WOGAS University of Warwick

Algebraic Statistics: a short review

Henry Wynn
(Kei Kobayashi, Giovanni Pistone, Eva Riccomagno)

London School of Economics
h.wynn@lse.ac.uk

6 April, 2010

Algebraic Statistical models

The aim is to describe how algebraic methods can help in defining and analysing statistical models.

1. x: a control (input) variable
2. θ a basic parameter
3. η a parameter which may (often) be considered as depending on x (eg a mean)

Definition

An algebraic statistical model is a statement that (η, x, θ) lie on an affine algebraic variety:

$$h(\eta, x, \theta) = 0,$$

together with a statement that the joint distribution of outputs $Y_1, \ldots Y_n$ depends on

$$\theta,\ (x_i, \eta_i),\ i = 1, \ldots, n$$

Explicit models

- Regression: if η is a mean:

$$\eta = f(x, \theta)$$

Then, if g is polynomial we can write

$$h = \eta - f(x, \theta) = 0$$

- Variance components: may need a double index $\gamma_{ij} = cov(Y_i, Y_j))$. But we can have a variety for the covariances, eg $(\Gamma^{-1})_{ij} = 0$ in conditional independence models.
- Loglinear models:

$$p_i = \exp(x^T \theta) = \exp\{\sum x_i \theta_i\}$$

It appears as if exp kills the algebraic forms but we can write

$$t_i = \exp(\theta_i)$$

giving the *power product* representation

$$p_i = \prod t_i^{x_i}$$

Implicit models: the use of elimination

Eliminate θ (typically) to get an implicit relationship between x_i and η_i.

- Regression. $\eta = X\theta$

$$\Leftrightarrow K^T \eta = 0$$

where $K = \{k_{ij}\}$ spans the kernel of X: $X^T K = 0$ eg

$$\eta_i = \theta_0 + \theta_1 x_i, \quad x = 0, 1, 2$$

$$\eta_1 - 2\theta_2 + \eta_3 = 0$$

- Toric ideals. $[\log p] = X\theta$

$$\Leftrightarrow K^T [\log p] = 0 \Leftrightarrow \sum_i k_{ij} \log p_i = 0, \Leftrightarrow \prod_i p_i^{k_{ij}} = 1$$

$$\Leftrightarrow \prod_i p_i^{k_{ij}^+} - \prod_i p_i^{k_{ij}^-} = 0, \; j = 1, \ldots, n-p$$

Ideals of points and design of experiments

1. A design is a finite set of distinct points, D, in R^d (Q^d) and can be expressed as the solution of a set of equations and can be thought of as a zero dimensional variety. The set of all polynomials with zeros on a D is the ideal, $I(D)$.
2. There is a Gröbner basis $\{g_j(x)\}$ for $I(D)$ for a given monomial ordering: $I(D) = < g_1(x), \ldots, g_m(x) >$.
3. The quotient ring
$$K[x_1, \ldots, x_k]/I(D)$$
of the ring of polynomials $K[x_1, \ldots, x_k]$ in $x_1, \ldots, x_k$ forms is a vector space spanned by a special set of monomials: $x^\alpha, \alpha \in L$. These are all the monomials not divisible by the leading terms of the G-basis and $|L| = |D|$.

6. The set of multi-indices L has the "order ideal" property: $\alpha \in L$ implies $\beta \in L$ for any $0 \leq \beta \leq \alpha$. For example, if $x_1^2 x_2$ in the model so is $1, x_1, x_2, x_1 x_2$.
7. Any function $y(x)$ on D has a unique polynomial interpolator given by

$$f(x) = \sum_{\alpha \in L} \theta_\alpha x^\alpha$$

such that $y(x) = f(x),\ x \in D$.
8. The X-matrix is $n \times n$, has rank n and has rows indexed by the design points and columns indexed by the basis:

$$X = \{x^\alpha\}_{x \in D, \alpha \in L}$$

Message: we can always construct a polynomial interpolator (saturated regression model) over a finite set of design points

.

One slide on multi-dimensional quadrature

Take a measure ξ, a monomial term ordering: $\prec$: and a design D and construct L. For any $p(x)$:

$$p(x) = \sum_i s_i(x) g_i(x) + \sum_{\alpha \in L} \theta_\alpha x^\alpha$$

We can rewrite $r(x)$ in terms of indicator functions:

$$r(x) = \sum_{z \in D} p(z) L_z(x), \text{ where } L_z(x) = \delta_{x,z}, \quad x, z \in D$$

If $\mathrm{E}_\xi(\sum s_i(x) g_i(x)) = 0$, we have quadrature:

$$\mathrm{E}_\xi(p(x)) = \mathrm{E}_\xi(r(x)) = \sum_{z \in D} p(z) \mathrm{E}(L_z(x)) = \sum_{z \in D} w_z p(z)$$

Choose $D : \{x : h_\alpha(x) = 0, \quad \alpha \in M\}$, where the $h_\alpha(x)$ are orthogonal polynomials wrt ξ in $\prec$ order?

