NUOVI METODI ALGEBRICI NELLA PROGRAMMAZIONE DEGLI ESPERIMENTI

New Algebraic Methods in Design of Experiments

GIOVANNI PISTONE
Politecnico di Torino
Dipartimento di Matematica

1. Strutture matematiche e e metodi statistici

Anche se molti statistici preferiscono concepire la propria Arte come ben distinta dalla Matematica Applicata, è un fatto riconosciuto che l'adozione di strutture matematiche complesse ha una grande influenza sulla metodologia statistica. Forse l'esempio storico più efficace di questa influenza è l'adozione del linguaggio geometrico vettoriale per i modelli di regressione, promosso da vari autori, ad esempio (Rao 1965).

Tale influenza opera su due livelli, tra di loro collegati: l'adozione di metodi matematici particolari è normalmente motivata dalla esigenza di dotarsi di strumenti di elaborazione e calcolo potenti; ma il linguaggio e le strutture matematiche adottate hanno anche una profonda influenza concettuale, evidenziando e sottolineando relazioni e derivazioni che, senza quel linguaggio e quel patrimonio di idee, non sarebbero stati visibili.

Un ulteriore esempio di questa dinamica è l'introduzione di strutture prese dalla teoria dei gruppi e in generale dall'algebra per il trattamento di concetti legati all'invarianza e alla simmetria, in particolare con riferimento alle applicazioni alla teoria delle frazioni regolari di piani fattoriali. In quest'ultima applicazione, più vicina al nostro soggetto, rileviamo presentazioni più caute, come per esempio nel classico (Box, Hunter e Hunter 1978), o in (Montgomery 1991), ovvero più esplicitamente matematizzate, come, ad esempio, in (Dey 1985) o in (Collombier 1996). È interessante infine notare che l'introduzione di strutture matematiche in statistica porta sempre con se un parallelo rinascere di interesse dalla parte matematica, e spesso a sviluppi nuovi. Si veda per esempio il caso delle matrici di Hadamard in (Geramita e Seberry 1979).

In questo lavoro, di carattere prevalentemente di rassegna, tentiamo un primo bilancio della applicazione di metodi, che vengono dall'algebra dei polinomi, in particolare della nozione di base di Gröbner di un ideale di un anello di polinomi, alla metodologia dei piani di esperimento, e in generale ai concetti legati alla modellazione con spazi di funzioni. L'applicazione in Statistica di tali metodi ha avuto inizio con i lavori (Diaconis e Sturmfels 1993), pubblicato poi in (Diaconis e Sturmfels 1998), e (Pistone e Wynn 1996). Cercheremo di mettere in evidenza sia le potenzialità pratiche, legate alla disponibilità di sistemi di calcolo simbolico, sia la rilevanza concettuale.

Nello sviluppo di questa metodologia hanno collaborato con noi nel corso di questi ultimi anni molte persone, che vorrei qui citare (in ordine alfabetico): Roberto Fontana (CRF), Eva Riccomagno (U. Warwick), Lorenzo Robbiano (U. Genova), Maria Piera Ro-

gantin (U. Genova), Henry Wynn (U. Warwick). Alcuni lavori qui citati e che sono ancora in forma preliminare sono accessibili da Internet [1]. I molti altri contributi, cui non abbiamo partecipato direttamente, sono citati nel corso dell'esposizione. Ringraziamo infine A. Giovagnoli, organizzatrice di questa Sessione Specializzata e lei stessa autrice di importanti contributi alla applicazione di metodi algebrici in Statistica, per l'opportunità offerta di presentare in ambito SIS questi risultati. Le ricerche qui descritte sono state fatte, a partire dal 1994, nell'ambito dei vari progetti di ricerca: in particolare il gruppo Sis coordinato da A. Zanella, il progetto Murst coordinato da U. Magagnoli, il progetto CNR coordinato da G. Leti e E. Regazzini, il progetto Brite-Euram CE^2 coordinato da R. Fontana.

L'organizzazione della esposizione è la seguente: dopo una introduzione alla singola nozione per noi più importante, quella di base di Gröbner, viene presentata l'applicazione alla identificabilità dei modelli polinomiali; seguono considerazioni legata alla effettiva implementabilità dei metodi e una particolare applicazione alla teoria delle frazioni di piani fattoriali; al termine del lavoro indichiamo anche l'esistenza di altre applicazioni (non legate alla programmazione degli esperimenti), e i progetti di ulteriori sviluppi, sopratutto nella direzione della probabilità, della modellazione di sistemi dinamici, della geometria della verosimiglianza.

2. Anelli di polinomi e basi di Gröbner

Presentazioni della teoria algebrica e delle relative metodologie di calcolo simbolico si trovano in molti manuali. In particolare sembrano più adatti allo specifico campo che interessa: (Adams e Loustaunau 1994) e (Cox, Little e O'Shea 1997). Quest'ultimo è forse il più diffuso, fin dalla prima edizione del 1992, e contiene una utile rassegna dei sistemi di calcolo simbolico utilizzabili. Noi in questa rassegna faremo riferimento prevalentemente al sistema commerciale Maple (Char, Geddes, Gonnet, Leong, Monogan e Watt 1991) e al sistema CoCoA (Capani, Niesi e Robbiano 1995), sviluppato presso l'Università di Genova da un gruppo diretto da L. Robbiano.

