Exercises on Algebraic Statistics

Hugo Maruri-Aguilar[*†], Eva Riccomagno[‡], Henry Wynn[*]

June 26, 2009

Contents

[*]Department of Statistics, London School of Economics, London WC2A 2AE, UK
[†]Email address: H.Maruri-Aguilar@lse.ac.uk
[‡]Dipartimento di Matematica, Università degli Studi di Genova, Genova 16146, Italy

1 Exercises

1. (Ideal intersection) Compute the intersection of the ideals generated by the two lines $y = 1 + 3x$ and $y = 2 - x$.

2. (Gaussian elimination) Parametrize all the solutions of the linear equations

$$\begin{aligned} x + 2y - 2z + w &= -1 \\ x + y + z - w &= 2 \end{aligned} \tag{1}$$

3. (Linear parametrization) Find the equation of the plane in $\mathbb{R}^3$ parametrized by

$$\begin{aligned} x &= 1 + u - v \\ y &= u + 2v \\ z &= -1 - u + v \end{aligned} \tag{2}$$

4. (Nonlinear parametrization) Descartes' *folium* has the following parametric representation

$$(x(t), y(t)) = \left(\frac{3at}{1+t^3}, \frac{3at^2}{1+t^3} \right) \tag{3}$$

Find an implicit equation in terms of the coordinates x, y and of a. Do the same for the *Cissoid* of Diocles

$$(x(t), y(t)) = \left(\frac{2at^2}{1+t^2}, \frac{2at^3}{1+t^2} \right) \tag{4}$$

5. (Nonlinear parametrization) Consider the parametric representation

$$x(t) = \frac{t}{1+t}, y(t) = 1 - \frac{1}{t^2}.$$

Find the equation determined by the above parametric representation.

6. (Nonlinear solving) Consider the system of equations

$$\begin{aligned} x^2 + 2y^2 &= 3 \\ x^2 + xy + y^2 &= 3 \end{aligned} \tag{5}$$

If I is the ideal generated by these equations, find bases of $I \cap \mathbb{R}[x]$ and $I \cap \mathbb{R}[y]$. Also find all the solutions to the equations.

7. (Nonlinear optimization) Find the extreme values of the function $F(x, y, z) = x^3 + y^3 + z^3$ over the sphere $x^2 + y^2 + z^2 = 4$.

8. (Nonlinear optimization) Find the extreme values of the function $F(x, y, z) = x^3 + y^3 + z^3$ over the sphere $x^2 + y^2 + z^2 = 4$ and the plane $x + y + z = 1$.

9. (Nonlinear optimization) Find extreme values for x^2y subject to the restriction $x^2 + y^2 = 3$.

10. (Nonlinear optimization) Find the maximum and minimum values of

$$f(x_1, x_2) = \int_{x_1}^{x_2} \frac{1}{1 + t^4} dt \tag{6}$$

over the region determined by $x_1^2 x_2^2 = 1$.

11. (Design of Experiments) Consider the following Latin Square design

A	B	C
C	A	B
B	C	A

(7)

Which model can you identify with it?

12. (Design of Experiments) Consider the design

$$D = \{(-1, -1), (-1, 1), (1, -1), (1, 1)\},$$

(a) Construct the design ideal $I(D)$.

(b) Identify a model for the response values observed at design points.

(c) Find the Hilbert function of the design ideal. How does the Hilbert function relates to question 12b?

13. (Design of Experiments) Consider D to be a 2^3 design with levels ± 1 and its design ideal. Now let F be the fraction defined by the generator $x_1 x_2 x_3 = 1$. Construct $I(F)$ by the following two methods:

(a) by directly adding the generator to the generators of $I(D)$.

(b) by removing those points not satisfying the generator from $I(D)$ (using the semicolon operator :).

(c) Verify that both methods give the same result.

14. (Design of Experiments) Using different term orders, build a list of different polynomial models identified by each of the designs listed below. Play with different term orders to identify as many models as possible.

    (a) $D_1 = \{(3,1),(2,3),(0,2),(1,0)\}$

    (b) $D_2 = \{(0,0),(1,0),(0,1),(-1,1)\}$.

    (c) D_3: add the point $(0,0)$ to the design in Problem 12.

    When you have your list of models then for every design, make a plot of the exponents of every model in integer grids in $\mathbb{Z}^2$. Identify unifying features for all the models identified by the designs, plotted in this form. What do you conclude?

