Nær-orthogonale eksperimenter i organiske syntesereaksjoner

KJE-3900

Masteroppgave i organisk kjemi

Geir Simonsen

April, 2010
FAKULTET FOR NATURVITENSKAP OG TEKNOLOGI
Institutt for kjemi
Universitetet i Tromsø

Forord

Jeg er lærer i den videregående skolen. Utdanningen min har jeg fra Universitetet i Tromsø, og jeg har holdt kontakt med Instituttfor kjemi gjennom skolebesøk med mine kjemielever og gjennom enkelte kurs arrangert av instituttet. Dette er alltid positive møter.

Som masterstudent nå opplever jeg stor velvilje og hjelpsomhet både fra instituttledelse, administrasjon og spesielt fra min veileder Rolf Carlson. Dette bekrefter mine tidligere positive erfaringer med Institutt for kjemi i Tromsø.

Det har selvfølgelig vært noe strevsomt å kombinere masterstudiet med full jobb, men Rolf Carlson har vært fantastisk i troen på at dette skulle la seg gjøre. *Design og optimalisering i organiske syntesereaksjoner* var et helt nytt felt for meg, og Rolfs tålmodighet har nok vært satt på prøve noen ganger, uten at han på noen måter har uttrykt det.

Som del av masterstudiet har jeg også fått anledning til å delta på to relevante kurs, et i London (*Medicinal Chemistry*) og et i Roma (*Chemical Development and Scale–Up in the Fine Chemical and Pharmaceutical Industries*).

Jeg vil rette en stor takk til spesielt Rolf Carlson, men også andre ansatte ved instituttet som jeg har vært i kontakt med.

Geir Simonsen

Sammendrag

Denne avhandlingen består av to artikler som sammen med avhandlingen utgjør selve masteroppgaven.

I avhandlingen prøver jeg å vise hvordan noen tradisjonelle former for design kan brukes i multivariable forsøk når formålet er å finne de viktigste variablene.

SVD-design brukes i en ny sammenheng som metode for å bestemme den eksperimentelle designen når formålet er å oppnå et forbedret utbytte.

Kombinasjonen av SVD-design og PLS-modell brukes i screeningforsøk når formålet er å identifisere de viktigste variablene med få forsøk.

Nøkkelord

Forsøksplanlegging (experimental design), Screening-forsøk (screening experiments), Nær-orthogonale eksperimenter (near-orthogonal experiments), Organisk syntese (organic synthesis).

Innholdsfortegnelse

Forord
Sammendrag
Nøkkelord

Mål 1
Nær-orthogonale eksperimenter i organiske syntesereaksjoner. 1

I-Screening av eksperimentelle variabler. 1
Introdukuksjon 1

II-Eksperimentell design 5
Screeningforsøk med fractional factorial design 5
Screeningforsøk med *Plackett-Burman design* 7
Screeningforsøk med D-optimal design 8

III-Bruk av tilnærmet orthogonale eksperimenter i screeningforsøk. 11
SVD-design og PLS-modell 11
Nærmere beskrivelse av metoden. 11
Prinsipper for konstruksjon av "nær-orthogonale eksperimenter" er følgende: 12

IV-Eksperimentelle studier og resultat: Brominering av et acetal 15
Diskusjon 20

V-Oppskalering 23
Diskusjon 24

VI-Enaminsyntesen Screeningforsøk med mange variabler. 27
Screening 29
Diskusjon 31

VII-Konklusjon 33

Eksperimentell del 34

Vedlegg 34

Mål

Nær-orthogonale eksperimenter i organiske syntesereaksjoner.

Målet for det arbeidet som beskrives i denne avhandlingen har vært å etablere en ny metode for forsøksplanlegging ved undersøkelse av organiske syntesereaksjoner. Metoden er basert på forsøksplaner der variasjonen i modellrommet er nær-orthogonal.

I-Screening av eksperimentelle variabler.

Introdukuksjon

Kjemiske reaksjoner er ofte komplekse mht. hvilke forhold som påvirker forløpet av en reaksjon. Slike forhold kan for eksempel være temperatur, trykk, pH, omrøringshastighet, partikkelstørrelse, løsemiddel, katalysatorer osv.

Det vil også ofte være samspillseffekter mellom faktorene. Det kan for eksempel være at pH og temperatur ikke påvirker reaksjonsforløpet uavhengig av hverandre. Videre er det ofte også slik at en ikke alltid har god forståelse av de indre egenskapene ved molekylene og det er derfor ofte vanskelig og kanskje umulig å se for seg hvilke egenskaper ved molekylene som har betydning for reaksjonsforløpet, enn videre hvordan endringer i reaksjonsbetingelsene vil påvirke disse indre egenskapene. Med indre egenskaper mener jeg elektrontetthet, molekylform og ioniserbarhet. Disse egenskapene vil ha betydning for hvordan substrat, reagenser og løsemidler påvirker hverandre. Ideelle reaksjonsbetingelser er derfor umulig å sette opp uten å gjøre forsøk.

Større utbytte, renere produkt og lavere kostnader er ofte forhold som en vil prøve å optimalisere i en kjemisk reaksjon. De resultatene man får kaller en responser og betegnes med y. De variablene som påvirker responsen betegner en med x og vi kan sette

$y = f(x)$

og fordi responsen, i de fleste tilfeller, påvirkes av flere variabler (for eksempel temperatur (X_1), omrøringshastighet (X_2), konsentrasjon (X_3)……) kan vi sette at responsen er en funksjon av alle variablene

$y = f(x_1,x_2,x_3,\ldots\ldots x_k)$

Ettersom resultatet, for eksempel utbyttet ved en kjemisk reaksjon, beror på energier kan en beskrive disse energiforholdene med en energiresponsflate. En reaksjonsvei blir da en vei over potensialflaten, fra et minimum til et annet minimum. Energiforskjellen (ΔG^o) mellom disse nivåene er relatert til likevektskonstanten. Den energibarrieren man må passere, aktiveringsenergien, er relatert til reaksjonsbetingelsene. Energipotensialflaten kan bestemmes fra kvantekjemiske beregninger ved å løse kompliserte differensialligninger. Det kan derfor antas at potensialflaten er kontinuerlig og differensierbar. Hvor dype disse energiminima er, hvor høy aktiveringsenergien er, beror på de detaljerte eksperimentelle betingelsene og at

$y = f(x_1,x_2,x_3,\ldots\ldots x_k)$

Et analytisk uttrykk for f er i de fleste tilfeller ukjent og det er vanskelig å utlede et slikt uttrykk fra fysikalsk-kjemiske modeller. En tilnærming for f kan en få via et Taylorutredning.

$$y = f(0) + \frac{df(0)}{dx_1} * x_1 + \frac{df(0)}{dx_2} * x_2 + \ldots\ldots + \frac{df(0)}{dx_k} * x_k + \frac{1}{2} * \frac{d^2f(0)}{dx_1 dx_2} * x_1 x_2 + \ldots\ldots\ldots + \frac{1}{2} * \frac{d^2f(0)}{dx_i dx_j} * x_i x_j + \ldots\ldots\ldots\ldots \frac{1}{2} * \frac{d^2f(0)}{dx_1^2} * x_1^2 + \ldots\ldots\ldots + \frac{1}{2} * \frac{d^2f(0)}{dx_k^2} * x_k^2 + \ldots . R(x) + \varepsilon.$$

(R(x) er en restterm som alltid blir mindre jo flere ledd som tas med i Taylor-utviklingen. ε er den eksperimentelle feilen)

Dette kan skrives på en enklere måte slik:

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots\ldots + \beta_{12} x_1 x_2 + \ldots\ldots + \beta_{ij} x_i x_j + \beta_{11} x_1^2 + \beta_{kk} x_k^2 + R(x) + \varepsilon$

I uttrykket beskriver modellparametrene (β_1, β_2 …) de partiellderiverte i Taylor uttrykket. De angir i hvilken grad variablene (x_1, x_2 ….) har innflytelse på responsen. De lineære koeffisientene β_1, β_2 …….. bestemmer hellinga av responsflaten i disse retningene. Modellparametrene β_{12} …. β_{ij} beskriver samspilleffekter mellom variabler og hvordan

responsflaten er vridd. De kvadratiske termene β_{ii} ……. viser krummina av responsflaten. Konstanten β_0 er gennomsnittsresponsen.

Koeffisientene i Taylorpolynomet kan estimeres ved å utføre eksperiment. Den estimerte modellen kan skrives:

$$y = b_0 + b_1x_1 + b_2x_2 + ……..+b_{12}x_1x_2 + ……+ b_{ij}x_ix_j + b_{11}x_1^2 + b_{kk}x_k^2 + R(x) + \varepsilon$$

hvor b er de estimerte koeffisientene. Disse blir bestemt fra en tilpasset forsøksplan.

En forsøksplan spesifiserer hvilke innstillinger de ulike variablene skal ha i de ulike eksperimentene. Det er viktig at variasjonen av variablene er uavhengige av hverandre over hele serien av eksperimenter. Det finnes forsøksplaner der variablene er helt ukorrelerte av hverandre og slike forsøksplaner kalles orthogonale.

Å undersøke forhold som har betydning for reaksjonsforløpet er viktig for å få økt innsikt i og forståelse for de kjemiske prosessene som påvirker et reaksjonsforløp.

I mange tilfeller er det imidlertid slik at tidsfaktoren er viktig og det finnes ikke tid til å gjennomføre hele forsøksserien. Hensikten med orthogonale eller nær-orthogonale eksperimenter, som beskrives i denne avhandlingen, vil i slike tilfeller ofte være å forbedre utbyttet eller renheten av et produkt fra et minimum av eksperiment. Altså ikke primært for å gi detaljert innsikt i de kjemiske prosessene som påvirker reaksjonsforløpet, ei heller å bestemme koeffisientene med maksimal presisjon. Det kan for eksempel innenfor medisinsk forskning være snakk om å produsere større mengder av et lovende legemiddel innenfor et kort tidsrom for å kunne komme videre i en fase 2-testing av legemiddelet.

Den situasjonen det gjelder kan beskrives slik: En kjent reaksjon skal brukes for å syntetisere en substans. Det er stort sett kjent hvor mye de eksperimentelle variablene kan tillates å variere. Det kan imidlertid antas at et forbedret resultat kan oppnås ved å justere de eksperimentelle variablene noe fra det som i dag er kjent som de beste eksperimentelle betingelsene. Man ønsker derfor å kunne beskrive hvordan responsen, y, varierer i det eksperimentelle domenet. Det kan derfor antas at en begrenset Taylormodell kan gjøre det. Til det kan vi bruker en andre ordens samspillsmodell

$$y = \beta_0 + \sum \beta_i X_i + \sum\sum \beta_{iy} X_i X_y$$

For å estimere koeffisientene bruker man en egnet design [3, 4, 5].

I beskrivelsen av en respons y som skissert ovenfor må en bestemme hvilke variabler som er hensiktsmessige å ta med og hvor mange ledd i Tayloruttrykket som skal tas med. En bestemmer seg for en modell.

Hvilken modell en velger vil være avhengig av hva en ønsker å få fram og hvilket problem en ønsker å belyse. I en tidlig fase av et prosjekt vil en kunne begrense modellen til bare å finne hovedeffekter, mens en senere i prosjektet vil søke etter de optimale betingelsene for reaksjonen. En må kanskje endre domenet og/eller ta med flere ledd i Taylor-uttrykket. Eller en kan innskrenke domenet eller eventuelt utvide med flere nivåer på parametrene. Dette forutsetter imidlertid at en er rimelig sikker på å finne optimale betingelser innenfor det aktuelle domenet. Med mange parametere og/eller mange nivåer på parametrene vil også screeningforsøkene ofte ta lang tid å gjennomføre. Det vil derfor være behov for screeningforsøk som på en rask måte kan finne mer optimale betingelser for en reaksjon.

Modellen en bestemmer seg for å bruke gir grunnlaget for den eksperimentelle designen.

Med en *multivariabel eksperimentell design* er det mulig å håndtere og variere flere variabler samtidig slik at en får ut den informasjonen en søker etter.

II-Eksperimentell design

Taylormodellen er uttrykt i skalerte (scaled) og sentrerte eksperimentelle variabler. Det betyr at Taylorkoeffisientene blir et mål på hvilken betydning variablene som undersøkes har.

Skalert (scaled)variabel. Eksempel:

Dersom temperaturen i et forsøk skal varieres mellom T_1 = 60 gr. og T_2 = 80 gr. kan vi beregne den konstruerte variabelen slik:

Midtpunkt $T_m = \frac{T_1 + T_2}{2} = \frac{60+80}{2} = 70$ og $\Delta T = T_2 - T_m = 80 - 70 = 10$

Scaled variabel: $T_{(høy)} = \frac{T_2 - T_m}{\Delta T} = \frac{80-70}{10} = 1$ og $T_{(lav)} = \frac{T_1 - T_m}{\Delta T} = \frac{60-70}{10} = -1$

I screeningforsøk vil en primært være interessert i å finne hvilke hovedeffekter og eventuelt også hvilke samsplillseffekter som har størst betydning for reaksjonsforløpet. Hvis resultatene fra screeningforsøkene er gode kan disse forsøkene være nok for å finne mer optimale betingelser for reaksjonen.

Det finnes mange former for design, men i screeningforsøk er det ofte tilstrekkelig å undersøke variablene på to nivåer. Til det bruker en ofte reduserte faktorforsøk (*fractional factorial design*) [3], *Plackett-Burman* [4] eller *D-optimal design*[5]. Under vil jeg gi noen eksempler på etablerte screeningforsøk for å finne viktige variabler.

Screeningforsøk med fractional factorial design [3]

Eksempel med 4 variabler (X_1, X_2, X_3 og X_4) på 2 nivåer.

Et slikt forsøk kan undersøkes i et redusert faktorforsøk (*fractional factorial design*) med 2^{4-1} = 8 eksperimenter.

Det tilsvarer et *full factorial de*sign hvor variabler og samspillseffekter mellom variablene kan settes opp slik:

X_1 X_2 X_3 X_1X_2 X_1X_3 X_2X_3 $X_1X_2X_3$

og hvor modellparametrene b_0, b_1, b_2, b_3, b_{12}, b_{13}, b_{23} og b_{123} vil inngå.

Setter vi $X_4 = X_1X_2X_3$ får vi generatoren I = 1234

Generatoren forteller oss hvordan koeffisientene er ”confounded”. Modellparametrene får følgende ”confoundings”:

Modellparameter	Confoundings
b_0	b_{1234}
b_1	b_{234}
b_2	b_{134}
b_3	b_{124}
b_{12}	b_{34}
b_{13}	b_{24}
b_{23}	b_{14}
b_{123}	b_4

Tabell 1: Modellparametre og confoundings

I dette eksemplet har vi en resolution IV, det vil si at hovedeffekter er counfounded med 3-faktor interaksjonseffekter og 2-faktor interaksjonseffekter er counfounded med hverandre.

Med antagelsen om at 3-faktor interaksonseffekter er ubetydelige, ser vi at vi får beregnet alle 4 hovedeffektene uten confoundings og at 2-faktor interaksjonseffekter er confounded med 2-faktor interaksjonseffekter. Dette vil ofte være tilstrekkelig i et screeningforsøk.

Imidlertid kan en også kjøre komplementære forsøk for å separere confoundings. I eksemplet ovenfor er b_{12} confounded med b_{34} (innstillingene for X_1X_2 varierer på samme måte som X_3X_4) og en kan ikke vite hvilken av disse som har størst betydning. En kan da kjøre eksperimenter hvor X_1X_2 varierer forskjellig fra X_3X_4. Dette kan gjøres i 2 forsøk som vist i tabellen nr. 2 under.

X_1X_2	X_3X_4
1	-1
1	1

Tabell 2: Separering av to-faktor interaksjonseffekter

Dette gir en modell hvor responsen $Y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_{12}x_1x_2 + b_{34}x_3x_4$

Da alle hovedeffektene er bestemt tidligere vil det være tilstrekkelig å kjøre ett eksperiment i tillegg for å separere confoundingene.

Screeningforsøk med *Plackett-Burman design* [4]

I screeningforsøk med mange variabler vil det i første rekke være interessant å finne hovedeffektene. Som illustrasjon tenker vi oss at vi har 17 variabler. I et redusert faktorforsøk er det minste antall eksperimenter vi må vi kjøre lik $2^{17-12} = 2^5 = 32$ for å utforske alle hovedeffektene. Det vil imidlertid være tilstrekkelig med 18 eksperimenter for å bestemme 17 hovedeffekter.

I *Plackett –Burman design* brukes en Hadamard-matrise H_n. En slik matrise har følgende egenskaper:

- H_n er en $n \times n$ matrise med elementene -1 og 1 og hvor
- $H^TH = n*I_n$ gjelder.