Esistono alcune esposizioni recenti di queste basi per lo specifico pubblico della statistica. In particolare una esposizione in forma di lezioni (non pubblicate) in (Pistone, Riccomagno e Wynn 1996), quella contenuta nella prima parte della tesi di dottorato di E. Riccomagno (1997). Una nostra monografia in collaborazione con E. Riccomagno e H. Wynn è in corso di preparazione su tutta questa materia.

Se k indica un campo numerico (nelle applicazioni che ci interessano i razionali $\mathbf{Q}$, o i reali $\mathbf{R}$, o i complessi $\mathbf{C}$), si indica con $k[x_1, \ldots, x_m]$ l'anello dei polinomi in m indeterminate $x_1, \ldots, x_m$ e coefficienti in k. Ricordiamo che un *ideale* dell'anello dei polinomi è un insieme di polinomi I che è chiuso per la somma (interna) e per la moltiplicazione per un qualunque polinomio dell'anello. Un ideale è generato da una base di polinomi $\{f_1, \ldots, f_k\}$ se è il più piccolo ideale che contiene quei polinomi, cioè se è costituito da tutti e soli i polinomi g che si scrivono nella forma

$$g = \sum_{i=1}^{k} h_i f_i \quad , \qquad [1]$$

[1] `http://www.polito.it/~pistone`

dove gli h_i sono generici polinomi dell'anello. Si noti che tale scrittura non è unica. Ad ogni base di ideale è associato un sitema di equazioni polinomiali

$$\begin{cases} f_1(x) &= 0 \\ &\vdots \\ f_k(x) &= 0 \end{cases} \quad . \qquad [2]$$

Usiamo la notazione vettoriale $x = (x_1, \ldots, x_m)$. L'insieme delle soluzioni del sistema [2] è detta *varietà* dell'ideale; infatti dipende dall'ideale, e non dalla particolare base usata.

In particolare sono rilevanti per l'applicazione alla programmazione degli esperimenti le cosidette varietà 0-dimensionali, che sono quelle costituite da un numero finito di punti. Ad esempio noi rappresentiamo il piano fattoriale $\{0,1\}^3$ come la varietà dell'ideale che ha per base i polinomi $f_1(x_1, x_2, x_3) = x_1^2 - x_1$, $f_2(x_1, x_2, x_3) = x_2^2 - x_2$, $f_3(x_1, x_2, x_3) = x_3^2 - x_3$. Questo significa che i punti del piano sono tutte e sole le soluzioni del sistema polinomiale

$$\begin{cases} x_1^2 - x_1 &= 0 \\ x_2^2 - x_2 &= 0 \\ x_3^2 - x_3 &= 0 \end{cases} \quad , \quad (x_1, x_2, x_3) \in \mathbf{Q}^3 \quad . \qquad [3]$$

Piani più complessi sono descritti naturalmente da sistemi più complessi; ad esempio posso cancellare il punto $(1, 1, 1)$ aggiungendo l'equazione $f_4(x_1, x_2, x_3) = x_1 x_2 x_3 = 0$ al sistema [3]. In generale, non c'è difficoltà, eventualmente usando un sistema di calcolo simbolico, a scrivere un sistema di equazioni polinomiali con radici assegnate, si veda al Par. 4.

Di tutte le possibili basi di un ideale di polinomi, alcune sono più interessanti. Sono le *basi di Gröbner*, che sono definite in corrispondenza di ogni ordinamento dei monomi, cioè ogni ordinamento totale $<$ dei monomi $x^\alpha = \prod_{i=1}^m x_i^{\alpha_i}$, $\alpha = (a_1, \ldots, \alpha_m)$ multi-esponente, tale che

1. $1 < x^\alpha$ se $\alpha \neq 0$.
2. se $x^\alpha < x^\beta$, allora per ogni γ vale $x^{\alpha+\gamma} < x^{\beta+\gamma}$.

Assegnato un ordinamento dei monomi, è possibile stabilire per ogni polinomio quale è il termine di testa; nel caso unidimensionale questa nozione si riduce a quella di termine di grado massimo. Questa nozione è essenziale per l'algoritmo di divisione.

Definizione di base di Gröbner *Una base $g_1, \ldots, g_k$ è una base di Gröbner per l'ordinamento $<$ se i termini di testa della base sono una base dell'ideale generato dai termini di testa dei membri dell'ideale.*

Questa definizione è molto meno astratta di quanto può sembrare ad un primo approccio, perchè permette di estendere immediatamente il calcolo del resto di un polinomio rispetto ad un ideale. Vale infatti il teorema seguente.

Teorema del resto *Sia I un ideale e sia $g_1, \ldots, g_k$ una sua base di Gröbner per l'ordinamento dei monomi $<$. Allora per ogni polinomio f, esiste un* unico *polinomio r, detto resto, tale che $f - r \in I$ e tale che nessun termine di r sia divisibile per qualche termine di testa dei polinomi della base di Gröbner.*

Si osservi che, essendo $f - r \in I$, ed essendo $g_1, \ldots, g_k$ una base di I, allora si può scrivere

$$f(x) = \sum_{i=1}^{k} q_i(x) g_i(x) + r(x) \quad , \tag{4}$$

mettendo in evidenza la formula della divisione. Ma, al contrario del caso univariato, solo r è unico, perchè uno stesso elemento dell'ideale può avere più rappresentazioni.