15. (Maximum likelihood) Consider a multinomial model with four cathegories

$$\Pr(x_1,x_2,x_3,x_4;\theta_1,\theta_2) = \frac{(x_1+x_2+x_3+x_4)!}{x_1!x_2!x_3!x_4!}\prod_{i=1}^{4} p_i^{x_i}, \tag{8}$$

    whose probabilities depend on parameters θ_1,θ_2 as follows:
    $p_1(\theta_1,\theta_2)=\theta_1$, $p_2(\theta_1,\theta_2)=\theta_2$, $p_3(\theta_1,\theta_2)=(\theta_1-1)^2+(\theta_2-1)^2-2$ and $p_4(\theta_1,\theta_2)=(\theta_1+1)^2+2(\theta_2-2)^2-9$.
    For the observed data $(x_1,x_2,x_3,x_4)=(2,1,9,1)$, compute the ML estimators for the parameters. Is it surprising what you get?

16. (Contingency table) The results of a Phase III randomized trial were recorded in a 2×3 contingency table. The figures are antiemetic response data after two days.

12	3	7
3	7	12

    (a) Compute the maximum likelihood estimators for p_{ij}.

    (b) Redo the ML computations for the log-linear hierarchical model

$$p_{r,c} = \exp\left(\sum_{(i,j)<(r,c)} \theta_{i,j}\right) \tag{9}$$

17. (Conditional independence) For binary random variables X_1, X_2, X_3, find the basis of the conditional independence ideal defined by $X_1 \perp\!\!\!\perp X_2 | X_3$.

18. (Moment aliasing) Consider a bivariate distribution for (X_1, X_2) associated with the table

1	p_{01}	p_{11}
0	p_{00}	p_{10}
	0	1

Under the independence model $p_{00}p_{11} = p_{10}p_{01}$, find implicit conditions in terms of the moments m_α for $\alpha \in L = \{(0,0),(1,0),(0,1),(1,1)\}$. Recall that for a d-variate random variable X with finite discrete support in $x_1, \ldots, x_n \in \mathbb{R}^d$, the α-th moment of X is

$$m_\alpha = E(X^\alpha) = \sum_{i=1}^{n} x_i^\alpha \Pr(X = x_i),$$

where $\alpha \in \mathbb{Z}^d_{\geq 0}$.

19. (Moment aliasing) Consider a bivariate distribution with discrete support according to the following table.

Point	$(1,0)$	$(1,1)$	$(0,1)$	$(-1,0)$	$(-1,-1)$	$(0,-1)$
Mass	p_1	p_2	p_3	p_4	p_5	p_6

Consider the usual independence model (as in Problem 18) and find implicit conditions in terms of the moments $m_\alpha = E(X^\alpha)$ for $\alpha \in L = \{(0,0), (1,0), (0,1), (1,1), (0,2), (1,2)\}$.

20. (Contingency table) Consider the probabilities associated to a 2×2 contingency table

$p_{0,0}$	$p_{1,0}$
$p_{0,1}$	$p_{1,1}$

and the ideals $I = \langle p_{0,1}p_{1,0} - p_{0,0}p_{1,1} \rangle$ and $J = \langle p_{i,j} - p_{i,\cdot}p_{\cdot,j} \rangle$. Explore the ideal relationship between the independence condition of J and the restriction in rank imposed by I.

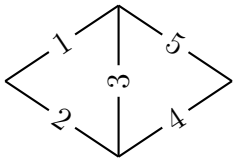
21. (Monomial ideal) Consider the following ideals $I = \langle x_1^3 x_2, x_1 x_2^2, x_2^4 \rangle$ and $J = \langle x_1^3 x_2, x_1 x_2^2, x_2^4, x_1^3 x_2^4 \rangle$.

    (a) Compute the Gröbner basis for each of them and determine the relation between I and J.

    (b) Compute the Hilbert function and Hilbert series for the quotient ring induced by I. What is the interpretation of Hilbert function and series? Which is the Hilbert polynomial? Which is the index of regularity?

22. (Monomial ideal) Consider the following design $D = \{(0,0), (1,0), (0,1), (1,1), (2,0), (0,2)\}$. Compute the Hilbert function and Hilbert series of $\mathbb{R}[x]/I(D)$. Determine the Hilbert polynomial and index of regularity.