Fra orthogonal design vet vi at X^TX er en diagonalmatrise og at $X^TX = n*I_n$. Det betyr at søylene i en Hadamard-matrise er orthogonale, og at modellmatrisene fra *fractional factorial design* også er Hadamard-matriser.

Plackett og Burman har vist hvordan en får Hadamard-matriser for $n = 4, 8, 12, 16, 20, 24, 28$ når n er en multippel av 4.

I en *Plackett-Burman design* [2] brukes Hadamard-matriser for å definere screeningeksperimenter hvor $n-1$ variabler kan studeres med n forsøk.

Konstruksjonen av designmatrisen får en ved en syklisk permutasjon av den første rekken. Rolf Carlsson og Johan E. Carlsson presenterer en tegntabell for den første rekken og konstruksjonen av en slik design er vist for n = 8. (2005:170)

Fra Hadamard matrisen kan koeffisientene i en lineær modell beregnes slik (eks. med n = 8):

$b = (H^TH)^{-1}H^Ty = \frac{1}{8}*I_8H^Ty = \frac{1}{8}*H^Ty$

Screeningforsøk med D-optimal design [5]

For å oppnå maksimal presisjon på modellparametrene må determinanten til spredningsmatrisen $|(X^TX)^{-1}|$ være så liten som mulig. Dette gjør at "the joint confidence region" blir så liten som mulig [8].

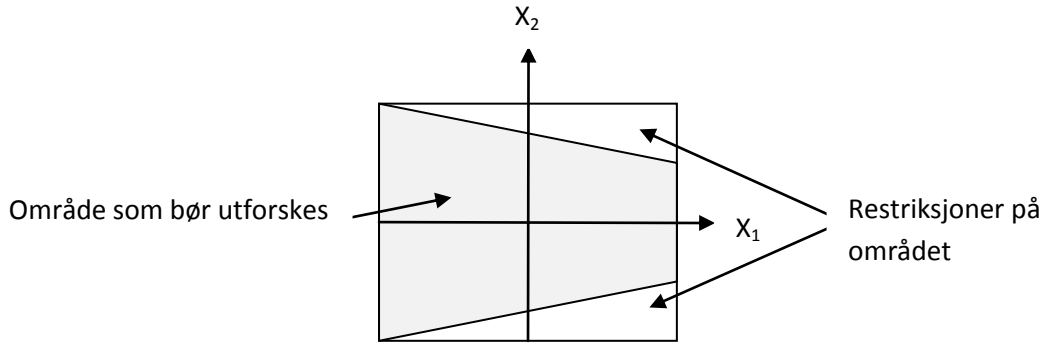
Da $|(X^TX)^{-1}| = \frac{1}{|X^TX|}$

er dette ekvivalent med å si at determinanten $|X^TX|$ skal være så stor som mulig. Dette oppnår vi både med *fractional factorial design* [1] og *Plakett.Burman design* [2]. Disse formene for design er derfor D-optimale.

I forsøk hvor:

- Det er restriksjoner på det aktuelle domenet (f. eks. på grunn av sikkerhet)
- Det er påkrevd med et minimum antall eksperimenter
- Vi vil kjøre komplementære eksperimenter

kan en risikere at store deler av domenet ikke dekkes opp av forsøkene. Se illustrasjonen under.



Figur 1: Eksperimentelt domene med restriksjoner.

For å få utforsket hele det aktuelle domenet kan en bruke en D-optimal design. Konstruksjonen av D-optimal design krever bruk av datamaskiner. Det er en ulempe. Det finnes PC-programvare som kan velge ut punkter på en slik måte at hele området er dekket og slik at determinanten $|X^TX|$ blir så stor som mulig. Dette kalles for en *D-optimal design*[5].

I D-optimal design vil variablene ikke være uavhengige av hverandre, selv om de er beregnet med maksimal presisjon og er så uavhengige som mulig. Dette gjør at screening med D-optimal design på ukjente reaksjoner ikke er å anbefale. I ukjente reaksjoner vil det være uklart hvilke effekter og samspilleffekter som bør være med. Få eksperimenter gjør også at antall frihetsgrader er begrenset mht. å kontrollere gyldigheten av modellen .

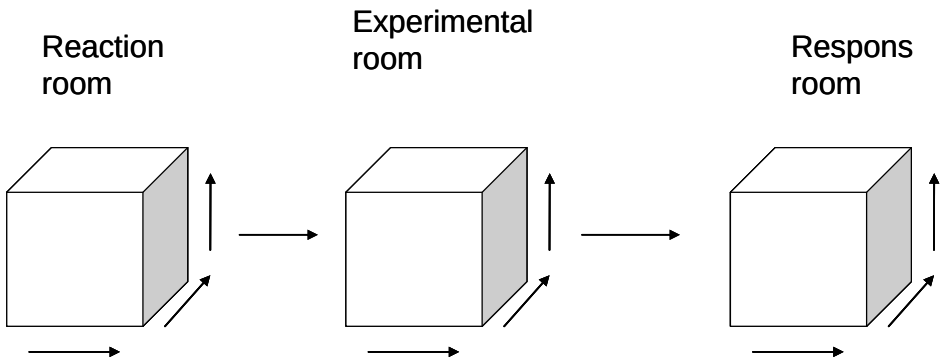
Felles for alle screeningforsøk er at en vil finne hvilke variabler som er av størst betydning for responsen. Hvilken design en velger å bruke er avhengig av flere forhold. Forhold som må tas hensyn til kan for eksempel være: - hvor godt en kjenner reaksjonen på forhånd, - hvor mange variabler vi har, - hvilke interaksjonseffekter som er sannsynlige, - hva formålet med eksperimentene er, - og om det er slik at en maksimal presisjon av modellparametrene er viktig .

III-Bruk av tilnærmet orthogonale eksperimenter i screeningforsøk.

SVD-design og PLS-modell

Nærmere beskrivelse av metoden.

Vi tar utgangspunkt i figuren under:

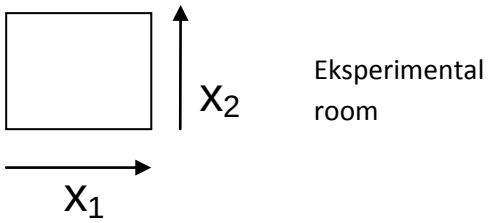


Figur 2: Reaksjonsrom, eksperimentelt rom og responsrom

Figur 2 viser at det er sammenhenger mellom de forskjellige ”rommene”, og at endringer i for eksempel reaksjonsrommet også vil føre til endringer i responsrommet.

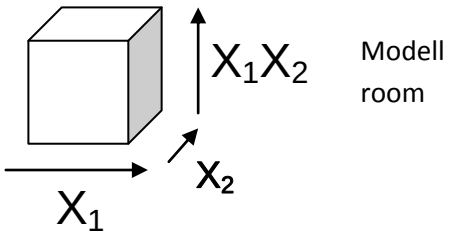
Men vi kan også tenke oss et modellrom (Modell room). For å klargjør vil jeg illustrere dette med et eksempel med en modell med 2 variabler. Dette fordi det er enkelt å illustrere den geometriske tolkningen av dette med figurer i 2 og 3 dimensjoner.

Det eksperimentelle rommet ”spennes altså ut” av variablene X_1 og X_2 slik som vist under:



Figur 3: Eksperimentelt rom utspent av to variabler

Har vi så en modell hvor det også er en interaksjon mellom variablene x_1 og x_2, vil vi ha et modellrom som "spennes ut" av 3 variabler. Geometrisk illustrasjon:



Figur 4: Modellrom utspent av tre variabler

Det gjelder derfor å velge eksperimenter slik at disse spenner ut modellrommet. Modellrommet defineres av variablene som inngår i Taylor-modellen. Taylor-modellen er

$$Y = b_0 + b_1x_1 + b_2x_2 + b_{12}\, x_1x_2$$

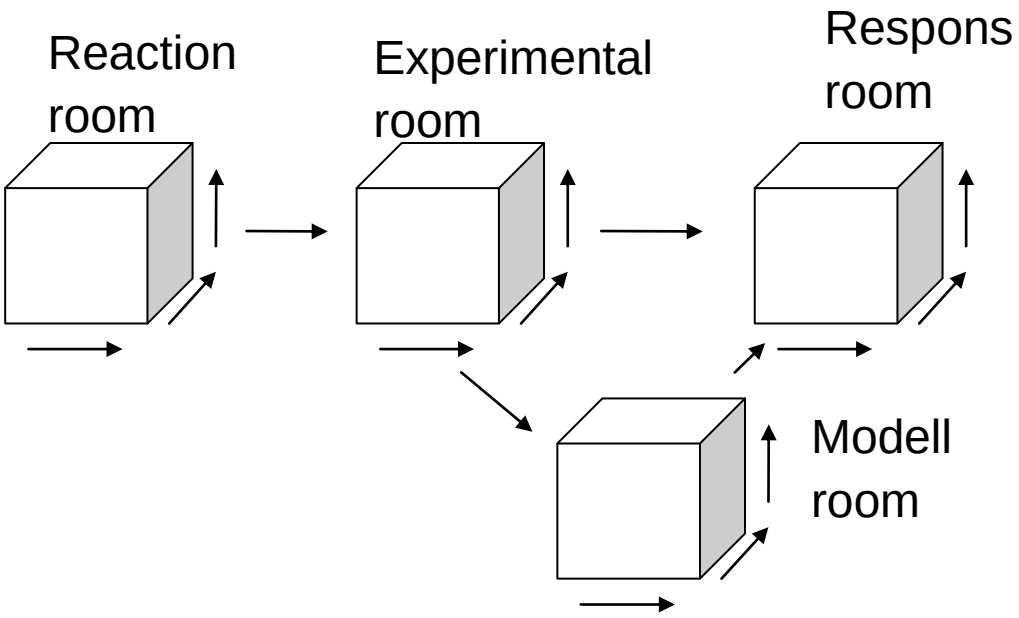
For ytterlige eksempler se ref. 1.

Prinsipper for konstruksjon av "nær-orthogonale eksperimenter" er følgende:

(1) Angi Taylormodellen – Denne definerer modellrommet.
(2) Legg ut en forsøksplan av mulige eksperiment slik at disse spenner ut variabelrommet.
(3) Utvid kandidat-designmatrisen til tilsvarende kandidat-modellmatrise (X_c).
(4) Bruk SVD til å velge ut nær-orthogonale rader i X. De utvalgte radene definerer siden forsøksplanen.

I ref. 1 vises forsøksplan for 3, 4 og 5 variabler for ulike Taylor-modeller. Kandidateksperimentene som ble brukt i bromineringsforsøket var et 11-nivåers *full factorial* forsøk.

Ethvert punkt i modellrommet vil ha ei bestemt innstilling i eksperimentrommet. Det er altså en sammenheng mellom punktene i modellrommet, eksperimentrommet og responsrommet:



Figur 5: Sammenheng mellom ”rommene”

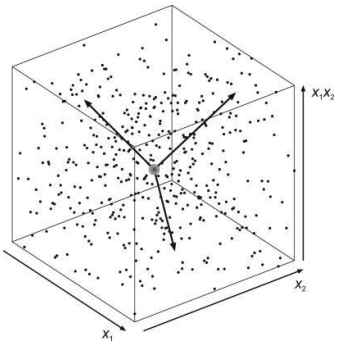
I bromineringsforsøket (brominering av acetal) har vi brukt en modell med 4 variabler på 11 nivåer. Det gir totalt $11^4 = 14641$ forskjellige variasjoner. Dette antar vi er tilstrekkelig for å spenne ut modellrommet på en god måte. I beskrivelsen under tar jeg utgangspunkt i dette eksperimentet for å beskrive metoden.

Med en modell hvor alle hovedeffekter og samspillseffekter skal bestemmes har vi:

$$y = b_0+b_1x_1+b_2x_2+b_3x_3+b_4x_4+b_{12}x_1x_2+b_{13}x_1x_3+b_{14}x_1x_4+b_{23}x_2x_3+b_{24}x_2x_4+b_{34}x_3x_4$$

Det er altså et modellrom som “spennes ut” av 10 variabler. Modellen har 11 koeffisienter. Det kreves derfor minst 11 eksperimenter for å estimere koeffisientene.

Ved bruk av SVD-design (Singular Value Decomposition) plukker en så ut ”retninger” i dette modellrommet hvor vi har størst variasjon. Med SVD design kan disse dimensjonene trekkes ut av modellrommet en for en. En detaljert beskrivelse for hvordan dette gjøres er gitt i ref.1.

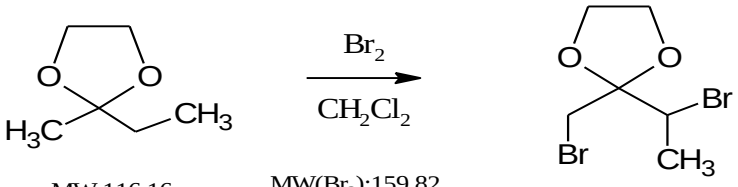


3-dimensjonalt rom som spennes ut av orthogonale vektorer. Vektorene angir retninger hvor vi har størst variasjon.

Figur 6: 3-dimensjonalt rom som spennes ut av orthogonale vektorer

Prinsippene, som beskrevet ovenfor, er anvendt på 2 forsøk: Brominering av et acetal og i syntese av enamin med molekylsikter.

IV-Eksperimentelle studier og resultat: Brominering av et acetal



MW:116,16 MW(Br_2):159,82 MW: 273,98

Br_2: d = 3,12 g/cm^3

I forsøket bruktes en modell med kryssprodukter og 4 variabler på 11 nivåer slik som beskrevet ovenfor under "Nærmere beskrivelse av metoden " og i vedlagte artikkel [2].

Variabler	Nivå			
	-1	0	0,2	1
X_1: Temperatur/ ^oC	0	15	-	30
X_2: Konsentrasjon av acetal	0,2	0,3	-	0,4
X_3: Omrøring/rpm	250	325	340	400
X_4: Bromtilsetning/meq min^{-1}	20	50	-	70

Tabell 3: Variabler og nivåer brukt i designmatrisen:

Fire variabler ble undersøkt, se tabell 3. Taylor-modellen var en andre ordens samspillmodell, se ovenfor.

Tabellen 4 under viser de 11 eksperimentene som ble plukket ut med SVD-design og tilhørende utbytte. Som nevnt tidligere har modellrommet 11 ukjente parametere, og det er derfor nødvendig å kjøre 11 eksperimenter hvis en vil kunne finne alle 11 parametrene ved bruk av et Taylorpolynom og en minste kvadrattilpasning. Som det framgår av tabell 2 fikk vi et svært bra resultat allerede i eksperiment nr. 2. Hadde det vært dårlig med tid kunne vi ha stoppet allerede der. For å kunne validere resultatet ble hele serien på 11 eksperimenter pluss

et eksperiment i senterpunktet kjørt. Senterpunktet brukes for å bestemme en responsmodell. Senterpunktet tilsvarer de hittil kjente ”beste betingelser”.

Eksperimentell design og utbytte av dibromoacetal (4 timer)					
Design					Utbytte (4h)
Eksp.nr	X_1	X_2	X_3	X_4	Y
1	1	1	1	1	87,4
2	1,0	-1,0	-1,0	-1,0	95,8
3	-1,0	1,0	-1,0	1,0	79,5
4	-1,0	-1,0	1,0	-1,0	63,7
5	-1,0	-1,0	0,2	1,0	53,9
6	-1,0	1,0	1,0	-1,0	68,7
7	1,0	1,0	1,0	-1,0	58,8
8	1,0	-1,0	1,0	-1,0	93,5
9	1,0	-1,0	-1,0	1,0	94,0
10	-1,0	-1,0	-1,0	-1,0	77,1
11	1,0	-1,0	1,0	1,0	80,9
12	0	0	0	0	88,6

Tabell 4: Eksperimentell design og utbytte av dibromoacetal etter fire timer.