È un importante risultato, ottenuto negli anni 80 dalla ricerca in algebra commutativa, che *una base di Gröbner finita esiste sempre, ed è calcolabile con un un algoritmo finito, detto di Buchberger*. Tale algoritmo è in effetti disponibile nella maggior parte dei sistemi di calcolo simbolico e permette di trasformare ogni base finita di un ideale in una base di Gröbner per un ordinamento assegnato.

3. Applicazione alla identificabilità nei modelli polinomiali

Un modello lineare polinomiale è assegnato tramite una sua base, nel senso degli spazi vettoriali, per esempio di monomi, o di altre funzioni polinomiali. Noi consideriamo nel modelli del tipo

$$\mathbf{E}\left[Y_x\right] = \sum_{\alpha \in B} \theta_\alpha x^\alpha \tag{5}$$

Se le osservazioni sono ristrette ad un insieme finito (piano) $\mathcal{D}$, allora sorgono i fenomeni ben noti di confondimento, colinearità . . . In (Pistone e Wynn 1996) viene proposta una particolare impostazione del problema, quando l'insieme B degli esponenti nel modello polinomiale in [5] è *completo* nel senso che, se $\beta \in B$ e $\alpha_i \leq \beta_i$, $i = 1, \ldots, m$, allora $\alpha \in B$. Il piano viene identifivato con le soluzioni di un sistema di equazioni polinomiali, cioè in termini algebrici, con una varietà algebrica 0-dimensionale. Si considera poi l'ideale generato e si osserva che due polinomi del modello sono *confusi* sul piano se la loro differenza appartiene all'ideale del piano. Dato che esiste un algoritmo finito per determinare le basi di Gröbner, ed una volta che la base di Gröbner è determinata è possibile determinare l'appartenenza all'ideale verificando che il resto è nullo (si confronti il precedente teorema del resto), ne segue che il problema del confondimento ha una soluzione algebrica del tutto generale.

Questa soluzione è superiore a quella dell'algebra lineare, consistente nel calcolo del rango di una matrice di piano

$$X = [x^\alpha] \quad , \quad x \in \mathcal{D} \quad , \quad \alpha \in B \quad , \tag{6}$$

perchè il teorema del resto implica anche il seguente enunciato, che caratterizza completamente i modelli polinomiali "di grado minimo", cioè soddisfacenti la condizione che abbiamo assunto su B, identificabili su di un piano.

Teorema della base di monomi *Sia $I(\mathcal{D})$ l'ideale di un piano $\mathcal{D}$ con n punti e m fattori a valori nella chiusura algebrica del campo numerico k, e sia $g_1, \ldots, g_k$ una sua base di Gröbner per un certo ordinamento. Allora il quoziente $k[x_1, \ldots, x_m]/I(\mathcal{D})$, cioè lo spazio vettoriale dei polinomi identificabili sul piano, ha una base di monomi costituita dagli n monomi x^α che* non *sono divisibili per i termini di testa degli elementi della base di Gröbner.*

Bisogna notare che dunque ci possono essere molti modelli diversi completi e identificabili su di un dato piano, tante quante sono le basi di Gröbner con termini di testa diversi. Si osservi anche che la teoria non si occupa della stima dell'errore, e che in questa formulazione non c'è trattamento per le replicazioni.

Nel lavoro (Pistone e Wynn 1996) si fa vedere che la non unicità del sistema di equazioni (la base) che definisce il piano si può mettere a frutto, dandone una opportuna interpretazione statistica che estende i metodi usati per la determinazione di piani fattoriali frazionari a partire da *equazioni di definizione*, da trasformare poi in *equazioni di confondimento*.

Una ulteriore discussione delle problematiche derivanti dalla molteplicità dei modelli identificabili con un determinato piano (il cosidetto *fan* di un piano) sono sviluppate in (Caboara, Pistone, Riccomagno e Wynn 1997). Un approfondimento dal punto di vista dell'algebra commutativa delle problematiche sollevate dalle applicazioni statistiche sono in (Caboara e Robbiano 1997) e (Robbiano 1998).

Anche interessante è il problema inverso, cioè la determinazione di piani con particolari proprietà rispetto alla identificabilità. Il relativo problema di identificabilià è stato affrontato in (Robbiano e Rogantin 1998), oltre che nel già citato (Caboara et al. 1997).

I metodi algebrici si estendono in modo naturale a tutte quelle strutture che sono analoghe ai modelli polinomiali nel senso di essere basate su una classe di funzioni modellanti chiusa per il prodotto. Il caso più importante è quello dei *polinomi di Fourier*, trattato in una serie di lavori frutto di una ricerca condotta presso l'Università di Warwick: (Riccomagno, Schwabe e Wynn 1995), (Caboara e Riccomagno 1997), (Riccomagno 1997).