23. (Monomial ideal) Consider the following five element network

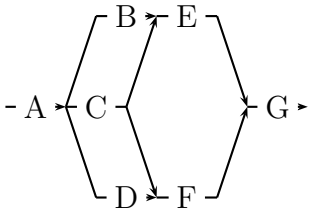


    The network functions if there is a connected path between the left and right extremes.

    (a) Construct the ideal representing the failure of the network.

    (b) Use multigraded Hilbert series of $\mathbb{R}[x]/I$ to identify Bonferroni bounds for the reliability of the graph. Explain the relation with Hilbert series.

24. (Monomial ideal) A complex telecommunication network system has the following reliability block diagram.



With the aid of a minimal cut set, generate the corresponding ideal. Then compute the multigraded Hilbert series of $\mathbb{R}[x]/I$ to study Bonferroni bounds for the reliability of the network.

25. (Monomial ideal) Using the same ideal of Problem 22, compute multigraded Hilbert series for $\mathbb{R}[x]/I$ using different weight matrices. Give an interpretation of your results.

26. (Markov basis) Consider a 3×3 contingency table.

    (a) Using the kernel method of the corresponding design matrix, find a Markov basis for the independence ideal of the table. This basis represents moves that retrieve the same table marginals.

    (b) With the data

$$\begin{bmatrix} 2 & 3 & 3 \\ 3 & 3 & 5 \\ 4 & 5 & 6 \end{bmatrix} \tag{10}$$

    and using Metropolis-Hastings algorithm, perform a χ^2 test for the independence model $H_0 : p_{i,j} = p_{i,\cdot}p_{\cdot,j}$.

    (c) Compare the results with those obtained with generalised linear models.

    (d) Repeat the computations for the following tables and compare your results.

$$\begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{bmatrix} \tag{11}$$

$$\begin{bmatrix} 1 & 6 & 4 \\ 2 & 5 & 15 \\ 2 & 12 & 30 \end{bmatrix} \tag{12}$$

$$\begin{bmatrix} 3 & 8 & 1 \\ 3 & 0 & 12 \\ 7 & 14 & 7 \end{bmatrix} \tag{13}$$

$$\begin{bmatrix} 20 & 30 & 40 \\ 29 & 38 & 52 \\ 51 & 64 & 70 \end{bmatrix} \tag{14}$$

27. (Design of Experiments) Let $\mathcal{D}$ be a binary full factorial design in k factors with level coding 0 and 1. Compute

    (a) Ideal($\mathcal{D}$),

    (b) the support for a hierarchical saturated polynomial regression model (call it SM_{01}),

    (c) its algebraic fan and statistical fan,

    (d) i. NormalForm($x_1^2x_2$, Ideal($\mathcal{D}$)) and NormalForm($x_1^2x_2^2$, Ideal($\mathcal{D}$)).

    ii. Give a formula for NormalForm(x^α, Ideal($\mathcal{D}$)) for $\alpha \in \mathbb{Z}^k_{\geq 0}$.

    (e) The "alias" table for the design in this example is given by

$$\begin{aligned} X_1^2 &= X_1 \\ &\vdots \\ X_k^2 &= X_k \end{aligned}$$

    where capital letters stand for factors/random variables, while lower case letters are the algebraic indeterminate. Identify the connection with Ideal($\mathcal{D}$). How could this connection be generalised to any design?

28. (Design of Experiments) Apply the change of coordinates $y_i = 2x_i - 1$ for all $i = 1, \ldots, k$ to $\mathcal{D}$ of Problem 27.

    (a) Identify the new design $\mathcal{D}'$. Give a practical example of when trasformations of the type $ax_i + b$ might be useful.

    (b) Compute a Gröbner basis of Ideal($\mathcal{D}'$).

    (c) Determine the standard basis of $\mathbb{Q}[x_1, \ldots, x_k]/\operatorname{Ideal}(\mathcal{D}')$ and call it $\mathrm{SM}_{\pm 1}$.

    (d) Determine the alias table/confounding relationships and compare with the results in Part I.

    (e) Compute NormalForm(x^α, Ideal($\mathcal{D}'$)) for all $\alpha \in \mathbb{Z}^k_{\geq 0}$.

    (f) Consider $f(x) = \sum_{x^\alpha \in \mathrm{SM}_{01}} \theta_\alpha x^\alpha$ and $f(x) = \sum_{x^\alpha \in \mathrm{SM}_{\pm 1}} \psi_\alpha x^\alpha$.

        i. Why could it be relevant in statistics to have different parametrization of the same model?

        ii. Show that there exists an invertible matrix A such that $[\psi_\beta]_\beta = A[\psi_\alpha]_\alpha$ holds.