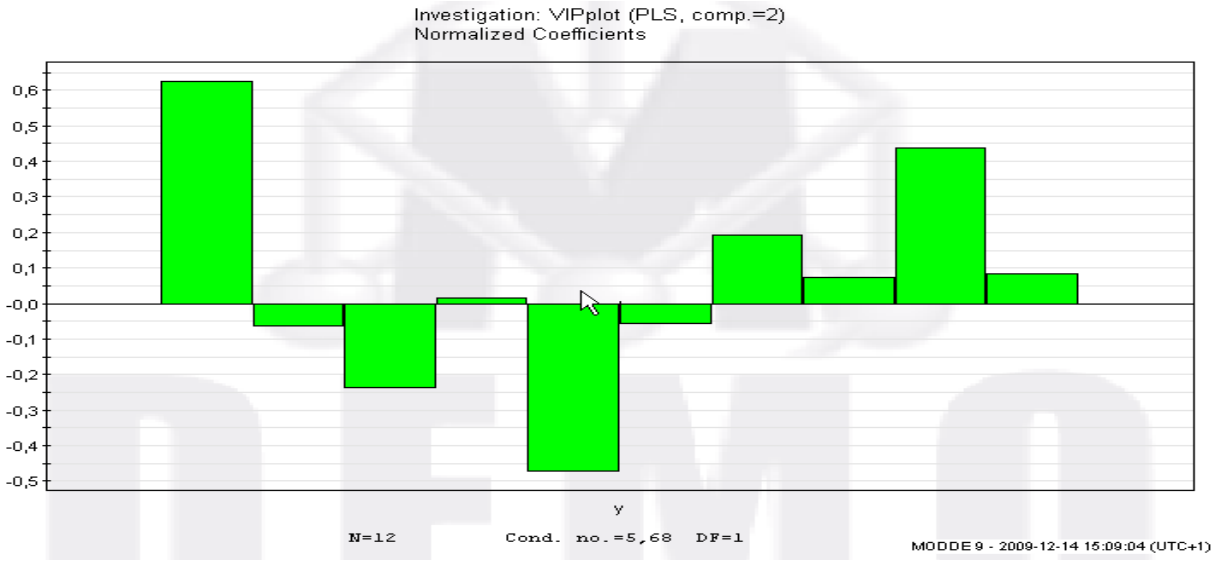
Designmatrisen og utbyttene etter 4 timer ble lagt inn i dataprogrammet ”Modde” [6], og koeffisientene i Taylorutrykket ble estimert med PLS-regressjon. Dette ga:

$y = 77,71 + 8,92\ x_1 - 0,71\ x_2 - 3,11\ x_3 - 0,18\ x_4 - 6,83\ x_1x_2 - 1,24\ x_1x_3 + 2,66\ x_1x_4 + 0,69\ x_2x_3 + 6,24\ x_2x_4 + 1,64\ x_3x_4 + e$

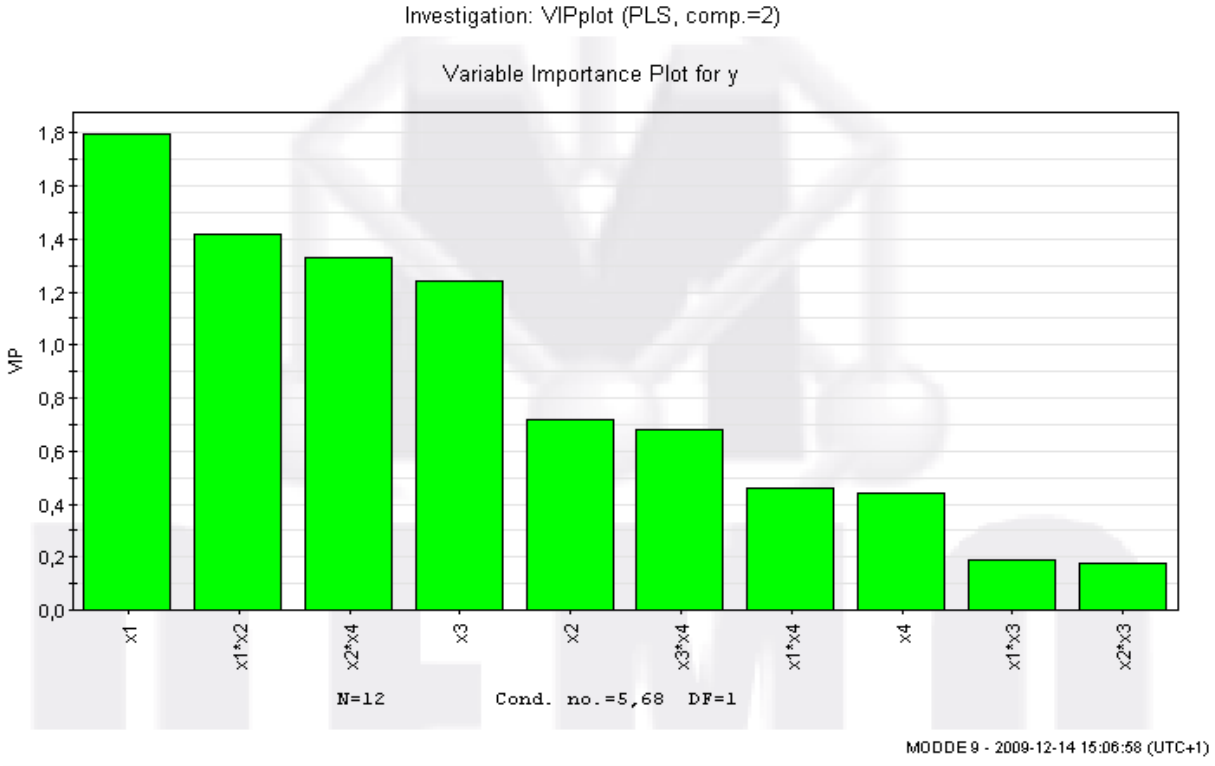
Tabell 2 i ref. 2 viser kumulativ normalfordelingsplot av koeffisientene etter at henholdsvis 5, 6 og alle eksperimentene var kjørt. Alle plottene viser at X_1 (temperatur) er en signifikant parameter.

Fra dataprogrammet kan en hente ut mange forskjellige tabeller og plot. Jeg har tatt med noen her:

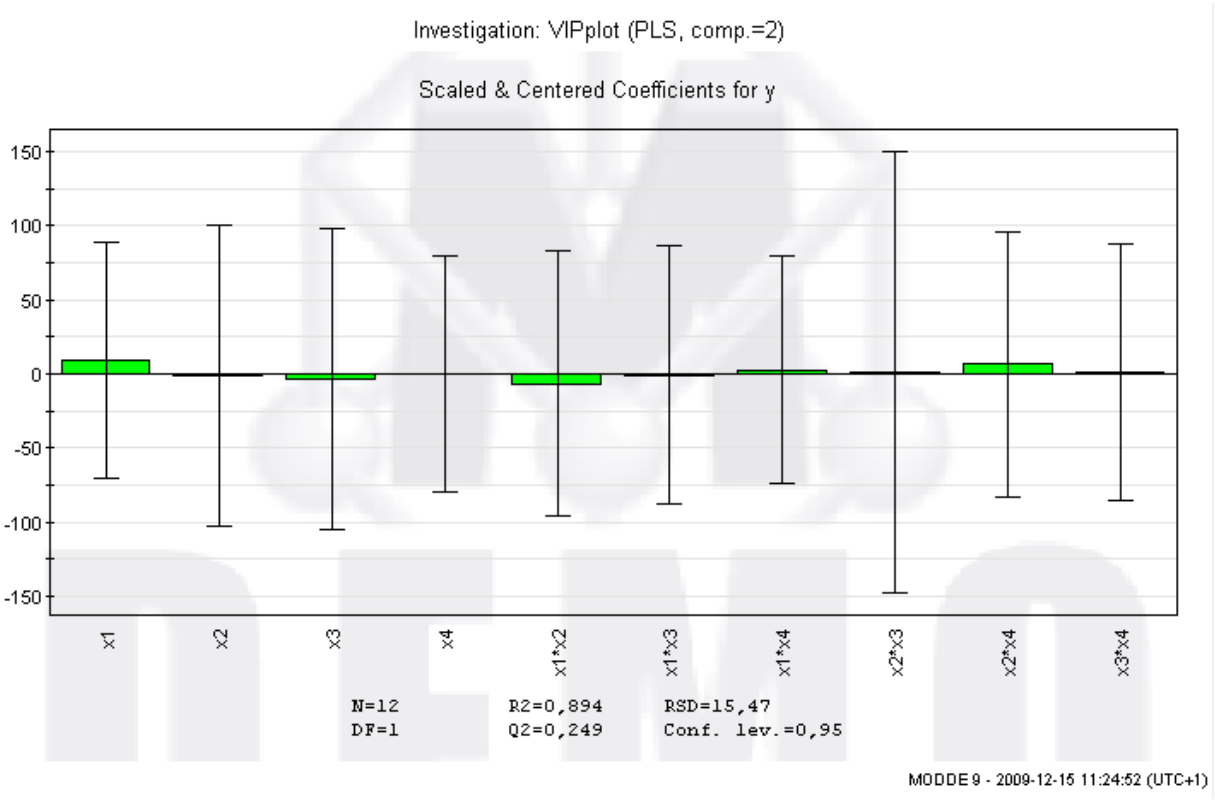
Variable importance plot:



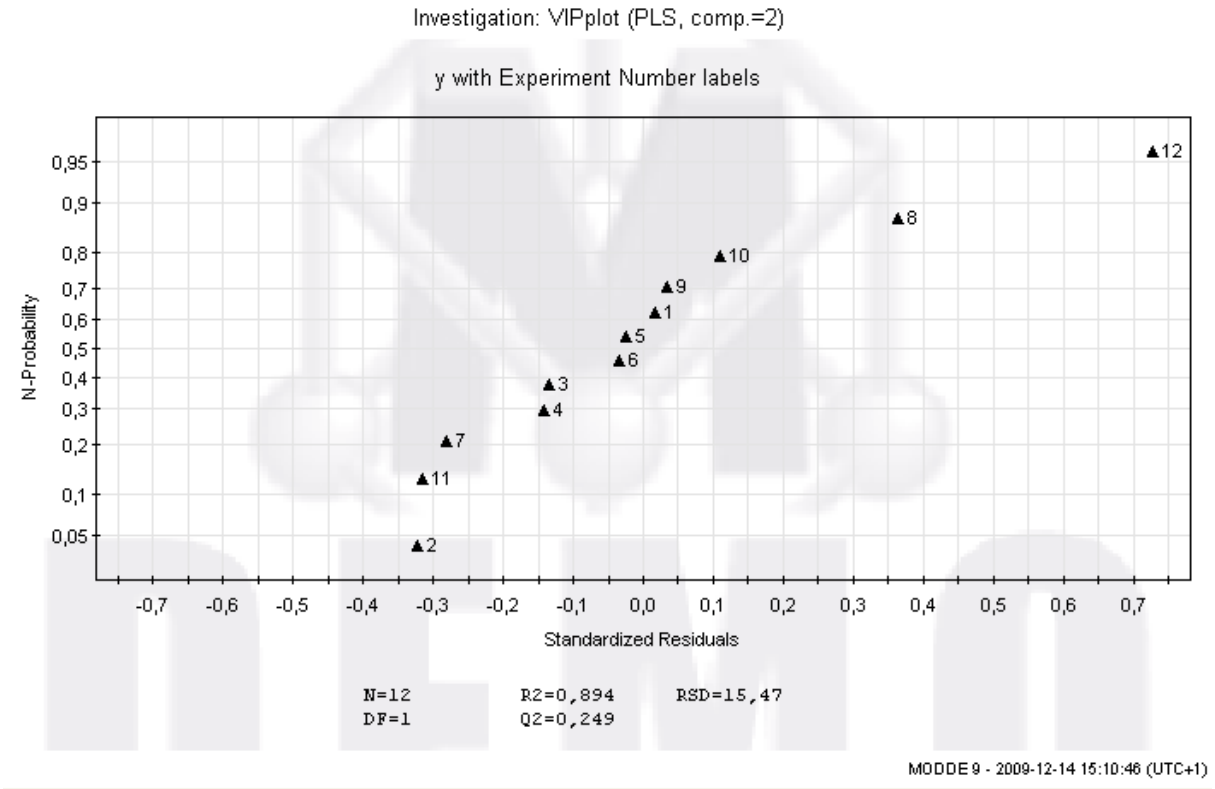
Coefficient plot:



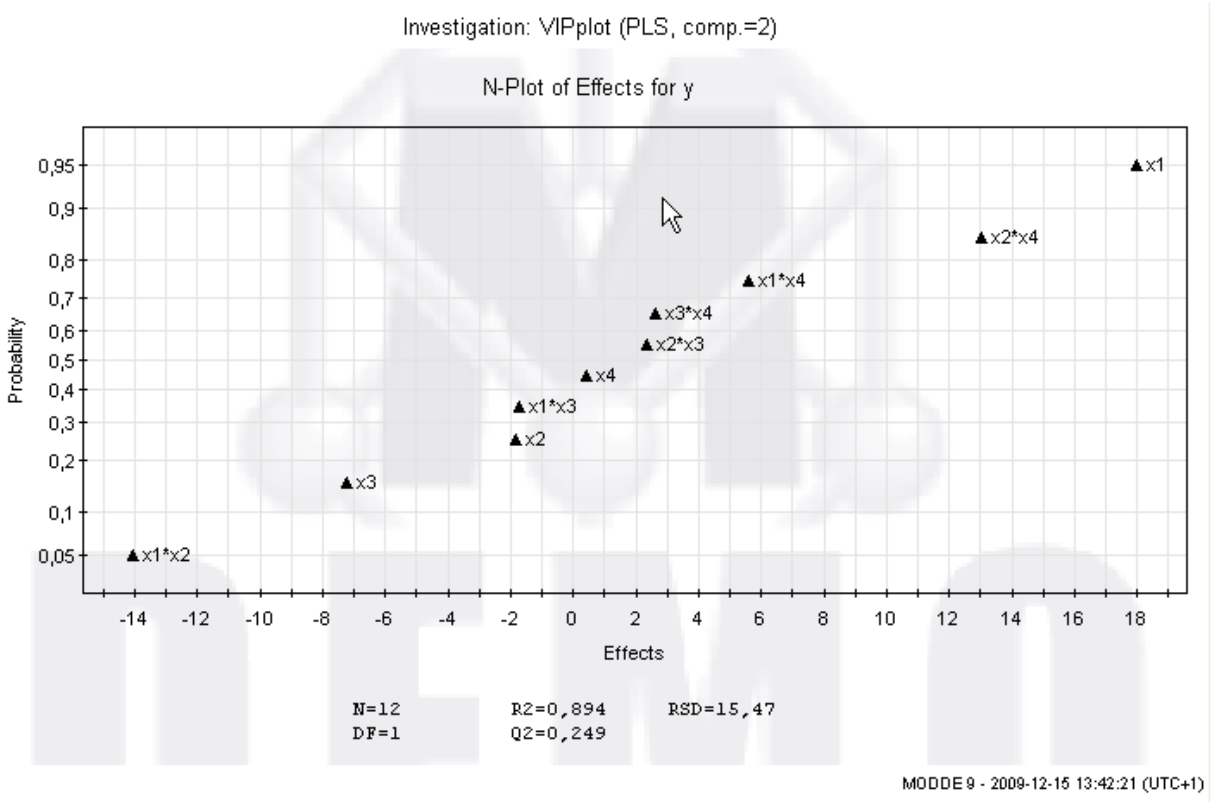
Scaled and centered coefficients for yeld

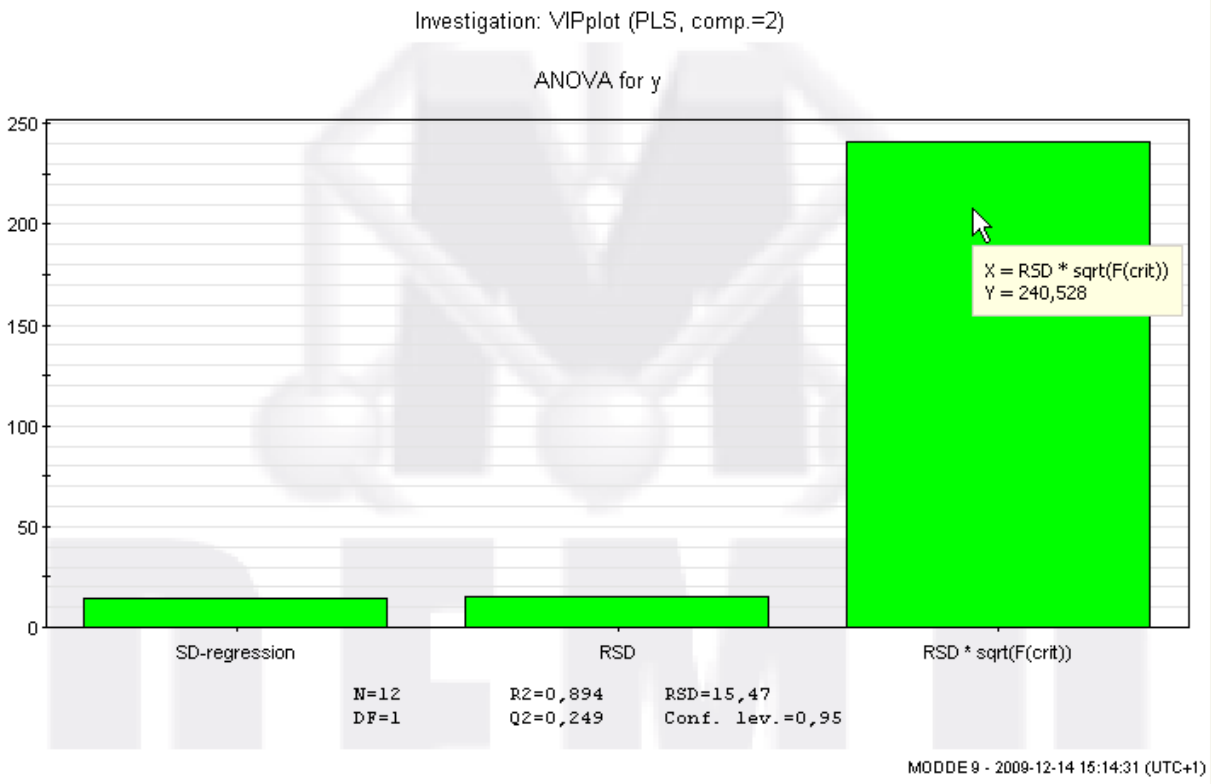


Normalfordelings plot of residuals:

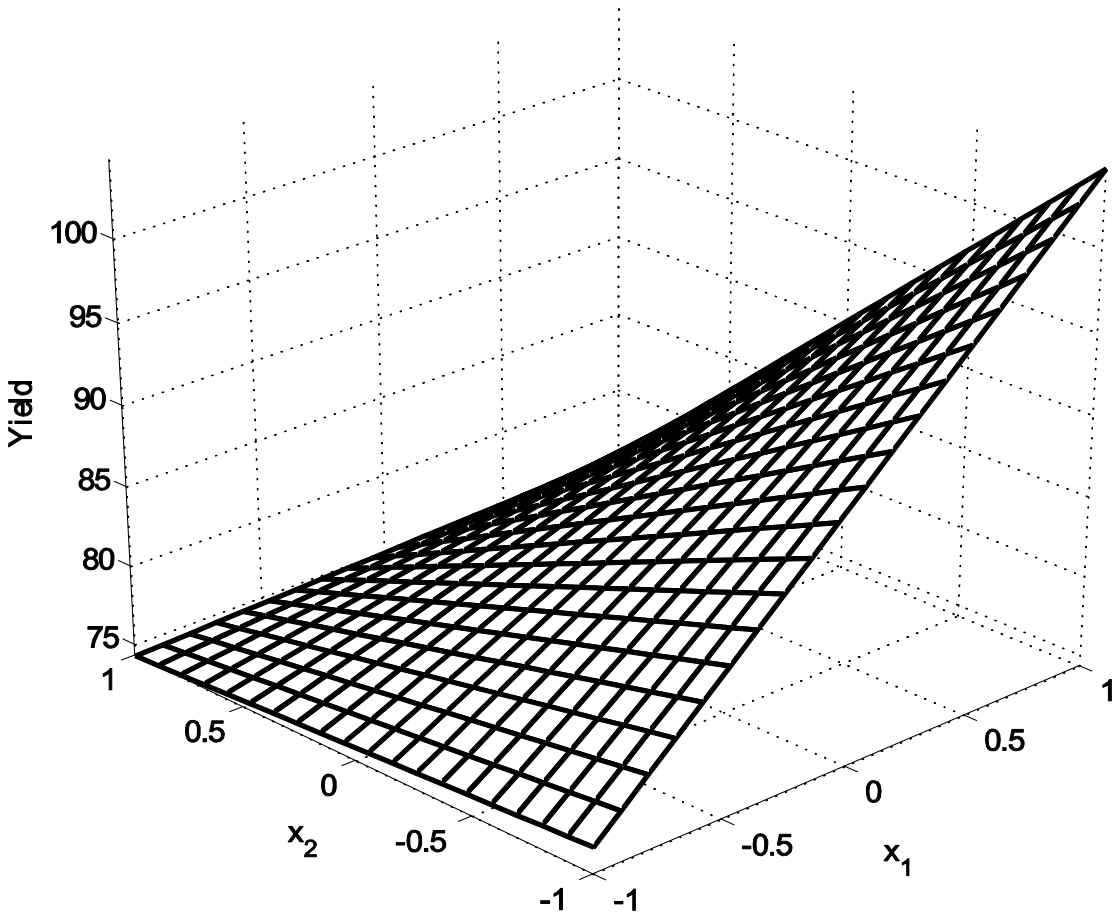


Normalfordelingsplot av effekter:





Responsflate



Diskusjon

Analysen i ”Modde” forteller oss at temperaturen X_1 har størst innflytelse. Dette så vi fra de kumulative normalfordelingsplottene allerede etter at 5 eksperimenter var kjørt. Interaksjonseffektene mellom temperatur og konsentrasjon X_1X_2, interaksjonseffekten mellom konsentrasjon av acetal og addisjonshastighet av brom X_2X_4, og omrøringshastighet X_3 ser også ut til å ha betydning.

Videre forteller analyseverktøyene i ”Modde” at modellen ikke har gode statistiske egenskaper. Dette er for så vidt ikke noen overraskelse. Orthoghonale forsøk med en design som gir gode statistiske egenskaper vil gi mer nøyaktige beregninger av modellparametrene. I søk etter mer optimale betingelser i et relativt sterkt avgrenset domene, burde en modell sannsynligvis også inneholde kvadratiske termer og derfor også kreve mange flere forsøk for

å gi en god modellfit. En må likevel ha med seg at en ikke har som mål å få en best mulig modell. Hensikten er å finne bedre reaksjonsbetingelser med få eksperimenter.

Etter at alle forsøkene var kjørt ble modellen slik:

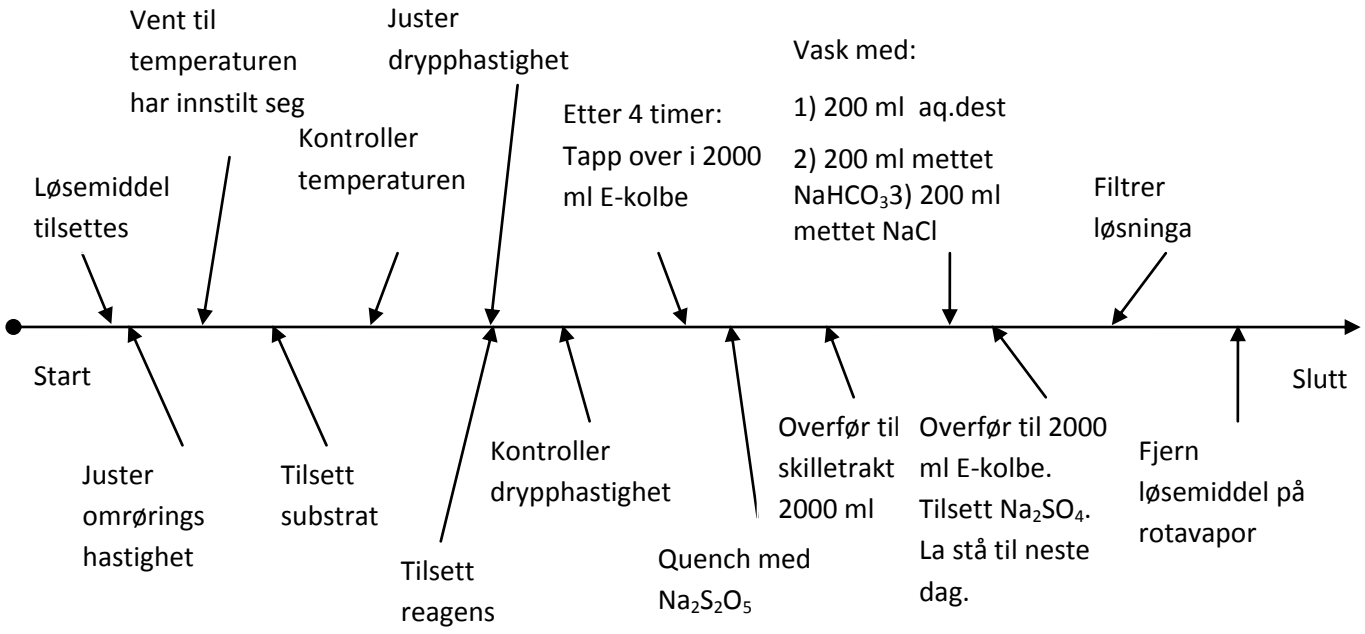
$$y = 77{,}71 + 8{,}92\ x_1 - 0{,}71\ x_2 - 3{,}11\ x_3 - 0{,}18\ x_4 - 6{,}83\ x_1x_2 - 1{,}24\ x_1x_3 + 2{,}66\ x_1x_4 + 0{,}69\ x_2x_3 + 6{,}24\ x_2x_4 + 1{,}64\ x_3x_4 + e$$

og tolkes slik at temperatur X_1 bør settes til høyt nivå, konsentrasjonen X_2 bør settes til lavt nivå, omrøringshastighet X_3 settes til lavt nivå og bromtillsetting X_4 bør settes til lavt nivå.

Med disse innstillingene vil interaksjonseffektene ha maksimal innflytelse.

V-Oppskalering

Etter at screeningeksperimentene var gjennomført ble det gjort et oppskaleringsforsøk (10X) hvor 2 batcher ble kjørt. Variabelinnstillingene ble gjort med utgangspunkt i screeningforsøkene fra PLS-modellen. Ved oppskalering er det imidlertid flere forhold som bør vurderes. Gode kunnskaper om reaksjonsmekansimer er viktige og bruk av et "Ishikawa diagram" bør vurderes.

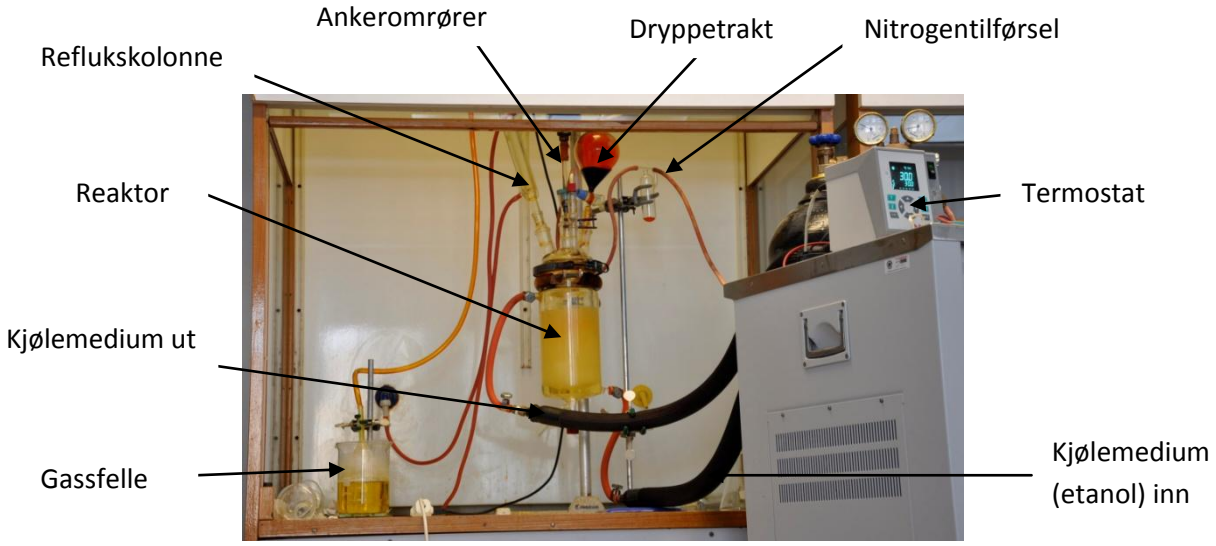


Figur 7: Ishikawa diagram:

	Variabler				Utbytte/%
	Temp./^{o}C	Kons. av acetal	Omrøring/rpm	Bromtilsetn./meq^{-1}	
Batch 1	30	1	220	20	88,3
Batch 2	30	1	220	20	98,2

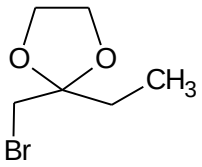
Tabell 5: Variabler, innstillinger og utbytte ved oppskaleringsforsøket:

Nærmere redegjørelse for prosedyren finnes i vedlagte artikkel. På batch 1 fikk jeg lekkasje av brom mellom reaktor og lokk. For å få brom til å dryppe rett ned i reaksjonsblandinga byttet jeg til ei dryppetrakt med lang spiss. Denne hadde imidlertid ikke trykkutjevning og jeg måtte derfor flytte nitrogentilsetninga til en annen hals på lokket. Bildet under viser oppsettet på batch 2.



Diskusjon

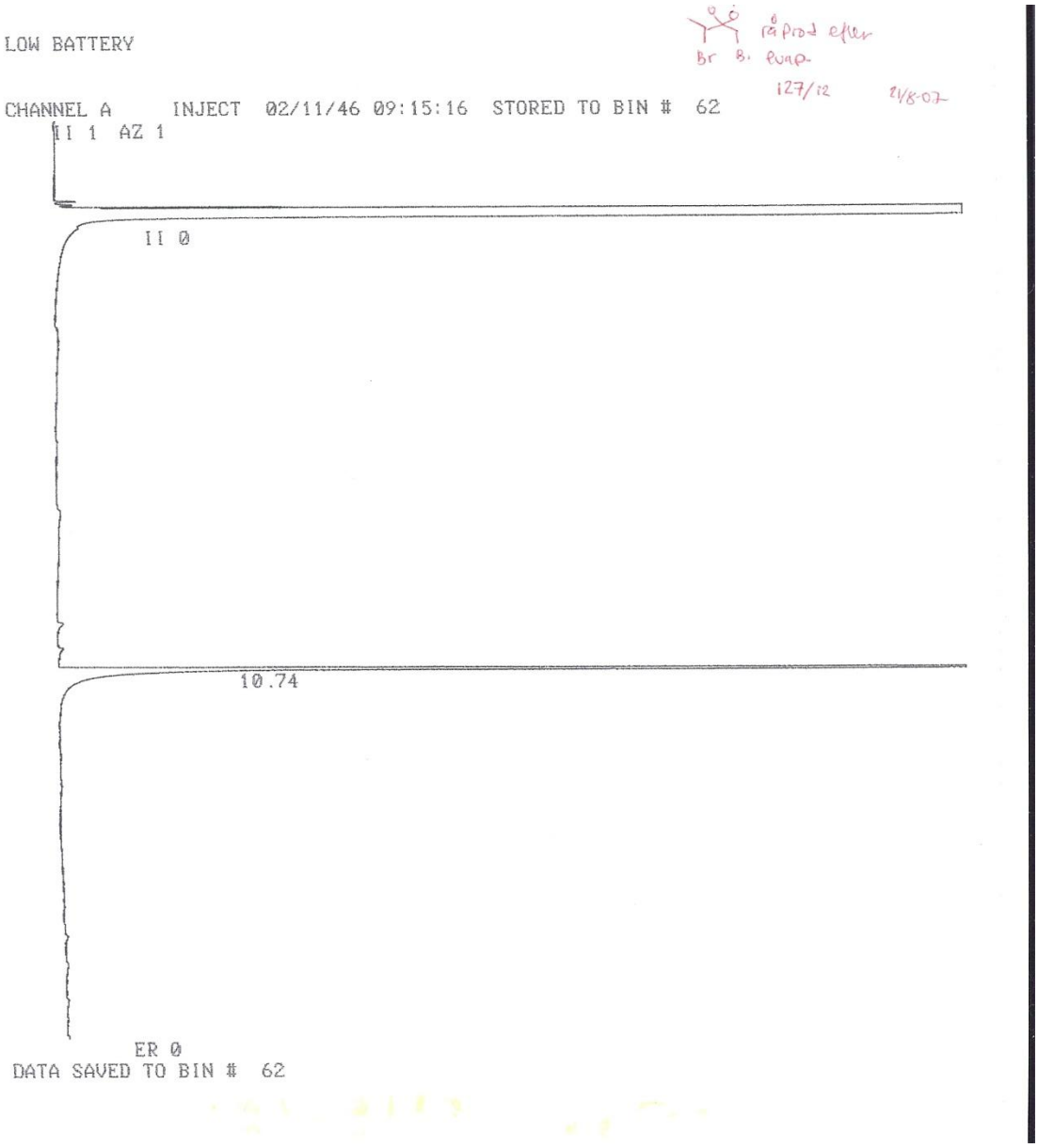
Det ble tatt prøver for både GC og NMR. Gasskromatogrammet fra batch 1 viste imidlertid forurensning av monobrominert acetal:



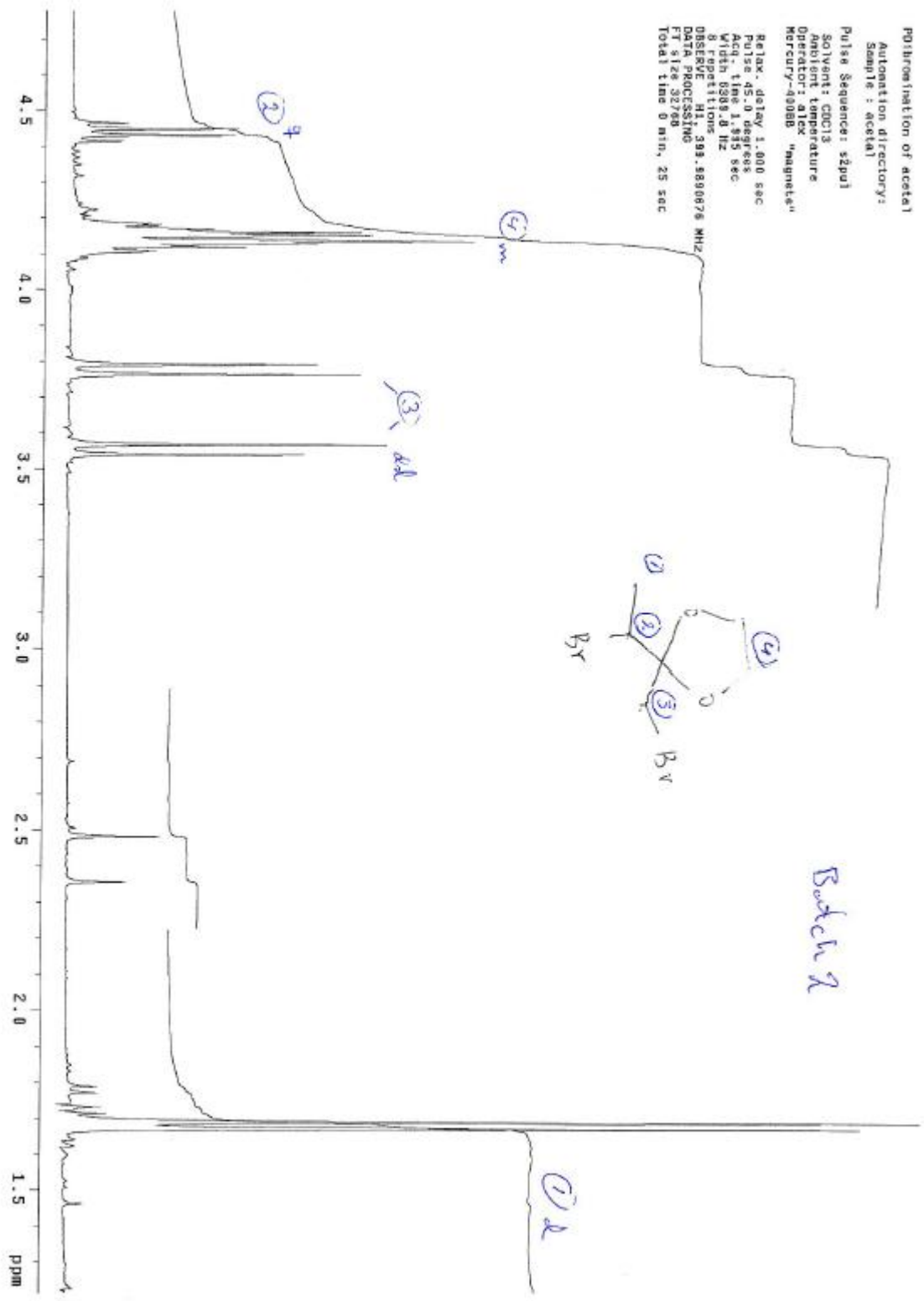
Det var derfor ikke aktuelt å kjøre NMR på batch 1.

I disse reaksjonene ble faktisk utbytte beregnet til 98%. Dette styrker påstanden om at PLS-modellen er god mht. å finne de viktigste variablene og gi nyttig informasjon om innstillinger for å få økt respons.

Under vises GC av råproduktet i fig. 8 og ^{1}HNM-spekter av råproduktet i fig. 9.

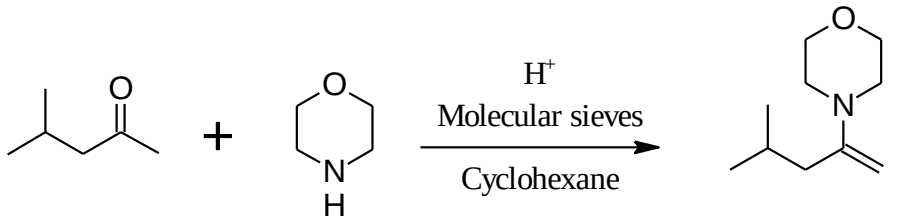


Figur 8: GC av råproduktet



Figur 9: [1]HNMR -spekter av råproduktet

VI-Enaminsyntesen Screeningforsøk med mange variabler.



MW = 100.16 MW = 87.12 MW = 169.26

Eksperimentene ble kjørt i 50 ml testrør ved hjelp av et "heating block reactor system", Bohdan 2080 Miniblock™, fra Mettler Toledo. 7 variabler ble undersøkt. Se tabell 6. Taylormodellen var en andre ordens samspillmodell med to diskrete variabler. *Kandidat designmatrisen* ble definert med en $2^3 * 5^4$ full faktorial design med totalt 5000 eksperimentelle innstillinger. Nærmere redegjørelse for modellen og prosedyren er gitt i vedlagte artikkel [2].

Variabler	Innstillinger				
	-1	-0,5	0	0,5	1
X1: Syre, type	Nafion				TFA
X2: Temperatur/^oC	0	10	20	30	40
X3: Mulecular sieve 5Å type	Pulver				Pellets
X4: Omrøring	0				300rpm
X5: Forhold morpholine/keton / mol/ml	1,0	1,5	2,0	2,5	3,0
X6: Forhold molecular sives/keton / g/mol	200	300	400	500	600
X7: Molar kons. av keton	2,5	2,9	3,3	4,0	5,0

Tabell 6: Eksperimentelle variabler og innstillinger i enaminsyntesen.

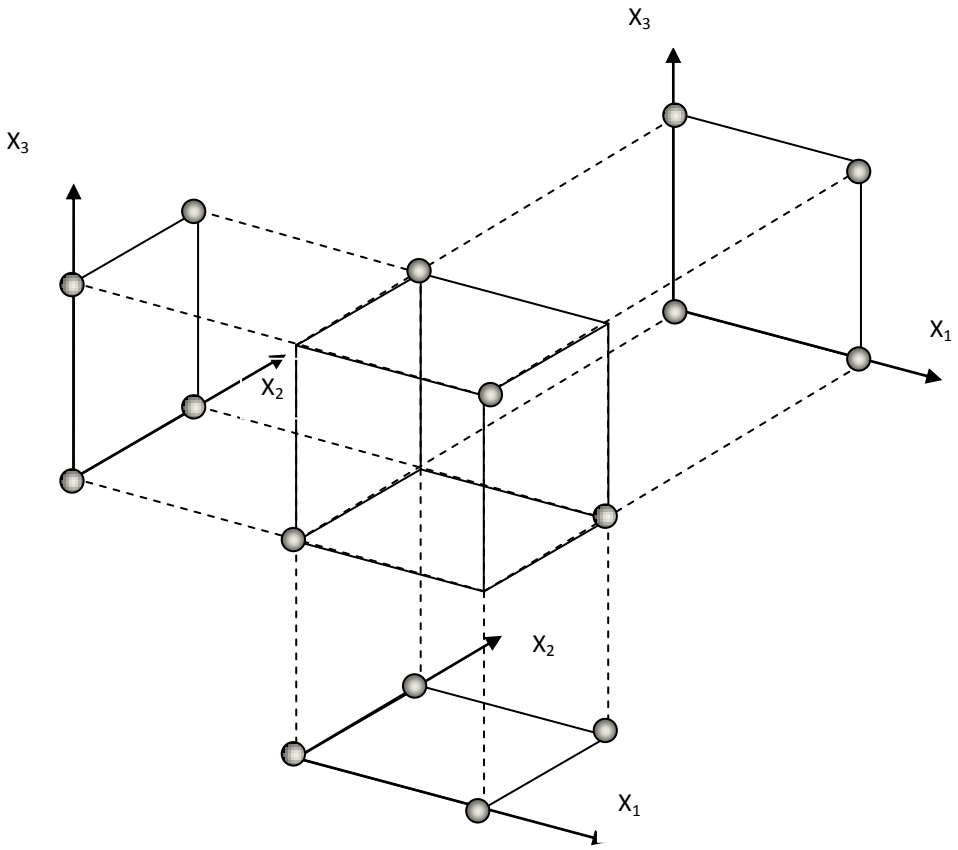
Eksperimentell design og utbytte er gitt i tabell 7.

	Variabler							Utbytte
Eksp. nr.	X_1	X_2	X_3	X_4	X_5	X_6	X_7	*Y*
1	-1	-1	-1	-1	-1	-1	1	0,4
2	1	1	1	1	1	1	1	44,2
3	-1	-1	-1	-1	1	1	1	0,9
4	-1	-1	-1	1	-1	1	1	1,7
5	1	1	1	-1	1	1	-1	51,9
6	1	0,5	1	-1	-1	-1	-1	27,0
7	1	1	1	1	-1	1	-1	30,9
8	1	1	1	-1	-1	1	1	26,3
9	-1	-1	-1	-1	1	-1	-1	0,2
10	-1	-1	-1	1	-1	-1	-1	0,2
11	-1	1	1	1	-1	-1	1	22,7
12	1	-1	-1	1	-1	-1	-1	5,2
13	-1	-1	-1	1	1	-1	1	0,8
14	-1	-1	1	-1	-1	1	-1	25,6
15	1	-1	-1	1	-1	-1	1	4,5
16	1	-1	-1	1	-1	-1	1	4,5
17	-1	1	-1	-1	-1	1	-1	4,8
18	1	-1	-1	-1	-1	1	-1	4,5
19	1	-1	-1	-1	1	-1	1	6,2
20	1	1	-1	1	1	-1	-1	27,4
21	-1	1	-1	1	-1	1	1	11,1
22	1	-1	-1	1	1	1	-1	5,6
23	-1	1	-1	1	1	1	-1	13,3
24	-1	1	1	-1	1	-1	-1	25,2
25	-1	-1	1	1	-1	-1	1	11,5
26	-1	1	1	-1	1	-1	-1	25,2
27	1	1	-1	-1	-1	-1	1	17,0
28	1	-1	1	1	1	1	1	25,4
29	-1	1	1	1	1	1	-1	43,6
30	-1	-1	-1	-1	-1	1	-1	0,4

Tabell 7: Eksperimentell design og utbytte i enaminsyntesen

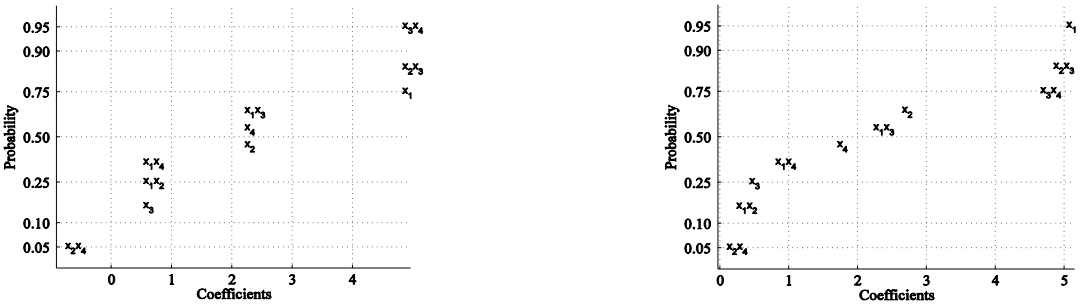
Screening

Når man undersøker mange variabler i et eksperiment er det ofte at man finner at kun et fåtall av de undersøkte variablene har en virkelig betydning. Bruker man reduserte faktorforsøk så kan man projisere de gjorte eksperimentene ned til de dimensjonene som er signifikante. Se figur 8.



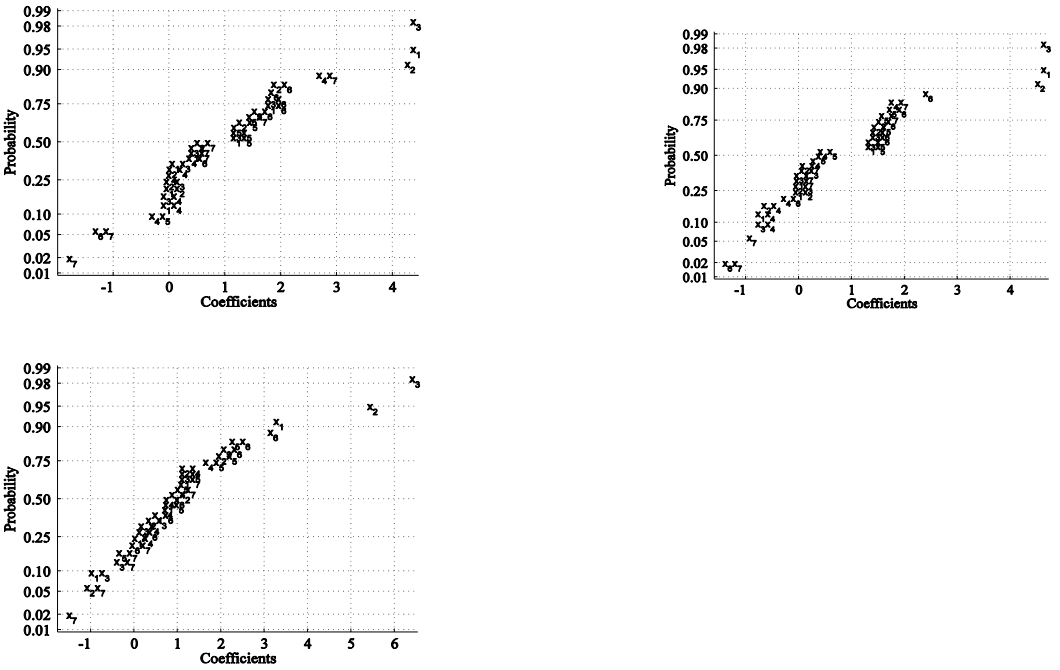
Figur 8: Reduserte faktorforsøk kan projiseres ned til et lavere dimensjons variabelrom når det viser seg at noen av variablene ikke er signifikante.

Med nær-orthogonale eksperimenter spenner hvert eksperiment opp en dimensjon i modellrommet. Er det slik at det er tilstrekkelig med kun et fåtall eksperiment for å spenne opp den signifikante variasjonen? Dette har blitt undersøkt (se ref.2]) og under oppsummeres resultatene.



Figur 9: Brominering av acetal. Kumulativ normalfordelingsplot av koeffisientene etter at 5 og 8 eksperimenter var kjørt. For ytterlige figurer se ref.2

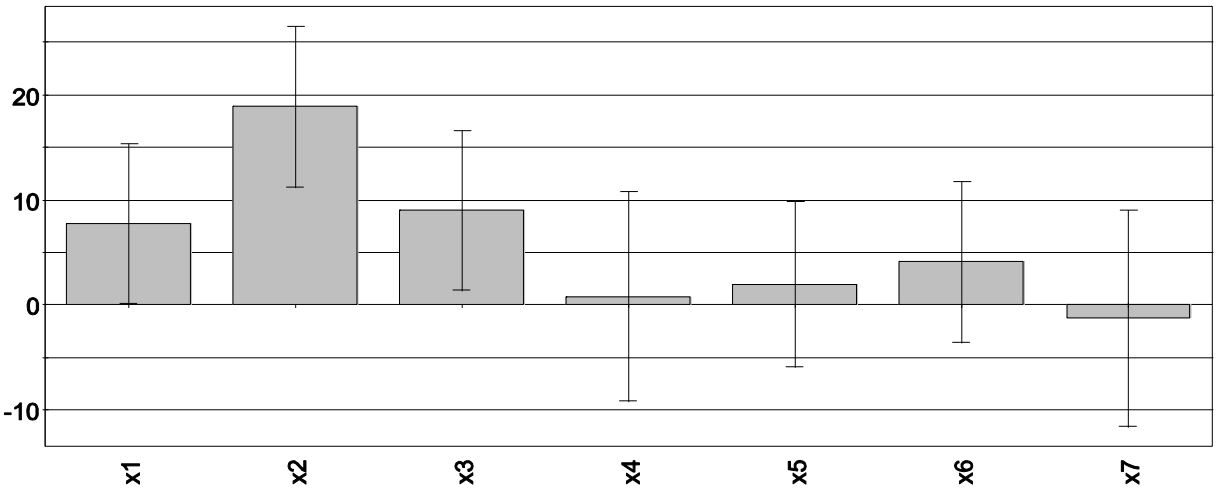
Figurene viser at variabelen X_1 i begge figurene ligger utenfor støylinja. Dette var også konklusjonen som ble presentert i ref.2.