4. Aspetti computazionali ed esempi

Il problema della determinazione dell'ideale di un assegnato disegno è trattato in (Buchberger e Möller 1982) ed ammette molte altre soluzioni; costituisce la base della applicazione del metodo algebrico. Un altro algoritmo generale ma non molto efficiente è stato usato in (Pistone e Wynn 1996). In generale si constata che il punto critico della applicazione dei metodi algebrici è l'efficienza computazionale. L'uso di algoritmi generali, peraltro disponibili ed affidabili, non permette di affrontare problemi di dimensione veramente significativa a causa della grande complessità computazionale di tali algoritmi generali.

Alcuni algoritmi specifici relativi a particolari problemi di calcolo sono stati sviluppati; ad esempio CoCoA contiene un particolare ordinamento polinomiale, detto **`elim`**, che è specialmente adattato alla soluzione del centrale problema dell'eliminazione. Si sta ora lavorando allo sviluppo di algoritmi specifici del caso di interesse (cioè le varietà di punti), nella speranza che questi permettano di passare a problemi più grandi di quelli affrontabili ora. Una prima raccolta di algoritmi specifici delle applicazioni statistiche è stato sviluppato da M. Caboara nell'ambito del sistema CoCoA ed è disponibile su Internet[2].

Alcuni lavori hanno affrontato le prime applicazioni significative, pur nelle limitazioni menzionate sopra. Il lavoro (Riccomagno e Wynn 1997*a*) presenta una rassegna di applicazioni alla statistica industriale, (Holliday, Riccomagno, Wynn e Pistone 1997) sviluppa completamente lo studio di un caso di sperimentazione industriale con dati mancanti e propone una procedura di identificazione alternativa alla *regressione stepwise*, mentre al-

[2] **`ftp://cocoa.dima.unige.it/cocoa/packages/contrib/caboara/`**

tre ricerche non pubblicate sono in corso presso l'Università di Warwick (H. Wynn) sulle applicazioni alla pianificazione nel caso di esperimenti di mistura.

In campo tecnologico, altre ricerche sono in corso nell'ambito di un Progetto Brite-Euram denominato CE^2, coordinato dal Centro Ricerche Fiat (R. Fontana) [3]. Questo progetto lavora allo sviluppo di emulatori di programmi di simulazione utilizzabili per studi di ottimizzazione del progetto di prodotti industriali. Una delle procedure di ottimizzazione è basata su un metodo non-parametrico di adattamento polinomiale in cui i termini sono scelti sulla base di metodi sia algebrici che non-parametrici. Lo sviluppo condurrà ad un confronto sul campo tra metodi di adattamento più classici e quelli algebrici.

Altre applicazioni, in particolare usando le procedure di eliminazione, si sono avute in campo biologico e chimico. Citiamo i lavori (Bayram, Bennet e Dewar 1993), (Kamenski e Pistone 1996), (Riccomagno e White 1997) in cui il metodo dell'eliminazione viene applicato, con opportuni artifici, alla eliminazione delle variabili non osservabili, dunque non stimabili, di sistemi di interesse chimico-biologico descritti da equazioni differenziali polinomiali. Queste ultime applicazioni, che in parte fanno riferimento ad una metodologia algebrica diversa da quella descritta in questo lavoro, cioè alla struttura di anello differenziale, si veda (Kaplansky 1957) e (Fliess e Glad 1993), saranno descritte in un prossimo lavoro di rassegna (Pistone 1998) che copre un ambito più vasto di quello presente, ma è meno dettagliato.

5. Applicazioni alle frazioni ortogonali

Il problema della ortogonolità di un piano introduce una struttura geometrica di tipo euclideo, che non rientra direttamente nelle strutture che sono controllabili tramite i concetti visti finora, tipicamente *affini*, in particolare privi di ogni concetto metrico. Per un inquadramento nell'ambito della trattazione con prodotti tensoriali, si veda (Raktoe, Hedayat e Federer 1981) e (Gupta e Mukerjee 1989). Questa impostazione, pur essendo adeguata alla descrizione, presenta alcuni problemi, il principale dei quali è il fatto che descrive il piano con una matrice, cioè con una applicazione lineare; questo non è naturale, e infatti non permette di inquadrare bene la problematica della descrizione di frazioni ortogonali di carattere generale, cioè non regolare.

È comunque possibile ricorrere, nel caso di piani fattoriali e delle loro frazioni, alla rappresentazione in spazi ortogonali di contrasti di interazioni, e di mettere in rapporto questa rappresentazione con la rappresentazione del piano tramite una funzione indicatrice polinomiale. Questa descrizione è stato usata in (Fontana, Pistone e Rogantin 1997*a*) e (Fontana, Pistone e Rogantin 1997*b*) per ottenere una descrizione completa delle frazioni di piani fattoriali binari con una assegnata struttura di ortogonalità.

Consideriamo m fattori binari e codifichiamo i livelli come $-1, 1$. Il piano fattoriale è $\mathcal{D} = \{-1, 1\}^m$, e il suo ideale Ideal($\mathcal{D}$) è generato dai polinomi $x_i^2 - x_i$, $i = 1 \ldots m$. L'insieme dei polinomi generatori è una base di Gröbner per ogni ordinamento, e la base dei monomi identificabili coincide l'insieme dei 2^m monomi $C_I(x) = \prod_{i \in i} x_i$, $I \subset \{1 \ldots m\}$.