29. (Design of Experiments) Consider the two following ideals $\text{Ideal}(\mathcal{D}) = \langle a^2-1, b^2-1, c^2-1\rangle$ and $\text{Ideal}(\mathcal{F}) = \langle a^2-1, b^2-1, c^2-1, abc-1\rangle$. The full-factorial design $\mathcal{D}$ with levels ± 1 is indicated as 2^3; while $\mathcal{F}$ is called a 2^{3-1} design (half fraction of $\mathcal{D}$).

    (a) Show that they are design ideals and that $\mathcal{D} \supset \mathcal{F}$.

    (b) Compute their algebraic fans and statistical fans.

    (c) How would you define their alias tables?

    (d) Compute the indicator function of $\mathcal{F} \subset \mathcal{D}$.

    (e) Apply the transformation of variables $2x-1$ for $x = a, b, c$. Answer to (b)-(d) above.

30. (Design of Experiments) The following data set comes from an experiment involving flour for making bread [Næs et al., 1998]. The mixture components are three types of flour, and two process variables were involved. The output is leaf volume.

Flour type			Responses		
Tjalve	Folke	Hard-Red-Spring			
x_1	x_2	x_3	y_1	y_2	y_3
1/4	3/4	0	378.89	396.67	392.22
1/2	1/2	0	388.89	423.33	416.11
3/4	1/4	0	426.11	483.33	389.44
1	0	0	386.11	459.11	423.33
1/4	1/2	1/4	417.78	437.22	444.56
1/2	1/4	1/4	389.44	447.22	415.00
3/4	0	1/4	448.33	459.44	455.56
1/4	1/4	1/2	413.89	485.56	462.22
1/2	0	1/2	415.56	514.44	437.78
1/4	0	3/4	432.78	498.33	517.22

    (a) Analyse the subset of the mixture component of the bread experiment for the following design points and output values.

    (b) Analyse the mixture component of the bread experiment.

    (c) Analyse the full bread experiment including process variables.

31. (Design of Experiments) [Holliday et al., 1999] describe a designed experiment in four factors (R, L, A and E) which was stopped after only 23 runs had been collected. The intended experiment was effectively a 3^4 in which there is no interest in the interactions between two of the factors (A and E). Figure 1 shows the designed experiment, $\mathcal{D}$, and which runs were collected, $\mathcal{F}$.

    (a) Show that ae is in no standard bases of both $\mathcal{F}$ and $\mathcal{D}$.

    (b) Consider the default term ordering in CoCoA and vary the initial ordering of the factors. How do you expect the standard bases to be?

    (c) Choose a suitable subset of the standard bases to be used as support of a non-saturated statistical model. Non-saturated means that the X matrix is not full rank having more rows (one for each design point) than columns (one for each monomial in the chosen subset of the standard bases).

32. (Design of Experiments) Draw the design $\mathcal{D}$ corresponding to the ideal generated by $x(x^2-1), y(y^2-1), (x-y)(x+y)$.

    (a) Compute the quotient space bases of this ideal for all term orderings.

    (b) Compute the statistical fan of $\mathcal{D}$.

    (c) Study the algebraic fan of the star composite designs in two dimensions. A central composite design has a central point, in $(0,\ldots,0) \in \mathbb{R}^d$, a 2^d full factorial component at levels ± 1, and $2d$ "axial" points, located on each axis at levels ± 2.

33. (Contingency table) Show that two binary random variables are independent if and only if the determinant of the contingency table matrix is zero. Recall that X and Y are independent in $(\Omega, \mathcal{F}, \mathrm{P})$ if $\mathrm{P}(X = x, Y = y) = \mathrm{P}(X = x)\,\mathrm{P}(Y = y)$ for all possible values of x, y.

34. (Probability interpolation) Let $f(a,b,c) = \theta_0 + \theta_1 a + \theta_2 b + \theta_3 c$ be a probability density function on the 2^{3-1} support with levels ± 1 including $(1,1,1)$.