Figur 10: Enaminsyntesen. Kumulativ normalfordelingsplot atter at 8,10 og alle eksperimenter var kjørt. For ytterlige figurer se ref.2.

Her ser vi at X_2 og X_3 og muligens også X_1 faller utenfor støylinja.

For å verifisere resultatene ble enaminsyntesen også kjørt med en 2^{7-4} *fractional factorial design* komplementert med en *fold-over design* for å få beregnet hovedeffektene fri fra to-faktor interaksjonseffekter. Se ref 2. Figur 11 viser de beregnede koeffisientene.



Figur 11: Beregnede koeffisienter i enaminsyntesen med *fractional factorial design* og *fold-over design*. Koeffisientene er fri for confoundings med tofaktor samspillseffekter.

Vi ser at X_2 og X_3 er klart signifikante og at X_1 er på grensa til å være signifikant. Det er det samme bilde som vi fikk fra PLS-modellen.

Diskusjon

7 variabler hvor hovedeffekter og 2-faktor interaksjonseffekter inngår gir et modellrom som spennes ut av 28 dimensjoner. Det betyr at en må kjøre i alt 29 eksperimenter for å kunne bestemme alle parametrene med en minste kvadrattillpasning. For å kunne bestemme modellparametrene med et begrenset antall forsøk har vi derfor brukt en PLS-modell (Projections to Latent Structures). Med PLS kan modellparametrene beregnes selv om antall forsøk er mindre enn antall parametere som skal bestemmes.

Kumulative normalfordelingsplott av koeffisientene etter at henholdsvis 8, 10, 16, 20 og alle eksperimentene var kjørt (Fig 3 i ref. 2) viser at de signifikante parametrene X_2 (temperatur) og X_3(type molecular sieve) er tydelige etter bare 8 eksperimenter. X_1 (type syre) og X_7 (molar konsentrasjon keton) sammen med noen interaksjonseffekter peker seg også tidlig ut som mulig signifikante parametere. Fra de samme normalfordelingsplottene ser vi også at X_4 (omrøring), X_5 (forhold morpholine/keton) og X_6 (forhold molecular sieves/keton) har liten betydning. Likevel kan noen interaksjonseffekter hvor noen av disse inngår ha betydning.

Tabell 6 i ref.2 viser designmatrisen og utbyttet i enaminsyntesen med *fractional factorial design* og *foldover-design*.

VII-Konklusjon

Kumulativ normalfordeling av koeffisientene både i bromineringsforsøket og enaminsyntesen viser at de viktige koeffisientene identifiseres allerede etter få forsøk. En design basert på tilnærmet orthogonale eksperimenter kan brukes som metode for å finne de viktige variablene i synteseforsøk hvor det er restriksjoner på antall forsøk som kan kjøres. Metoden gir ikke gode modellparametere, men er tilstrekkelig for å avsløre de viktige variablene, og dermed hvilke endringer som kan gjøres for å oppnå en bedre respons. Fordi metoden er sekvensiell er det mulig å avslutte forsøkene når en har fått et klart bilde av hvilke parametere som er viktige.

Fra bromineringsforsøket ser en at 5 eksperimenter av totalt 12 var tilstrekkelig for å finne de viktigste variablene. Fra enaminsyntesen ser en at 8 eksperimenter av totalt 30 var tilstrekkelig.

Metoden vil være svært nyttig når en har mange variabler og når det er klar begrensning i tilgjengelig tid. En må likevel ha klare formeninger om hvordan det eksperimentelle domenet er avgrenset. Resultatene fra oppskaleringsforsøket (brominering av acetal) styrker påstanden om at SVD-design i kombinasjon med PLS-modellering er en bra metode.

SVD-design vil være et tillegg til øvrige screeningforsøk hvor antall eksperimenter som kan gjøres er svært begrenset.

Eksperimentell del

Brominering av acetal

Programvare, kjemikalier, analyseutstyr og generell prosedyre er gitt i ref.1.

Enaminsyntesen

Programvare, kjemikalier, analyseutstyr og generell prosedyre er gitt i ref. 2.

Oppskaleringsforsøket (brominering av acetal)

Kjemikalier og utstyr er som gitt i ref.1..

Analyse I oppskaleringsforsøket ble faktisk utbytte beregnet etter vakumdestillasjon på rotavapor.

GC Analyse som gitt i ref.1 og ref.2.

NMR-analyse. ^{1}H NMR spekter ble tatt opp ved 400 MHz ved hjelp av et Varian Mercury spektrometer.

Generell prosedyre: Samme som beskrevet i ref.1. Det bør likevel nevnes at håndtering av utstyr i denne målestokken (f.eks 2000 ml skilletrakt) krever spesiell oppmerksomhet.

Referanser

[1] Artikkel 1, *Orthogonal Experiments in the evelopment of organic Synthetic Processes*

[2] Artikkel 2, *Identification of important experimental variables in organic synthetic procedures by near-orthogonal experiments.*

[3] *Fractional factorial design*

(a) G. E. P. Box; J. S. Hunter
Techometrics 3 (1961) 311 – 351.

(b) G. E. P. Box; J. S. Hunter
Techometrics 3 (1961) 449 – 458.

(c) R. Carlson; J. E. Carlson
Design and Optimisation in Organic Synthesis Amsterdam (2005) Chapter 6.

[4] *Plackett-Burman-design*

R. L. Plackett; J. O. Burman
Biometrics 33 (1946) 305 – 325

[5] D-optimal design

V. V. Fedorov
Theory of optimal Experiments, Academic Press, New York, (1972)

[6] Data program, for eksempel:

Modde 8.0. Umetrics AB, Umeå

[7] PLS regresjon

S. Wold; M. Sjøstrøm; L. Eriksson

Chemometrics Intel. Lab.Syst. 58 (2001) 109 - 130

[8] R.Carlson; J. E. Carlson

Design and Optimisation in Organic Synthesis Amsterdam (2005) pp. 110 - 115

I

Rolf Carlson, Geir Simonsen, Alexandre Descomps og Johan E. Carlson

ORTHOGONAL EXPERIMENTS IN THE DEVELOPMENT OF ORGANIC SYNTHETIC PROCESSES

Organic Process Research & Development **13** (2009), 798–803.

Organic Process Research & Development **2009,** *13,* 798–803

Summary of Lecture Transcripts

Orthogonal Experiments in the Development of Organic Synthetic Processes

Rolf Carlson,* Geir Simonsen, and Alexandre Descomps

Department of Chemistry, Faculty of Science, University of Tromsoe, NO-9037 Tromsoe, Norway

Johan E. Carlson

Department of Computer Science and Electrical Engineering, Luleå University of Technology, SE-971 87 Luleå, Sweden

Abstract:

A new strategy is presented for the design of explorative experiments in synthetic chemistry when the objective is to identify the important experimental variables. The methodology is based on Taylor expansion (response surface) models, and the principles are: A grid of possible settings of the experimental variables is laid out in the experimental domain. These experiments define a candidate design matrix, D_C. From D_C, a candidate model matrix, X_C is defined by appending columns for each variable in the Taylor model X_C is then factored by singular value decomposition (SVD), and $X_C = USV^T$. The rows in X_C that are most parallel to the singular column vectors in V are selected, and the corresponding experiments in D_C are identified. This gives the experimental design. The selected experiments are nearly orthogonal, and they span the dimensions of the model space. The experiments can be run in sequence, and thus, they allow for a systematic search, one experiment at a time. The design principles are illustrated by an example of the dibromination of an acetal. Four variables were studied, and from 12 experiments, all the main effects and all two-factor interaction effects were estimated. From the response surface model, conditions for quantitative yield were predicted, and a mol-scale synthesis carried out under these conditions afforded 98% yield of the isolated pure, >97% product.

Introduction

When a synthetic procedure is to be developed into an optimum process procedure it is often necessary to identify the important experimental variables by a screening design and then to adjust the procedure to an optimum performance by response surface modelling or some kind of gradient search. This can, however, be a tedious task that usually requires a large number of individual experimental runs, and sometimes, there is not time enough to do it.

This paper describes a strategy for designing experiments in organic synthesis when the objective is to find experimental conditions that can give improved yields. The procedure described is intended as a tool when syntheses are transformed from gram scale to hundreds of grams scale or to kilogram scale.

The strategy is based on experiments for which the variable settings in each experiment are near-orthogonal to each other. This allows for a systematic search of the experimental conditions, including also possible interaction effects. The new feature is that the experiments are run sequentially to peel off the dimensions of the search space one by one. It is therefore possible to stop the search when sufficiently good experimental conditions have been found. This is to be contrasted with factorial and fractional factorial designs for which all experimental runs must be completed before the experiment can be evaluated.

Requisites. It is supposed that the experimental procedure that has been used on gram scale affords *promising* results and the experimenter can assign which experimental variables are likely to be influential. It is also assumed that the experimenter can assign a possible operational domain and that it is believed that improved experimental conditions are likely to be found in the vicinity of the hitherto used conditions but that the knowledge of the reactions is insufficient for making any detailed predictions in this sense.

Taylor Expansion Approximation of the Response Function. The outcome y (for example the yield) of a synthetic reaction is dependent on the experimental conditions. These conditions can be specified by the settings, x_i, of the experimental variables (*temperature, concentrations, feed rates, stirring rate, etc*.). We can therefore assume that there is some kind of functional dependence between the result, y, and the experimental settings, x_1, x_2,..., x_k, and that

$$y = f(x_1, x_2, ..., x_k)$$

In most cases it is very difficult to derive an analytical expression for the function f, but if the experimental domain is not too vast, it is reasonable to assume that a truncated Taylor expansion can give a sufficiently good approximation of f, i.e.

* To whom correspondence should be addressed: E-mail: rolf.carlson@chem.uit.no.

10.1021/op800322h CCC: $40.75
Published on Web 05/07/2009

$$y = f(0) + \sum_{i=1}^{k} \frac{\partial f(0)}{\partial x_i} \times x_i + \frac{1}{2!} \sum_{i=1}^{k} \sum_{j=1}^{k} \frac{\partial^2 f(0)}{\partial x_i \partial x_j} \times x_i x_j + \text{higher order terms} + R(0) + e$$

in which $R(\mathbf{0})$ is a remainder term due to the truncation, and e is a random error term. $R(\mathbf{0})$ contains the model error due to truncation, and it becomes smaller the more terms are included in the model.

This expression is more conveniently written as a polynomial response surface model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \beta_{12} x_1 x_2 + ... + \beta_{jk} x_j x_k + \beta_{11} x_1^2 + ... + \beta_{kk} x_k^2 + e$$

To assess the roles played by the experimental variables it will be necessary to obtain estimates of the polynomial coefficients. Interaction effects are often highly significant and should be accounted for in the experimentation. In spite of an increased use of statistically designed experiments in research and production, it is still a common practice, unfortunately, to vary one experimental variable at a time. Such experiments cannot account for any interaction effect, and conclusions from such experiment are often highly erroneous. To avoid this pitfall, it is necessary to run multivariate statistical designs so that possible interaction effects can be identified. In screening experiments, when the objective is to identify the most important variables, it is often sufficient to estimate the linear effects and the two-factor interaction effects. To localise the optimum experimental conditions it is sometimes necessary also to estimate the quadratic coefficients. This is an area where traditional experimental designs (factorial designs, and fractional factorial designs,[2] D-optimal designs,[3] response surface designs[4]) are highly efficient. However, in explorative synthetic chemistry the chemists are quite reluctant to use statistical designs mainly due to the misconception that such designs will contain an excessive number of experimental runs. Still today many new methods that have been established from poor experimental designs are presented. It is in this context the near-orthogonal experiments will play their roles.

Experimental Space and Model Space. The experimenter assigns a tentative Taylor expansion model. We should now distinguish between the *experimental space* and the *model space*. The *experimental space* is defined by the possible settings of the experimental variables. With two variables, x_1 and x_2, this space is two-dimensional and with three variables it is three-dimensional, see Figure 1.

The *model space* is defined by the possible variation of the variables in the Taylor expansion model. Assume that three experimental variables are to be analysed and assume also that it is necessary to consider two-factor interaction effects. The corresponding Taylor model will be

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + e$$

and the model space in this case will be six-dimensional and spanned by $\{x_1, x_2, x_3, x_1x_2, x_1x_3, x_2x_3\}$. With a full quadratic Taylor polynomial, the model space will be nine-dimensional and spanned by $\{x_1, x_2, x_3, x_1x_2, x_1x_3, x_2x_3, x_1^2, x_2^2, x_3^2\}$.

Near-Orthogonal Experiments by SVD Design. The following iterative procedure is used to generate the experimental design:

(1) Select a set of candidate experiments that define a grid of points in the experimental domain, i.e. the space spanned by the variable axes. In our first attempts we have used 11 levels of each variable, and the sets of candidate experiments are given by the full 11-level factorial design. For two variables, the grid contains 121 candidate experiments, for three variables, 1331 candidates; for four variables 14641 candidates; for five variables, 161051 candidates; and for six variables 1771561 candidates. We assume that this gives a sufficient spread of the candidate experiments in the experimental domain. This defines the candidate design matrix $\mathbf{D}_c$.

(2) Suggest the response surface model. A candidate model matrix, $\mathbf{X}_c$, is then constructed by appending columns corresponding to each term in the model (cross-products (interaction) and squares). The columns of $\mathbf{X}_c$ define the *model space*. The matrix $\mathbf{X}_c$ is usually very large. $\mathbf{X}_c$ is then factored by singular value decomposition, SVD

$$\mathbf{X}_c = \mathbf{USV}^{\mathrm{T}}$$

The vectors in $\mathbf{U}$ and $\mathbf{V}$ are orthonormal, $\mathbf{S}$ is a diagonal matrix of the singular values, σ_i. The vectors in $\mathbf{V}$ are the eigenvectors of the variance−covariance matrix, $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, and the vectors in $\mathbf{U}$ are the eigenvectors of the correlation matrix, $\mathbf{XX}^{\mathrm{T}}$. The columns of $\mathbf{U}$ define an orthonormal basis for the column space of $\mathbf{X}_c$, and the columns of $\mathbf{V}$ define an orthonormal basis for the row space of $\mathbf{X}_c$. The singular values have the following properties: the eigenvalues of the information matrix, $\mathbf{X}_c^{\mathrm{T}}\mathbf{X}_c$ are equal to σ_i^2 and the eigenvalues to the dispersion matrix $(\mathbf{X}_c^{\mathrm{T}}\mathbf{X}_c)^{-1}$ are equal to σ_i^{-2}. Another important property is that the eigenvector in $\mathbf{V}$ corresponding to the largest singular value points in the direction of the largest variance of the row space of $\mathbf{X}_c$, i.e. the model space.

When the number of candidate experiments (rows in $\mathbf{X}_c$), is larger than the number of columns (the dimension of the model space) the maximum rank, r, of $\mathbf{X}_c$ equals the dimensions of the model space. In that case, when all singular values, σ_1,..., σ_r are distinctively different from zero, the singular vectors, $\mathbf{v}_i$, ($i = 1,..., r$) will span the model space. It was shown by Eckhart

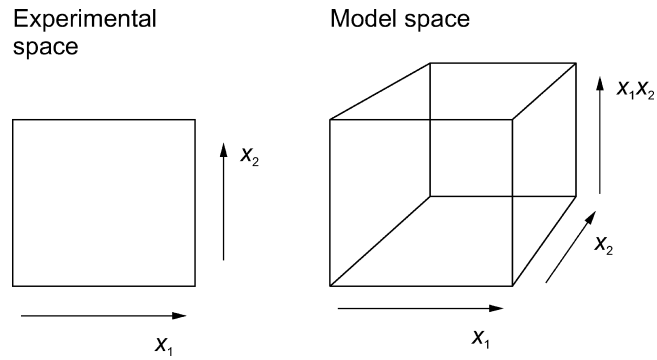


Figure 1. **Experimental space with two variables and model space with three variables.**

and Young[5] as early as in 1936 that SVD gives an optimal low-rank approximation of any matrix.