Si osservi che $C_\emptyset(x) = 1$, se $I = \{i\}$ allora si ottengono gli m *effetti semplici* $C_i(x) = x_i$, e se I ha più elementi si ottengono le classiche *interazioni*. L'insieme delle

[3] `http://www.polito.it/~pistone/projects.html`

possibili risposte $y(x)$ sui punti $x \in \mathcal{D}$ è dunque identificabile con l'insieme dei polinomi del tipo

$$Y(x) = \sum_{I \subset \{1,\ldots,m\}} \alpha_I C_I(x) \quad . \qquad [7]$$

Inoltre, se introduciamo su $\mathcal{D}$ la media uniforme $\mathbf{E}_{\mathcal{D}}[Y] = 2^{-m} \sum_{x \in \mathcal{D}} Y(x)$, allora per ogni $I \neq \emptyset$, $\mathbf{E}_{\mathcal{D}}[C_I] = 0$, cioè in questo caso le C_I sono *contrasti*. Inoltre, essendo $C_I C_J = C_{I \triangle J}$, e dato che $I = J$ se e solo se $I \triangle J = \emptyset$, allora

$$I \neq J \quad \Longrightarrow \quad \mathbf{E}_{\mathcal{D}}[C_I C_J] = \mathbf{E}_{\mathcal{D}}[C_{I \triangle J}] = 0 \quad , \qquad [8]$$

dunque i contrasti sono *ortogonali* tra di loro. Qualche autore chiama per questa ragione la formula [7] *rappresentazione ANOVA* della risposta Y, si veda (Owen 1994).

Consideriamo ora una *frazione* $\mathcal{F}$ di $\mathcal{D}$. Noi non poniamo nessuna restrizione alla nozione di frazione, dunque $\mathcal{F}$ è semplicemente un sottoinsieme di di $\mathcal{D}$. In questa applicazione consideriamo una particolare descrizione dell'ideale della frazione: anzichè assegnare un ordinamento dei monomi (che ora sono 1, gli effetti semplici e le interazioni, consideriamo il polinomio indicatore di $\mathcal{F}$, cioè quell'unico polinomio F in forma ANOVA tale che

$$F(x) = \begin{cases} 1 & \text{se} \quad x \in \mathcal{F} \\ 0 & \text{se} \quad x \notin \mathcal{F} \end{cases} \quad . \qquad [9]$$

Se

$$F(x) = \sum_{I \subset \{1,\ldots,m\}} b_I C_I(x) \quad , \qquad [10]$$

allora dal fatto che una indicatrice deve soddisfare l'equazione $F^2(x) = F(x)$ per ogni $x \in \mathcal{D}$, se ne ricava un sistema del secondo ordine nelle b_I, $I \subset \{1, \ldots, m\}$, che definisce tutte le possibili frazioni di $\mathcal{D}$ in forma algebrica.

Questo è utile, perchè è possibile interpretare sulle $(b_I)_I$, molte proprietà interessanti della frazione, tra cui l'ortogonalità tra effetti e interazioni, la risoluzione e altre. Trattiamo qui brevemente la questione dell'*ortogonalità*.

Diciamo che due effetti semplici o interazioni C_I e C_J sono ortogonali su $\mathcal{F}$ se esse sono ortogonali per la media uniforme su $\mathcal{F}$, definita da

$$\mathbf{E}_{\mathcal{F}}[Y] = 2^{-\sharp\mathcal{F}} \sum_{x \in \mathcal{F}} Y(x) = 2^{m-\sharp\mathcal{F}} \mathbf{E}_{\mathcal{D}}[FY] \quad , \qquad [11]$$

dove $\sharp$ denota la numerosità dell'insieme e F è definita da [9]. Ne segue, introducendo la rappresentazione ANOVA in [10], che un effetto semplice o interazione C_J è un contrasto sulla frazione $\mathcal{F}$ se $0 = \mathbf{E}_{\mathcal{F}}[C_I]$, dunque, usando la [11] e la [10],

$$0 = \mathbf{E}_{\mathcal{D}}[F C_J] = \sum_{I \subset \{1,\ldots,m\}} b_I \mathbf{E}_{\mathcal{D}}[C_{I \triangle J}] = b_J \quad , \qquad [12]$$

e questo impone una condizione su b_J. Analogamente è facile verificare che due effetti semplici o interazioni C_J e C_K sono ortogonali sulla frazione se e solo se $b_{I \triangle J} = 0$. Questo permette di caratterizzare algebricamente tutte le frazioni con una certa struttura di ortogonalità sulla base della funzione F. Rimandiamo per gli altri sviluppi e la discussione della applicabilità pratica ai lavori citati.

