, ..., z_m$ where z_i represents $f_i^{-1} = \frac{1}{f_i}$. So, we have two polynomial rings $\mathbb{Q}[\theta]$ and $\mathbb{Q}[\theta, z]$; clearly

$$\mathbb{Q}[\theta] \hookrightarrow \mathbb{Q}[\theta, z].$$

Consider the ideal J_u generated by $d + m$ polynomials in $\mathbb{Q}[\theta, z]$

$$J_u := \left(\sum_{i=1}^{m} u_i z_i \frac{\partial f_i}{\partial \theta_1}, \dots, \sum_{i=1}^{m} u_i z_i \frac{\partial f_i}{\partial \theta_d}, z_1 f_1 - 1, \dots, z_m f_m - 1 \right).$$

A point $(\theta, z) \in \mathbb{C}^{d+m}$ lies in the variety $V(J_u)$ if and only if

1. θ is a solution to the critical equations,
2. $f_i(\theta) \neq 0$, and
3. $z_i = f_i^{-1}(\theta)$.

Next, we compute the **elimination ideal** of J_u in $\mathbb{Q}[\theta]$, that is

$$I_u := J_u \cap \mathbb{Q}[\theta]$$

We call I_u the **likelihood ideal** of the model $\mathbf{f}$ with respect to the data u. A point $\theta \in \mathbb{C}^d$ with $f_i(\theta) \neq 0$ lies in $V(I_u)$ if and only if θ is solution to the critical equations. **Thus, $V(I_u)$ is the likelihood variety.**

Remark 2.4 (Algorithm).

1. Compute the likelihood ideal: $I_u = J_u \cap \mathbb{Q}[\theta]$

2. Compute $V(I_u)$ (for example by computing a Gröbner basis).

3. Compute $S = V(I_u) \cap \mathbf{f}^{-1}(\Delta)$, where Δ is the $(m-1)$-dimensional probability simplex.

4. For each $\theta \in S$ check if $\mathbf{f}(\theta)$ is a local maxima (for example by examining the Hessian matrix).

2.2 Maximum likelihood degree

An important question for computational statistics is this:

What happens to the estimate $\hat{\theta}$ when we vary u?

Definition 2.5 (Algebraic model). We say a model $\mathbf{f}$ is **algebraic** if all the f_i are polynomials or rational functions.

Proposition 2.6 ($\hat{\theta}$ is an algebraic function of the data u). The maximum likelihood estimate $\hat{\theta}$ is an algebraic function of the data u if $\mathbf{f}$ is algebraic. That is, $\hat{\theta}_i$ is a zero of a polynomial of the following form

$$a_r(u)x^r + a_{r-1}(u)x^{r-1} + ... + a_i(u)x + a_0(u),$$

where each $a_i \in \mathbb{Q}[u]$.

Without loss of generality, we can assume that the polynomial is an *irreducible element* of $\mathbb{Q}[u, x]$. This means that **the discriminant is a nonzero polynomial in** $\mathbb{Q}[u]$**.**

Definition 2.7 (Generic). We say that $u \in \mathbb{R}^m$ is **generic** if no discriminant vanishes at u for all $i = 1, ..., m$. Hence, there exist no multiple roots in any field extension (see Wikipedia, section *Zero discriminant*). The generic vectors are dense in $\mathbb{R}^m$.

Definition 2.8 (Maximum likelihood degree). The **maximum likelihood degree** or **ML degree** of an algebraic statistical model is the *number of solutions to the critical equations* for generic data point $u \in \mathbb{R}^m$.