(3) The next step is to identify which row vector, $\mathbf{x}_i$, in $\mathbf{X}_c$ is most parallel to the first singular vector, $\mathbf{v}$, (i.e., corresponding to the largest singular value), in $\mathbf{V}$ as evaluated from the maximum absolute value of the scalar product $|\mathbf{x}_j\ \mathbf{v}_1|_{\mathrm{Max}}$. Then, identify which row in the candidate design matrix, $\mathbf{D}_c$, corresponds to this first selected row, $\mathbf{x}_i$, in $\mathbf{X}_c$. This yields the first experiment in the experimental design matrix. This experiment will represent a direction through the candidate design accounting for the largest variance, thus being of importance when finding a minimum set of experiments that efficiently span the variations of the model space.

(4) When the first experiment has been chosen, the next step is to remove the component in this direction from all remaining rows in $\mathbf{X}_c$. The resulting matrix, $\mathbf{X}_{c-1}$ will have the rank exactly one less than $\mathbf{X}_c$ and the corresponding rows, $_k$ are computed as

$$\hat{x}_k = \mathbf{x}_k - (\mathbf{x}_i,\mathbf{x}_k^{\mathrm{T}})/(\mathbf{x}_i,\mathbf{x}_i^{\mathrm{T}})\cdot\mathbf{x}_i$$

$\mathbf{X}_{c-1}$ is then factored by SVD and the row that is most parallel to the first singular vector is determined. The corresponding row in $\mathbf{D}_c$ is identified. This gives the second experiment in the design.

This procedure is repeated until the desired experiments have been selected. When r experiments have been selected experiments, they will span the model space.

The singular vectors, $\mathbf{v}_i$, are orthogonal, and the selected rows in $\mathbf{X}_c$ will be as orthogonal as possible. The selected experiments will thus peel off the dimensions of the model space, one experiment by one. Since the experiments are near-orthogonal, each new experiment will provide as much new information as possible. This permits a systematic search of the model space. The design is interruptible, and the experimenter can stop when a satisfactory result has been obtained. When enough experiments have been run, it is possible to fit the suggested model.

The principle for the selection of experiments is illustrated in Figure 2.

The algorithm for generating the design is illustrated in Figures 3 and 4.

We have up to now determined designs with 3, 4, and 5 variables for fitting linear, second-order interaction models, and quadratic models. The candidate experiments were defined by 11-level full factorial designs. These designs are summarised in the Appendix.

A Note on Computations. The selection procedure described above is new and has not yet been implemented in any commercial software. We have used the MATLAB software[6] for determining the design matrices. The singular vectors, $\mathbf{v}_i$ in

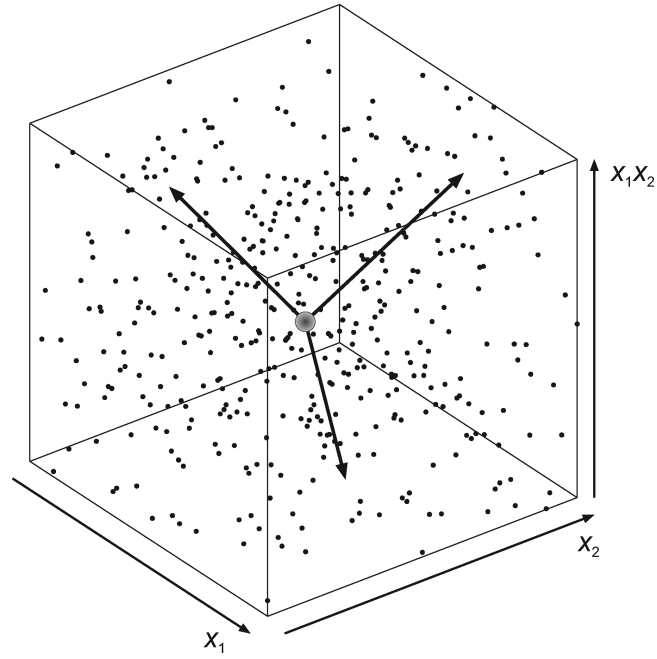


Figure 2. **Orthogonal vectors defining experiments in a three-dimensional model space.**

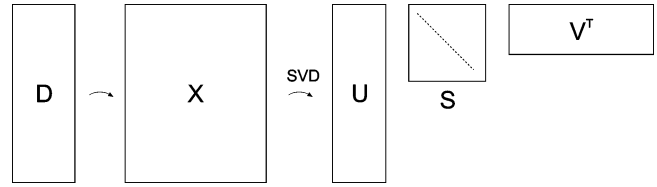


Figure 3. **Singular value decomposition of the candidate model matrix $\mathbf{X}_c$.**

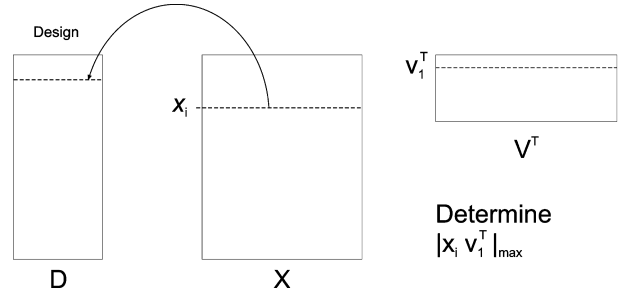


Figure 4. **Selection of experiments that are parallel to the singular vectors.**

$\mathbf{V}$ are identical to the loading vectors $\mathbf{p}_i$ obtained in principle component decomposition of a matrix $\mathbf{X}$ and $\mathbf{X} = \mathbf{TP}^{\mathrm{T}}$. The matrix $\mathbf{P}$ is defíned by the loading vectors, $\mathbf{P} = [\mathbf{p}_1\ \mathbf{p}_2 ...\ \mathbf{p}_r]$. For this reason, any commercial software that can perform principal component analysis[7] can be used to determine the singular vectors.

Distribution of the Selected Experimental Points in the Model Space. We show an example with three experimental variables. The distribution of the experimental points in the

(1) *Optimising Organic Reactions*, presented at the Scientific Update Conference, Basel, Switzerland, 29−30 October , 2007.
(2) Box, G. E. P.; Hunter, J. S.; Hunter, W. G. *Statistics for the Experimenters: Design, Innovation, and Discovery*; Wiley-Intersciences: Hoboke, NJ, 2005.
(3) (a) Nalimov, V. V.; Golikova, T. I.; Mikeshina, N. G. *Technometrics* **1970**, *12*, 799–812. (b) FedorovV. V. *Theory of Optimal Experiments*; Academic Press: New York, 1972.
(4) Box, G. E. P.; Draper, N. R. *Response Surfaces, Mixtures, and Ridge Analysis*; Wiley-Intersciences: Hoboken, NJ, 2007.
(5) Eckhart, C.; Young, G. *Psychometrika* **1936**, *1*, 211–218.
(6) *MATLAB*; The MathWorks, Inc.: Natick, MA 01760, U.S.A, 2007.
(7) Some examples of commercial software are: *SIMCA,* available from Umetrics Inc. 17 Kiel Avenue, Kinnelon, NJ 07405, U.S.A.; *Unscrambler*, available from CAMO Smart, 1480 Route 9 North Suite 209, Woodbrodge, NJ 07405, U.S.A.; *SIRIUS*, available from Pattern Recognition Systems AS, Bergen High_Tech Center, Thorm. Gt 55, NO-5008 Bergen, Norway.

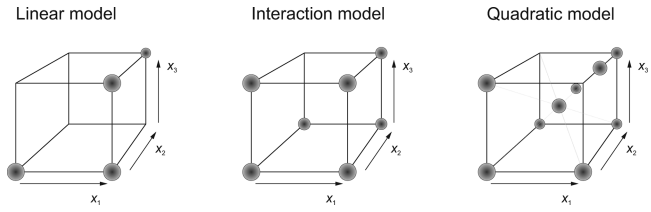


Figure 5. **Distribution of experimental points in SVD designs.**

Scheme 1

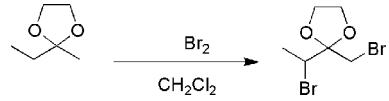


Table 1. **Experimental variables and the levels of their settings**

variables	levels of the settings −1	0	+1
x_1: reaction temperature/°C	0	1.5	30
x_2: concentration of acetal/M	0.2	0.3	0.4
x_3: stirring rate/rpm	250	325	400
x_4: rate of bromine addition/meq min^{-1}	20	50	70

Table 2. **Experimental design and yields obtained**

	design				yield
exp. no.	x_1	x_2	x_3	x_4	y
1	1.0	1.0	1.0	1.0	87.4
2	1.0	−1.0	−1.0	−1.0	95.8
3	−1.0	1.0	−1.0	1.0	79.5
4	−1.0	−1.0	1.0	−1.0	63.7
5	−1.0	−1.0	0.2	1.0	53.9
6	−1.0	1.0	1.0	−1.0	68.7
7	1.0	1.0	1.0	−1.0	58.8
8	1.0	−1.0	1.0	−1.0	93.5
9	1.0	−1.0	−1.0	1.0	94.0
10	−1.0	−1.0	−1.0	−1.0	77.1
11	1.0	−1.0	1.0	1.0	80.9
12	0	0	0	0	88.6

experimental domain of SVD designs for a linear model, an interaction model, and a quadratic model are shown in Figure 5.

From Figure 5 it is seen how such designs in this case (three variables) can be used in a sequential manner; a linear model can be fitted from four experiments. If this is unsatisfactory, an interaction model can be established by adding a few complementary experiments. A quadratic model can be established from the interaction model design by adding a few complementary experiments in the interior of the search space.

An Example: Bromination of an Acetal. We show an example of a SVD design in the bromination of the ethylene acetal from 2-butanone, see Scheme 1.

Laboratory-scale (10 mmol) experiments had afforded yields in the range 80−84%. Four variables were investigated, and their variations were chosen to embrace the hitherto known best conditions. The variables and their settings are given in Table 1. As interactions are likely, a second-order interaction Taylor model was assigned. The design and the yields obtained are given in Table 2. The experiments carried out by the design were run on larger scale (0.1−0.2 mol). The evolution of the yield was monitored by gas chromatography (internal standard technique). After 4 h the increase in yield had become insignificant, and the yields given in Table 2 were obtained after 4 h.

The second orthogonal experiment, no. 2, gave a highly increased yield compared to what was previously known as the "best" conditions. Under severe time constraint, the study could have stopped here. By using all the experiments in the design, the coefficients of the Taylor polynomial were determined using PLS regression[8], and the estimated model is

$$y = 77.71 + 8.92x_1 - 0.71x_2 - 3.11x_3 - 0.18x_4 - 6.83x_1x_2 - 1.24x_1x_3 + 2.66x_1x_4 + +0.69x_2x_3 + 6.27x_2x_4 + 1.64x_3x_4 + e$$

where e is a random error term.

The model is interpreted as follows. To increase the yield: The temperature, x_1 should be adjusted to its high level (30 °C); the concentration, x_2 should be low; the stirring rate, x_3, should be low; and the rate of addition of bromine, x_4, should be low. With these setting, the interaction effect would have a maximum beneficial influence. The predicted yield is actually 102%. We can understand the model as follows: The reaction is slightly exothermal, and to prevent unwanted temperature increase, bromine should be added slowly to the acetal at a not too high concentration. To dissipate heat from the reaction mixture, stirring is necessary, but it is probably sufficient at any level in the experimental domain. With a rapid bromine addition to a concentrated solution of the substrate, minor amounts <5% of higher brominated products were observed. A response surface projection showing the variation in yield vs x_1 and x_2 when x_3 and x_4 were set to their low level is seen in Figure 6

We have tested the suggested improved conditions in a scale-up run using 1 mol of substrate, see Experimental Section. The isolated yield was 98%, and the purity was >97% (GC, ^{1}H NMR).

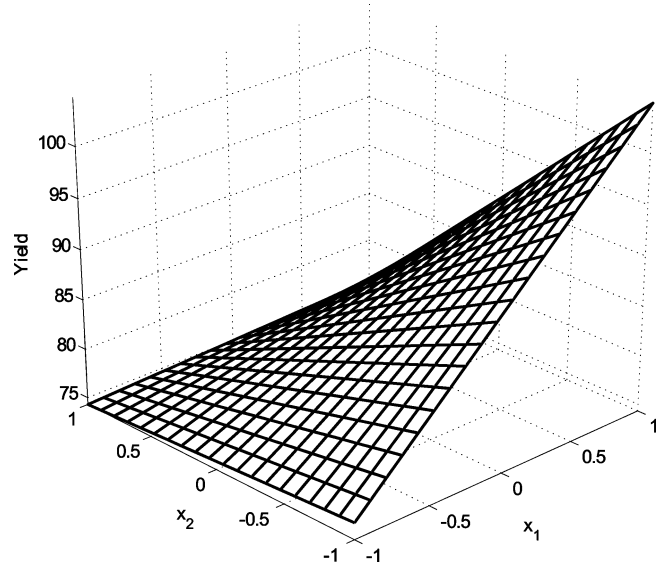


Figure 6. **Response surface projection: yield, y, vs the reaction temperature, x_1, and the initial concentration of the acetal, x_2. The stirring rate, x_3, and the rate of bromine addition, x_4, are set to their low values.**

Discussion

The experimental designs based on near-orthogonal experiments are intended as tools in explorative synthetic experimentation when the objective is to rapidly determine useful experimental conditions. Since the experimental settings in different experimental runs are nearly orthogonal to each other, the suggested strategy makes it possible to run the experiments sequentially, one by one, in order to systematically investigate the experimental space. It may well be possible that sufficiently good experimental conditions can be found after only a few experimental runs. In this respect, the designs based on orthogonal experiments are interruptible. We assume that this feature will make the suggested strategy attractive when time constraints impose limitations as to the number of possible experiments. We have previously shown that a design based on orthogonal experiments can be used for designing combinatorial libraries.[9] In this context it was demonstrated that such designs are A-Optimal: they minimise the trace of the dispersion matrix $(\mathbf{X}^T\mathbf{X})^{-1}$. If the experimental settings are adjusted exactly as specified by the singular vectors in $\mathbf{V}^T$, the designs become D-Optimal. This is possible when all variables are continuous over their range of variation and when a Taylor expansion model with only linear terms is attempted. If some variables are discrete and investigated on only two levels, ± 1, or if a higher-order model is attempted, it is unlikely that the experimental vectors can be adjusted to be parallel to the singular vectors. In such cases, the algorithm presented above can be used.

It was pointed out by one reviewer that the designs presented in this paper have inferior statistical properties compared with fractional factorial designs and D-Optimal designs. We agree with this criticism. When compared with fractional factorial designs or D-Optimal designs, the designs based on near-orthogonal model vectors have larger condition numbers, $\lambda_{Max}/\lambda_{Min}$ (the ratio of the largest and smallest eigenvalues of the dispersion matrix $(\mathbf{X}^T\mathbf{X})^{-1}$. It should, however, be borne in mind when and where an experimental design is laid out. If the objective is to fit a model with high precision in the estimated model parameters, factorial, designs, fractional factorial designs, composite response surface designs or D-Optimal designs should be used. The objective is then the model fit. If, on the other hand, the objective is to rapidly find improved experimental conditions and to have some information as to the most influencing variables, the designs based on near-orthogonal experiments are likely to be sufficiently good.

Experimental Section

Chemicals. 2-Ethyl-2-methyl-1,3-dioxoloane (99%) was obtained from Aldrich, dichloromethane (*Puriss.*), and bromine (*Puriss.*) were obtained from Merck, and 1,2-dichlorobenzene (*Puriss.*) was obtained from Fluka and used as delivered.

GC Analyses. A Varian 3400 gas chromatograph equipped with a flame ionisation detector coupled to a Varian 4400 integrator was used. The column was SPB-5, 30 m, 0.35 mm i.d., operated with the following temperature program: 70 °C, 5 min; 10 °C min^{-1}; 180 °C. The yields in the screening experiments were determined from the peak areas using 1,2-dichlorobenzene as an internal standard.

^{1}H NMR spectra were recorded at 400 MHz and ^{13}C NMR at 100 MHz using a Varian Mercury spectrometer.

General Procedure for the Screening Experiments in Table 2. The settings of the experimental variables, x_1-x_4 are given in Table 1.

The reactions were run in a four-necked 1 L mantled cylindrical reactor. The reaction temperature was controlled by circulating ethanol through the cooling mantle using a Julabo F70 thermostat. The flask was mounted with an anchor-shaped Teflon stirrer for which the stirring rate was adjusted using a Peaktech 2780 Tachometer, a reflux condenser connected to a HBr trap, a 250 mL pressure-equalised dropping funnel with a nitrogen inlet, and a temperature probe (Pt 100 sensor) dipping into the reaction mixture.