L'estensione al caso non binario, usando una rappresentazione complessa, come già in (Fontana 1996) e (Lombardo 1997), è in corso di sviluppo in (Fontana, Pistone e Rogantin 1998). Questo approccio alla pianificazioneè stato anche preso in considerazione da J. Wu (Ann Arbor, Michigan), che ne ha suggerito l'applicazione allo studio delle matrici di Hadamard di ordine 16, sviluppando (Hall 1961) e (Sun e Wu 1993).

6. Altre applicazioni e possibili sviluppi.

6.1. Miscellanea

Le applicazioni nel campo della cinetica chimica e della teoria delle popolazioni in biologia sono già state menzionate. È in corso di sviluppo un progetto, denominato WANDA, basato presso l'Università di Warwick dedicato specificamente allo sviluppo delle metodologie basate sull'algebra differenziale. infatti questo tipo di applicazione ha condotto anche alla considerazione di strutture matematiche più particolari, gli anelli differenziali, si veda in (Riccomagno e White 1997) per un primo esempio di applicazione, che sembrano particolarmente adatte per i modelli dinamici.

Sono state considerate anche possibili applicazioni alla modellista probabilistica e bayesiana, in (Riccomagno e Wynn 1997*b*) e altri lavori non ancora pubblicati. L'applicazione al calcolo dei cumulanti è considerata in (Pistone e Wynn 1997).

6.2. Geometria differenziale della verosimiglianza

Menzioniamo infine il collegamento, ancora da sviluppare completamente, con i metodi geometrici nello studio della verosimiglianza e la cosidetta *Information Geometry*: la letteratura è amplissima, e citiamo in particolare il classico (Amari 1985), il recente libro di Kass e Vos (1997) e l'impostazione non parametrica sviluppata in (Pistone e Sempi 1995), (Pistone e Rogantin 1998), (Gibilisco e Pistone 1997). Il concetto più importante è quello di *modello esponenziale non parametrico*, descritto come la famiglia di densità di probabilità

$$p_u(x) = \exp(u(x) - K_p(u)) \cdot p(x) \qquad [13]$$

sullo spazio campionario $(X, \mathcal{X}, \mu)$, dove p è una densità di riferimento, u è una variabile casuale centrata e esponenzialmente integrabile che funge da parametro infinito-dimensionale, e

$$K_p(u) = \mathbf{E}_p\left(e^u\right) = \mathbf{E}_p\left(\log \frac{p_u}{p}\right) \qquad [14]$$

si interpreta sia come funzionale dei cumulanti di u che come *entropia relativa* di p rispetto a p_u. Nell'ambito di questo *super*-modello, gli altri modelli statistici vengono definiti come sotto-modelli (siano essi parametrici o non-parametrici) e si trova che molte nozioni interessanti, sopratutto asintotiche, ma anche bayesiane, hanno una interpretazione geometrica. Si veda (Barndorff-Nielsen e Cox 1989, Barndorff-Nielsen e Cox 1994) per una impostazione che corrisponde bene ad una interpretazione geometrica.

Si dimostra che il funzionale K_p può essere derivato in senso funzionale infinite volte, e che valgono per questo caso non parametrico proprietà ben note nel caso dei modelli esponenziali parametrici, si veda ad esempio (Barndorff-Nielsen 1978) o (Letac 1992). In particolare, se $q = e^{u-K_p(u)} \cdot p$, allora le derivate direzionali funzionali sono

$$\begin{aligned} DK_p(u)v &= \mathbf{E}_q[v] \quad , & [15] \\ D^2K_p(u)vw &= \mathrm{Cov}_q(v,w) \quad . & [16] \end{aligned}$$

e ed esiste un *funzionale di varianza* V tale che

$$\text{Hess } K_p = V(\text{grad} K_p) \quad . \qquad [17]$$

In (Pistone e Wynn 1997) si esamina il caso in cui la dipendenza delle varianze dal parametro di media sia in effetti algebrica, cioè per un sottomodello esponenziale parametrico ottenuto come modello esponenziale naturale di una distribuzione di riferimento, la dipendenza della varianza dal parametro di media si esprimibile tramite un polinomio F. In questo caso si parla di distribuzioni *finitamente generate*, e si ottiene un analogo delle varietà algebriche. Risulta che la maggior parte delle distribuzioni classiche è di questo tipo, come risulta ad esempio dalla tabella sequente.

Normale λ, σ^2:	$K'' - \sigma^2 = 0$
Poisson λ:	$K'' - K' = 0$
Gamma r, λ:	$rK'' - (K')^2 = 0$
Binomiale r, p:	$rK'' - K'(r - K') = 0$
Binomiale Negativa r, p:	$rK'' - K'(r + K') = 0$
GHS r, λ:	$rK'' - (K')^2 - r^2 = 0$

Inoltre la proprietà di essere finitamente generata è stabile per una serie di operazioni di tipo campionario.

Riferimenti bibliografici

ADAMS, W. W., LOUSTAUNAU, P. (1994), *An Introduction to Gröbner Bases*, Graduate Studies in Mathematics, American Mathematical Society.

AMARI, S. (1985), *Differential-geometrical methods in statistics*, no. 28 *in* "Lecture Notes in Statistics", 2nd printing 1990 corrected edn, Springer-Verlag, New York-Berlin.

BARNDORFF-NIELSEN, O. E. (1978), *Information and Exponential Families in Statistical Theory*, John Wiley & Sons, New York.