Discrete probability models

Assume that we have a discrete probability distribution with *support* at the design points:

$$p(x) > 0, \quad x \in D$$

The we can interpolate

$$\log p(x) = \sum_{\alpha \in L} \theta_\alpha x^\alpha,$$

giving a saturated models in the exponential family:

$$p(x) = \exp\left(\sum_{\alpha \in L} \theta_\alpha x^\alpha\right) p_0(x)$$

More generally:

$$p(x) = \exp\left(\sum_{\alpha \in L_0} \theta_\alpha x^\alpha - \phi(\theta)\right),$$

where L_0 is $L \setminus \{0\}$ and θ excludes $\theta_{\{0\}}$

Five parametrizations

At the heart of the algebraic statistics of discrete distributions is the interplay between five important parameterizations

- θ_α
- $p(x)$
- $t_\alpha = \exp(\theta_\alpha)$
- Moments $\mu_\alpha = \mathrm{E}(X^\alpha)$
- Cumulants κ_α.

In the saturated case we can write $\alpha \in L$. But note importantly: L in general depends on the monomial order we use.

Relations

The relations between p, t, μ, κ are all algebraic

- p to μ is linear: $\mu = X^T p$
- μ to κ are the "exp-log" formula.
  - Start with the "square free" moments: $\alpha : \alpha_i = 0, 1$

$$\mu_\alpha = \sum_{\sigma \in \mathcal{L}} \prod_{\tau \in \sigma} \kappa_\tau$$

$$\sigma = [\beta_1 | \beta_2 | \ldots]$$

  - "Dummy" to get higher order moments, eg:

$$\mu_{2,0} = \mathrm{E}(X_1 X_1' X_2), \; X_1' \equiv X_1$$

Moment and cumulant aliasing

$\mu_\beta, \kappa_\beta,\ \beta \notin L$ can be expressed in terms of $\mu_\alpha, \kappa_\alpha,\ \alpha \in L$

$$x^\beta = \mathrm{NF}\left(x^\beta\right) = \sum_{\alpha \in L} c_{\alpha,\beta}\, x^\alpha,\ x \in D$$

Taking expectations:

$$\mu_\beta = \mathrm{E}\left(x^\beta\right) = \sum_{\alpha \in L} c_{\alpha,\beta}\, \mu_\alpha$$

For cumulants:

$$\kappa_\beta \longrightarrow \mu_\beta \longrightarrow \mu_\alpha \longrightarrow \kappa_\alpha$$

Submodels 1: sufficient statistics and MLE

Take a subset L' of monomials: $f(x) = \sum_{\alpha \in L' \subset L} \theta_\alpha x^\alpha$

For the probability models we get exponential families:

$$p(x) = \exp\left(\sum_{\alpha \in L' \subset L} \theta_\alpha x^\alpha \right)$$

$$p(x) = \exp\left(\sum_{\alpha \in L'_0 \subset L_0} \theta_\alpha x^\alpha - \phi(\theta) \right)$$

Then, under the usual iid assumptions the *sufficient statistics* are:

$$T_\alpha = \sum_{sample} x^\alpha, \quad \alpha \in L'$$

and the likelihood equations are

$$X^T m_\alpha = X^T \mu_\alpha, \quad \alpha \in L'$$

Submodels 2: Kernels and toric ideals

The interplay between the kernel K, toric ideals, Markov bases for submodels has been well developed

- Graphical models are well represented by particular choices of the sub model: eg conditional independence

$$p(x) = \exp(\theta_{000} + \theta_{100}x_1 + \theta_{010}x_2 + \theta_{001}x_3 + \theta_{101}x_1x_3 + \theta_{011}x_2x_3)$$

- Decomposable graphical models $\Leftrightarrow$ square free quadratic toric ideals $\Leftrightarrow$ closed form MLEs.
- Sufficient statistics are (generalised) margins. MCMC methods simulate from tables with given margins to give exact conditional tests.
- Kernel ideals plus "saturation" gives G-bases and Markov bases
- Live research to taylor Markov bases to the problem at hand
- Alternatives to MCMC: linear/integer programming, importance sampling, lattice point enumeration (latte)

Boundary models

How to obtain boundary models in which certain are $p(x) = 0$ limits of the $p(x) > 0$?

$$p(x) = \exp(\sum_{\alpha \in L'} \theta_\alpha x^\alpha) = \exp(\sum_{z \in D} \phi_z L_z(x)),$$

where $\phi_z = \log p_z$ and $-\infty < \phi_z \leq 1$.