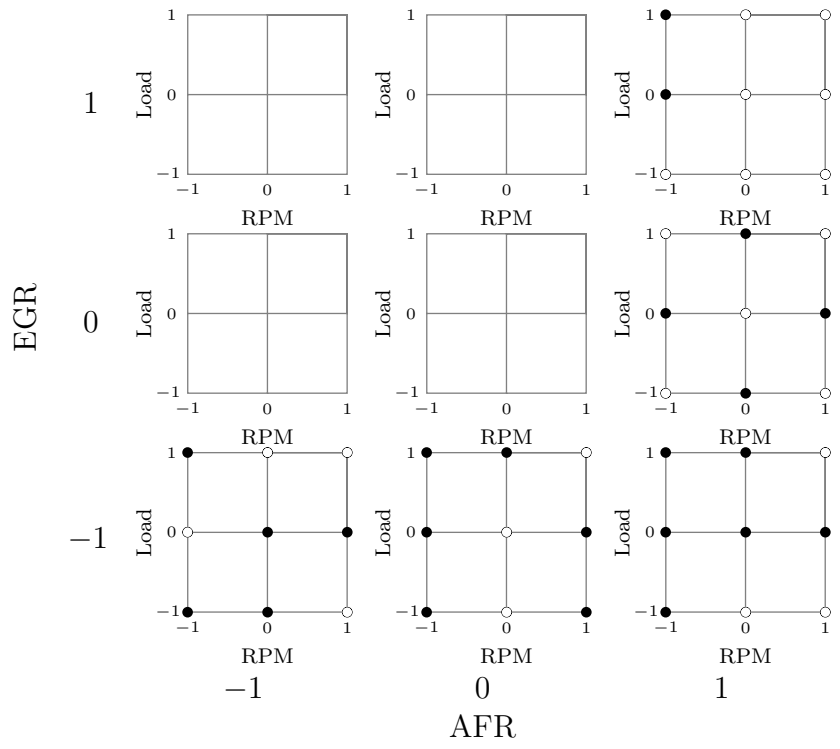


Figure 1: Planned design $\mathcal{D}$ ($\circ, \bullet$) and performed fraction $\mathcal{F}$ ($\bullet$) of Problem 31.

(a) Determine the parameters θ for a uniform distribution on the support.

(b) Determine the parameters θ for a probability P such that the probability of $(1, 1, 1)$ is $1/2$, the mass probability of $(-1, 1, -1)$ is $1/4$ and the other points have equal mass.

35. (Probability interpolation) Consider the random variable $Y = A + B + C$, where A is the random variable taking values ± 1 and corresponding to a (Problem 34 above), likewise for B and C.

(a) Compute the expectation of Y with respect to the uniform distribution. That is $\mathrm{E}_0(Y)$.

(b) Compute the second moment of Y with respect to the uniform distribution. That is $\mathrm{E}_0(Y^2)$.

(c) Eliminate a, b, c from $\{y-(a+b+c), a^2-1, b^2-1, c^2-1, abc-1\}$ to obtain the "image of Y". Interpret your result.

(d) Show that a generic form of the image probability of Y is $p_Y = \theta_0 + \theta_1 Y$.

(e) Consider the P distribution over the 2^{3-1}. Compute the image probability of Y.

(f) Consider the uniform distribution over the 2^{3-1}. Compute the image probability of Y and its polynomial representation.

(g) Compute the density of the image probability with respect to the uniform distribution over D^*.

36. (Probability interpolation) Consider a generic probability distribution P on 2^{3-1}

$(1,1,1)$	$(1,-1,-1)$	$(-1,1,-1)$	$(-1,-1,1)$
$p_{1,1,1}$	$p_{1,-1,-1}$	$p_{-1,1,-1}$	$p_{-1,-1,1}$

(a) Compute its polynomial representation over the quotient space.

(b) Compute the image probability of Y.

(c) Compute the density of the image probability of Y with respect to the uniform distribution.

37. (Conditional independence) Let A, B, C be three random variables. Determine an implicit representation of the statistical model corresponding to the graph A—B—C.

38. (Conditional independence) Let the graph G give an undirected graphical model on the random variables $X_1, \ldots, X_n$. Define $I_{pairwise(G)}$ to be the ideal generated by the quadratic binomials corresponding to the saturated conditional independence statements of the type

$$X_i \perp\!\!\!\perp X_j | \{X_1, \ldots, X_n\} \setminus \{X_i, X_j\} \text{ for all } i \neq j$$

if there is no edge between the nodes i and j in the graph G. Define $I_{global(G)}$ to be the ideal generated by the quadratic binomials corresponding to the saturated conditional independence statements

$$A \perp\!\!\!\perp B | C$$