BARNDORFF-NIELSEN, O. E., COX, D. R. (1989), *Asymptotic Techniques for use in Statistics*, Chapman & Hall, London.

BARNDORFF-NIELSEN, O. E., COX, D. R. (1994), *Inference and asymptotics*, Chapman & Hall, London.

BAYRAM, M., BENNET, J. P., DEWAR, M. C. (1993), "Using computer algebra to determine rate constants in biochemistry", *Acta Biotheoretica* **41**(1-2), pp. 53–62.

BOX, G. E. P., HUNTER, W. C., HUNTER, J. S. (1978), *Statistics for Experimenters*, John Wiley & Sons, New York.

BUCHBERGER, B., MÖLLER, H. M. (1982), The construction of multivariate polynomials with preassigned zeros, *in* Springer, ed., "Computer algebra (Marseille, 1982)", no. 144 *in* "Lecture Notes in Comput. Sci.", Berlin, pp. 24–31.

CABOARA, M., PISTONE, G., RICCOMAGNO, E., WYNN, H. P. (1997), The fan of an experimental design, Technical report, Department of Statistics, University of Warwick. Submitted to *The Annals of Statistics*.

CABOARA, M., RICCOMAGNO, E. (1997), "An algebraic computational approach to the identifiability of Fourier models", *Journal of Simbolic Computation*. To appear.

CABOARA, M., ROBBIANO, L. (1997), Families of ideals in Statistics, *in* Küchlin, ed., "Proc. ISSAC '97", ACM Press, New York, pp. 404–409.

CAPANI, A., NIESI, G., ROBBIANO, L. (1995), CoCoA*3, a system for doing Computations in Commutative Algebra*. Available via anonymous ftp from `lancelot.dima.unige.it`.

CHAR, B., GEDDES, K., GONNET, G., LEONG, B., MONOGAN, M., WATT, S. (1991), *MAPLE V Library Reference Manual*, Springer-Verlag, New York.

COLLOMBIER, D. (1996), *Plans D' Expérience Factoriels. Construction et propriétés des fractions de plans*, no. 21 *in* "Mathématiques et Applications", Springer, Paris.

COX, D., LITTLE, J., O'SHEA, D. (1997), *Ideal, Varieties, and Algorithms*, 2nd edn, Springer-Verlag, New York. 1st ed. 1992.

DEY, A. (1985), *Orthogonal Fractional Factorial Designs*, John Wiley.

DIACONIS, P., STURMFELS, B. (1993), Algebraic algorithms for sampling from conditional distributions, Preprint SC 93-17, ZIB, Berlin.

DIACONIS, P., STURMFELS, B. (1998), "Algebraic algorithms for sampling from conditional distributions", *The Annals of Statistics* **26**(1), pp. xx.

FLIESS, M., GLAD, S. T. (1993), An algebraic approach to linear and nonlinear control, *in* H. T. Trentelman, J. C. Willems, eds, "Essays on Control: Perspectives in the Theory and its Applications", no. 14 *in* "Progress in Systems and Control Theory", Birkäuser, Boston, pp. 223–267.

FONTANA, R. (1996), An algorithmic approach to doe - algebraic methods, *in* O. Mahrenholtz, K. Marti, R. Mennicken, eds, "ICIAM/GAMM. Issue 3: Applied Statistics and Optimization", Akademie Verlag, pp. 97–100.

FONTANA, R., PISTONE, G., ROGANTIN, M.-P. (1997*a*), "Algebraic analysis and generation of two-levels designs", *Statistica Applicata* **9**(1), pp. 15–29.

FONTANA, R., PISTONE, G., ROGANTIN, M.-P. (1997*b*), Classification of two-level factorial fractions, Preprint 349, Dipartimento di Matematica dell'Università di Genova. Submitted to *Journal of Statistical Planning and Inference* - MS 97-347.

FONTANA, R., PISTONE, G., ROGANTIN, M.-P. (1998), Algebraic description of mixed-level fractional factorial designs, Submitted at the Moda98 Conference.

GERAMITA, A. V., SEBERRY, J. (1979), *Orthogonal Designs, Quadratic Forms and Hadamard matrices*, no. 45 *in* "Lecture Notes in Pure and Applied mathematics", Marcel Dekker, Inc., New York.

GIBILISCO, P., PISTONE, G. (1997), Connections on non-parametric statistical manifolds by Orlicz space geometry, Rapporto Interno 8/97, Politecnico di Torino, Dipartimento di Matematica.

GUPTA, S., MUKERJEE, R. (1989), *A Calculus for Factorial Arrangements*, Vol. 59 of *Lecture Notes in Statistics*, Springer-Verlag, Berlin.

HALL, M. J. (1961), Research summary, Technical Report 36-10, p. 21–26, Jet Propulsion Laboratory, Pasadena, California.

HOLLIDAY, T., RICCOMAGNO, E., WYNN, H. P., PISTONE, G. (1997), "The application of algebraic geometry to the design and analysis of experiments: a case study", *Computational Statistics*. Accepted.

KAMENSKI, PISTONE, G. (1996), Elimination in chemical kinetics, Manuscript.