- Problem 1: it may be that ϕ does not cover all extremal rays of the *recession cone* (see LP).
- Solution 1: extend X to $[X : \tilde{X}]$ to include all extremal rays.
- Solution 2: Find where solutions to $K^T \phi = 0$ cut the coordinate hyperplanes.
- Problem 2: We also want to have integer solutions in order to be able to extend the $t_\alpha = \exp \theta_\alpha$, power product parametrization.
- Solution A: Hilbert basis
- Solution B (better): Only the integer generators of the extremal rays

Example: 2×2 table on $[0,1]^2$

Binary independence model

:

$$p(x) = \exp(\theta_{00} + \theta_{10}x_1 + \theta_{20}x_2),\; X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

Extremal rays: $\begin{array}{|c|c|} \hline p_{01} & p_{11} \\ \hline p_{00} & p_{10} \\ \hline \end{array} = \begin{array}{|c|c|} \hline 1 & 0 \\ \hline 1 & 0 \\ \hline \end{array} \begin{array}{|c|c|} \hline 0 & 1 \\ \hline 0 & 1 \\ \hline \end{array} \begin{array}{|c|c|} \hline 0 & 0 \\ \hline 1 & 1 \\ \hline \end{array} \begin{array}{|c|c|} \hline 1 & 1 \\ \hline 0 & 0 \\ \hline \end{array}$

$$[X : \tilde{X}] = \begin{pmatrix} 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 \end{pmatrix}, \quad \begin{array}{lclcl} p_{00} & = & t_0 & & t_0 t_3 t_4 \\ p_{10} & = & t_0 t_1 & \rightarrow & t_0 t_1 t_4 \\ p_{01} & = & t_0 t_2 & & t_0 t_2 t_3 \\ p_{11} & = & t_{t1} t_2 & & t_0 t_1 t_2 \end{array}$$

Classical indicator notation not so bad!: $\log p_{ij} = \mu + \alpha_i + \beta_j$

Curved exponential families

How far can the algebraic methods be used in information geometry and asymptotics?

- MLE, U-statistics, Fisher information,
- First, second, ... order efficiency
- Test statistics,...
- Diff geometry entities, curvature, connections etc

A beginning: second order efficiency

$$p(x,\theta) = p(\theta^T x - \phi(\theta))p_0(x)$$

dim $\theta = n$. Want to have a submodel parametrized by u (dim $u = p < n$: $\theta(u)$. Consider x to be the sufficient statistic. Start with a 1-1 function into (u,v) space :

$$\theta = F(u,v)$$

.

- Model: $\theta(u) = F(u,0)$
- Estimation: take the MLE of θ under the full model: $\hat{\theta}$
- Invert: find $(\hat{u},\hat{v})$ so that

$$\hat{\theta} = F(\hat{u},\hat{v})$$

- Consider the class $\tilde{\theta} = F(\hat{u},0)$
- In Amari there are conditions for first and second order efficiency. Try to "resolve" these conditions algebraically.

Using η can be easier

$\eta = \mathrm{E}(x) = \bigtriangledown\phi(\theta)$

Construction via η. Note we have $\eta(u)$, for the model.

$$\eta_i(u,v) = \eta_i(u) + \sum_j f_j(u,v)v_j,$$

- Finding u. Start with *explicit* algebraic curved exponential family or *implicit* variety for θ and eliminate.
- Find $f_j(u,v)$
- First and second order efficiency conditions induce conditions on the $f_j(u,v)$.
- Theorem:

$$\eta(u,v) = \eta(u) + \sum_j Q_j(u,v)z_j,$$

  Where $\{z_j\}$ is a basis for the kernel of $\eta(u)$ wrt Fisher metric.

Conclusions

- More on basics: relationship between the parametrizations
- Beyond graphical models: eg marginal models, the whole lattice.
- Fast algorithms for MCMC and alternatives
- Model building
- Link to differential geometry
- More algebra: monomial ideals, lattices, toric, ...
- Computational geometry/topology: eg persistent homology.

References

Algebraic Statistics. Pistone, Riccomagno, W. Chapman and Hall/CRC. (2001)

Algebraic Statistics for Computational Biology. Pacher and Sturmfels (eds). Cambridge. (2005).

Lectures on Algebraic Statistics. Drton, Sturmfels and Sullivant. Birkhuser. (2009).

Algebraic and Geometric Methods in Statistics. Gibilisco, Riccomagno, Rogantin, Wynn. (eds). Cambridge. (2010).

Algebraic and Geometric Methods in Statistics. Viana and W. (eds). (2010).