KAPLANSKY, I. (1957), *An Introduction to Differential Algebra*, Actualités Scientifiques et Industrielles, Hermann, Paris.

KASS, R. E., VOS, P. W. (1997), *Geometrical Foundations of Asymptotic Inference*, Wiley Series in Probability and Statistics, John Wiley, New York.

LETAC, G. (1992), *Lectures on Natural Exponential Families and their Variance Functions*, no. 50 *in* "Monografias de Matemática", Istituto de Matemática Pura e Aplicada, Rio de Janeiro.

LOMBARDO, A. (1997), "Product-array and combined-array for robust design: problems of heteroscedasticity", *Statistica Applicata*. In press.

MONTGOMERY, D. C. (1991), *Design and Analysis of Experiments*, John Wiley & Sons, New York.

OWEN, A. B. (1994), "Lattice sampling revisited: Monte Carlo variance of means over randomized orthogonal arrays", *The Annals of Statistics* **22**(2), pp. 930–945.

PISTONE, G. (1998), Symbolic computations in statistical design theory: An overview, SIMAI — IV Congresso Nazionale, Giardini Naxos (Messina) 1–5 giugno 1998.

PISTONE, G., RICCOMAGNO, E., WYNN, H. P. (1996), Applications of Gröbner bases theory in statistics, 4 lectures at the Department of Statistics, University of Warwick, April-May 1996.

PISTONE, G., ROGANTIN, M.-P. (1998), "The exponential statistical manifold: Mean parameters, orthogonality, and space transformation", *Bernoulli* **4**(1), pp. 1–39. To appear.

PISTONE, G., SEMPI, C. (1995), "An infinite dimensional geometric structure on the space of all the probability measures equivalent to a given one", *The Annals of Statistics* **33**(5), pp. 1543–1561.

PISTONE, G., WYNN, H. P. (1996), "Generalised confounding with Gröbner bases", *Biometrika* **83**(1), pp. 653–666.

PISTONE, G., WYNN, H. P. (1997), Finitely generated cumulants, Technical Report 3/97, Dipartimento di Matematica del Politecnico di Torino. Submitted to *Statistica Sinica*.

RAKTOE, B. L., HEDAYAT, A., FEDERER, W. T. (1981), *Factorial designs*, John Wiley & Sons Inc., New York. Wiley Series in Probability and Mathematical Statistics.

RAO, C. R. (1965), *Linear Statistical Inference and its Applications*, John Wiley, New-York.

RICCOMAGNO, E. M. (1997), Algebraic Geometry in Experimental Design and Related Fields, Phd thesis, Department of Statistics, University of Warwick.

RICCOMAGNO, E., SCHWABE, R., WYNN, H. P. (1995), "Lattice–based optimum design for Fourier regression", *Journal of Statististical Plannning and Inference*.

RICCOMAGNO, E., WHITE, L. J. (1997), "Multi-strain species modeling via differential algebra reduction", *Journal of Symbolic Computation*. In press.

RICCOMAGNO, E., WYNN, H. P. (1997*a*), Computational algebraic geometry in industrial experimental design, *in* C. P. Kitsos, L. Edler, eds, "Proceedings of the International Conference on Industrial Statistics: Aims and Computational Aspects", Physica-Verlag, Athens, pp. 193–197.

RICCOMAGNO, E., WYNN, H. P. (1997*b*), Gröbner bases and factorisation in discrete probability and bayes, Presented at Workshop on Symbolic Computation, September 21-27, 1997 at the Centre de recherches mathématiques of the Université de Montréal (CRM).

ROBBIANO, L. (1998), Gröbner bases and statistics, *in* B. Buchberger, F. Winkler, eds, "Gröbner Bases and Applications (Proc. of the Conf. 33 Years of Gröbner Bases)", Vol. 251 of *London Mathematical Society Lecture Notes*, Cambridge University Press. To appear.

ROBBIANO, L., ROGANTIN, M.-P. (1998), Full factorial designs and distracted fractions, *in* B. Buchberger, F. Winkler, eds, "Gröbner Bases and Applications (Proc. of the Conf. 33 Years of Gröbner Bases)", Vol. 251 of *London Mathematical Society Lecture Notes Series*, Cambridge University Press. To appear.

SUN, X. D., WU, C. F. J. (1993), Statistical properties of Hadamard matrices of order 16, *in* W. Kuo, ed., "Quality through Engineering Design", Elsevier Science Publishers, pp. 169–179.

New Algebraic Methods in Design of Experiments

GIOVANNI PISTONE
Department of Mathematics
Politecnico di Torino

Summary

The present paper presents a review of the applications of commutative algebra methods, expecially Gröbner basis, to the statistical methodology of design of experiments. The main reference is to the joint work with H. Wynn, started in 1995, see (Pistone e Wynn 1996). This direction of research is currently active in a number of directions, including symbolic computation methods in statistics and probability, application to industrial design, modelling and design in chemical and biological applications, fractional factorial design, algebraic statistical manifolds. After a brief description of the basic principle, current developments and ongoing research work, an extensive reference list is included, in the aim to give a fair and consisten representation of the state of current research in this area.