2-Ethyl-2-methyl-1,3-dioxolane (11.62 or 23.24 g, 0.10 or 0.20 mol, respectively) and an accurately weighed amount (ca. 6 g) of 1,2-dichlorobenzene (internal standard) were placed in the reactor and dissolved in dichloromethane to give the initial concentration, x_2, of the acetal.

The stirring rate was adjusted to x_3, and the temperature was adjusted to x_1. After 10−15 min, the temperature had reached the set value. Bromine (2 equiv) dissolved in 50 mL of dichloromethane was placed in the dropping funnel, and a slow stream of nitrogen was passed through the flask via the side arm of the dropping funnel. The rate of bromine addition was adjusted to x_4. Samples, 0.5 mL, were withdrawn at regular time intervals, washed with 5% aqueous sodium bisulfite, filtered through a plug of cotton, diluted with dichloromethane (2 mL) and analysed by GC. After 4 h (measured from the start of bromine addition) the changes in yields had become insignificant. These results are shown in Table 2.

Synthesis of 2(1-Bromoethyl)-2-(bromomethyl)-1,3-dioxoloane. The reactor was a four-necked, 2-L mantled cylindrical flask equipped as for the screening experiments, but using a 500 mL dropping funnel. The flask was charged with 2-ethyl-2-methyl-1,3-dioxolane (116.2 g, 1.00 mol) and 1 L of dichloromethane. The stirring rate was adjusted to 300 rpm, and the temperature was adjusted to 30 °C. When the temperature was stabilised, bromine (100 mL, 2 mol) dissolved in 250 mL of dichloromethane was added over 20 min, and the mixture was stirred at 30 °C for 4 h.

Workup: Water (300 mL) was added, and the mixture was stirred. Powdered sodium bisulfite was added carefully until the yellowish colour of unreacted bromine had disappeared. The organic layer was separated and washed a second time with 300 mL of water, and finally with 300 mL of saturated aqueous sodium bicarbonate to remove any remaining trace of Hbr. The organic layer was dried over anhydrous magnesium sulfate overnight. Filtration and evaporation of the solvent gave 267.9 g (98%) of 2-(1-bromoethyl)-2-bromomethyl-1,3-dioxolane. The product was >97% pure. The product can be further purified by distillation, bp 82 °C/13 mbar. Elemental analysis: Calcd. (C 26.30%, H 3.68%, Br 58.33%). Found (on the crude product) (C 25.71%, H 3.77%, Br 57.89%). ^{1}H NMR δ 1.69 (d, $J = 7.0$ Hz, 3H), 3.57 (d, $J = 11.1$ Hz., 1 H), 3.79 (d, $J = 11.1$ Hz, 1H), 4.13−4.18 (m, 4 H), 4.45 (q, $J = 7.0$ Hz, 1H); ^{13}C NMR δ 20.8, 35.0, 50.6, 67.7, 109.2.

(8) *Pirouette for Windows*, available from Infometrix Inc., P.O.Box 1528, Woodinville, WA 98072, U.S.A. The *MODDE 8.0* program was used. It is available from Umetrics Inc., 17, Kiel Ave, Kinnelon. NJ 07404, U.S.A.

(9) Carlson, R.; Carlson, J. E.; Grennberg, A. *J. Chemom.* **2001**, *15*, 455–474.

Appendix: A Design Matrices

A.1. Linear Models

Table A.1.1. **Three variables**

exp. no.	x_1	x_2	x_3
1	1	1	1
2	−1	−1	−1
3	1	−1	1
4	1	1	−1

Table A.1.2. **Four variables**

exp. no.	x_1	x_2	x_3	x_4
1	1	1	1	1
2	−1	−1	−1	−1
3	1	−1	1	−1
4	1	1	−1	−1
5	1	−1	−1	1

Table A.1.3. **Five variables**

exp. no.	x_1	x_2	x_3	x_4	x_5
1	1	1	1	1	1
2	−1	−1	−1	−1	−1
3	1	−1	−1	−1	−1
4	−1	1	−1	1	−1
5	1	1	−1	−1	1
6	1	−1	−1	1	1

A.2. Models with Linear Terms and Cross Terms

Table A.2.1. **Three variables**

exp. no.	x_1	x_2	x_3
1	1	1	1
2	−1	1	−1
3	1	−1	−1
4	−1	−1	1
5	−1	−1	−1
6	1	−1	1
7	1	1	−1

Table A.2.2. **Four variables**

exp. no.	x_1	x_2	x_3	x_4
1	1	1	1	1
2	1	−1	−1	−1
3	−1	1	−1	1
4	−1	−1	1	−1
5	−1	−1	0	1
6	−1	1	1	−1
7	1	1	−1	−1
8	1	−1	1	−1
9	1	−1	−1	1
10	−1	−1	−1	−1
11	1	−1	1	1

Table A.2.3. **Five variables**

exp. no.	x_1	x_2	x_3	x_4	x_5
1	1	1	1	1	1
2	−1	−1	−1	1	1
3	−1	0	1	−1	−1
4	1	1	−1	−1	−1
5	1	−1	1	−1	1
6	1	−1	−1	1	−1
7	−1	1	−1	−1	1
8	−1	1	1	1	−1
9	−1	−1	−1	−1	−1
10	1	1	−1	1	1
11	−1	−1	1	1	−1
12	−1	1	−1	1	−1
13	−1	−1	1	−1	1
14	1	1	1	1	−1
15	1	1	1	−1	1
16	1	−1	1	−1	−1

A.3. Quadratic Models, Including Linear Terms and Cross Terms

Table A.3.1. **Three variables**

exp. no.	x_1	x_2	x_3
1	1	1	1
2	1	−1	−1
3	−1	1	−1
4	−1	−1	1
5	0	0	0
6	1	1	−1
7	−1	−1	−1
8	−1	1	1
9	1	0	1
10	1	0	0

Table A.3.2. **Four variables**

exp. no.	x_1	x_2	x_3	x_4
1	−1	1	1	−1
2	−1	−1	−1	1
3	1	−1	1	1
4	1	1	−1	−1
5	0	1	0	1
6	1	−1	−1	−1
7	−1	−1	1	−1
8	−1	0	−1	−1
9	1	1	1	−1
10	−1	0	1	1
11	1	0	−1	1
12	1	1	1	1
13	−1	1	−1	1
14	−1	1	1	0
15	1	0	0	−1

Table A.3.3. **Five variables**

exp. no.	x_1	x_2	x_3	x_4	x_5
1	1	1	1	1	1
2	−1	1	1	−1	−1
3	1	−1	−1	1	−1
4	−1	−1	−1	−1	1
5	1	−1	1	−1	−1
6	−1	1	−1	1	−1
7	−1	−1	1	1	1
8	1	1	−1	−1	1
9	0	0	0	0	0
10	−1	1	1	−1	1
11	1	−1	−1	1	1
12	1	−1	1	−1	1
13	1	1	−1	−1	−1
14	1	1	1	1	−1
15	1	1	−1	1	1
16	−1	1	1	1	1
17	−1	−1	−1	1	0
18	−1	1	−1	−1	0
19	0	1	1	1	1
20	1	1	1	0	−1
21	1	0	−1	−1	−1

Received for review December 30, 2008.

OP800322H

II

Rolf Carlson, Geir Simonsen, Alexandre Descomps og Johan E. Carlson

IDENTIFICATION OF IMPORTANT EXPERIMENTAL VARIABLES IN ORGANIC SYNTHETIC PROCEDURES BY NEAR-ORTHOGONAL EXPERIMENTS

Organic Process Research & Development (submitted manuscript)

Manuscript to *Organic Process Research & Development*

The manuscript is an original research article.

Number of text pages including tables	23
Number of schemes	2
Number of figures	4
Graphical abstract	1

Title:

Identification of important experimental variables in organic synthetic procedures by near-orthogonal experiments.

Authors:

Rolf Carlson*, Geir Simonsen, Alexandre Descomps
Department of Chemistry, Faculty of Science, University of Tromsoe, NO-9037 Tromsoe, Norway

Johan E. Carlson
Department of Computer Science and Electrical Engineering, Luleå University of Technology, SE-971 87 Luleå, Sweden

* To whom correspondence should be addressed: E-mail <rolf.carlson@uit.no>

Running titles:

Carlson *et al.*
Near-orthogonal experiments for screening

Correspondence and proofs should be sent to

Rolf Carlson, see address above.

Identification of important experimental variables in organic synthetic procedures by near-orthogonal experiments

Abstract

A new strategy is presented for the design of screening experiments in synthetic chemistry when the objective is to identify the important experimental variables. The methodology is based on Taylor expansion (response surface) models. The principles are: a grid of possible settings of the experimental variables is laid out in the experimental domain. These experiments define a candidate design matrix, $\mathbf{D}_C$. From $\mathbf{D}_C$, a candidate model matrix, $\mathbf{X}_C$ is defined by appending columns for each variable in the Taylor model. The matrix $\mathbf{X}_C$ is then factored by Singular Value Decomposition (SVD), and $\mathbf{X}_C = \mathbf{U}\,\mathbf{S}\,\mathbf{V}^T$. The rows in $\mathbf{X}_C$ that are most parallel to the singular column vectors in $\mathbf{V}$ are selected, and the corresponding experiments in $\mathbf{D}_C$ are identified. This gives the experimental design. The selected experiments are nearly orthogonal and they span the dimensions of the model space. The experiments can be run in sequence and thus they allow for a systematic search, one experiment at a time. It is shown that subset selections from such designs in combination with PLS modelling can be used to identify the important variables. The principles are illustrated with two examples: (a) a dibromination of an acetal with four experimental variables, and (b) a synthesis of an enamine by condensing a ketone and morpholine in the presence of molecular sieves in which seven experimental variables are involved. In the acetal bromination, it was found that five experiments out of twelve were sufficient for identifying the most important variables. In the enamine example, eight experiments out of thirty were sufficient.

Introduction

When an experimental procedure is to be developed into a reliable *method*, an early and important step is to identify the critical experimental factors as well as their possible interaction effects. To this end, a variety of different statistical experimental designs are available: Factorial and fractional designs[1], D-Optimal designs[2], Plackett-Burman designs[3]. The use of such designs in organic synthesis is thoroughly described in Ref. 4.

There are, however, situations in which severe time-constraints preclude any attempt to run a screening design with many experimental runs. Two examples are: (1) A new compound turned out to have interesting pharmaceutical properties. For more testing, 200 g of the compound is needed within four weeks. The testings are expensive and no delay can be tolerated. The chemists have to produce the necessary quantity within the time limits.
(2) Outsourcing is now very common to produce the active ingredients in drugs. A chemical company is contracted by the customer to make some test experiments of a given procedure and to deliver 200 g of the desired compound. The time limits are strict and it is not possible to run more than a handful of tests. Common to this problems is that the chemist should run a reaction that is known and has already been used to make small quantities of the desired compound. It can therefore be assumed that a useful experimental domain is known (i.e. the possible ranges of variation of the experimental factors). It can also be assumed that improved results can be obtained in the vicinity of the known experimental conditions.

Under these circumstances it is reasonable to assume that the observed response, y, can be modelled by a truncated Taylor expansion in the scaled experimental variables, x_i, centred around the known experimental conditions and that

$$y = \beta_o + \Sigma\, \beta_i\, x_i + \Sigma\, \Sigma\ \beta_{ij}\, x_i\, x_j + e$$

in which β_0 is the intercept of the response model at the centre point of the experimental domain, and β_i and β_{ij} are the values of the partial derivatives along the variable axes at the centre point. Least squares estimates (b_0, b_i, and b_{ij}) of the Taylor coefficients can be obtained by fitting the polynomial to the experimental results obtained by a proper design.
We have previously shown in This Journal[5] that response surface models can be established from designs constructed in such a way that the rows the model matrix are nearly orthogonal. The construction of such designs is described in Ref. 5 and we will not repeat these details here. The essence of these designs is that each new experiment selected for the design spans a new dimension of the model space, i.e. the space spanned by the variables in the Taylor polynomial, and that it is possible to investigate the roles played by the variables and their interactions by a sequential approach in which the experiments are added one by one until the variations in the models space have been mapped.

In a screening, the task is to determine which experimental variables have a real influence on the result. It is often the case that out of many variables initially considered to be potentially important, there are only a few of them that really matter. A discussion of this is given in Ref. 6

It came to our mind that a design based on near-orthogonal experiments might be useful in a screening situation. The reasons are the following: The very first experiment in such a design describes the direction showing the largest variation of the variable settings in the model space. The important variables will exert their influence in this experiment. The second experiment is near-orthogonal to the first one and the important variables will influence in this experiment too, *but in a different way*. If there are only a handful of important variables, it might be possible that these can be identified from a handful of experimental runs. In this paper, we show two examples along these principles.

Examples

The first example is the bromination of an acetal fully described in Ref. 5, The reaction is portrayed in Scheme 1. The variables explored and the design are given in Tables 1 and 2, respectively.

Scheme 1 is placed here.

Tables 1 and 2 are placed here

The second example is the synthesis of an enamine by a condensation between 4-methyl-2-pentanone and morpholine in the presence of molecular sieves, see Scheme 2.

Sceme 2 is placed here

The variables explored and their settings are shown in Table 3. The experimental design and the yields obtained are shown in Table 4. There are three discrete variables at two levels and four variables at five levels. The candidate experiments were defined by the full $2^3 * 5^4$ full factorial design with a total of 5 000 runs. The design was expanded to the candidate model

matrix by appending columns of the cross-product terms. The design was then generated by singular value decomposition as described in Ref. 5.

Tables 3 and 4 are placed here

Data

It was assumed that second order interaction models would be sufficent for describing the variation yield, *y*, as functions of the experimental settings, x_i and the interactions, $x_i\, x_j$:

$$y = \beta_0 + \Sigma\beta_i x_i + \Sigma\Sigma\, \beta_{ij}\, x_i\, x_j + e \quad (i \neq j)$$

where *e* is a random error term.

In the bromination example there are four variables and 11 unknown parameters in the model. In the enamine synthesis, there are seven variables and 29 unknown parameters in the model.. For these reasons, the model spaces will have 11 ad 29 dimensions, respectively and the corresponding designs for fitting the Taylor polynomial by least squares multiple regression must have at least 11 and 29 experimental runs, respectively.

Data analysis

We wished to know whether or not a limited number of experimental runs would be sufficient for identifying the important variables. Hence, the number of experiments will be less than the number of coefficients in the model and it will not be possible to estimate the coefficient by least-squares multiple regression. Instead, we have used PLS-modelling to estimate the coefficients. (PLS is an acronym for Projections to Latent Structures.) This is possible since PLS is based on projections. PLS is a computational method by which it is possible to obtain quantitative relations between a matrix of independent variables, **X** (X-block) and the corresponding matrix **Y** of the response variable(s), (Y-block), $\mathbf{Y} = [\mathbf{y}_1 \quad \mathbf{y}_2 \ldots \mathbf{y}_m]$, If there is only one response to consider, the Y-block is a vector, **y**. Thorough treatments

of PLS modelling are given in Ref. 7. Here follows only a brief summary of PLS and how the coefficients in the models were estimated.

Computations

For each of the design matrices given in Tables 2 and 4, the corresponding model matrix, $\mathbf{X}_0$ was constructed by appending columns of the cross-product terms in the models. From the matrix $\mathbf{X}_0$, the designs are generated by the procedure described in Ref 5. The selected designs are then converted to the corresponding X-block matrix, $\mathbf{X}_1$, by mean centring. i.e. the average of each columns is subtracted from the elements of that column so that. $\mathbf{X}_1$ describes the variations of the variable settings around the average point. The response vector, **y**, was also mean centred.

Define the first weight vector, $\mathbf{w}_1$, as

$$\mathbf{w}_1 = \mathbf{X}_1^T \mathbf{y} / \| (\mathbf{X}_1^T \mathbf{y})^T (\mathbf{X}_1^T \mathbf{y}) \|$$

i.e. the normalised projection of $\mathbf{X}_1$ onto **y**.

Another way of defining **w** is as follows.

Let

$$\mathbf{C}_{YX} = \mathbf{y}^T \mathbf{X}_1.$$

The cross variance-covariance matrix in then

$$\mathbf{C}_{YX}^T \mathbf{C}_{YX} = \mathbf{X}_1^T \mathbf{y}\, \mathbf{y}^T \mathbf{X}_1.$$

Determine the first eigenvector $\mathbf{w}_1$ of the cross variance-covariance matrix, i.e the eigenvector corresponding to the largest eigenvalue. Normalise $\mathbf{w}_1$. The first weight vector describes the direction through the model space in which the projected experimental points have the maximum correlation with the response. This vector is defined by the cosines of the angle between the vector and the axes of the models space, see Figure 1.

Figure 1 is placed here.

From $\mathbf{w}_1$ and $\mathbf{X}$, the first score vector, $\mathbf{t}_1$ is computed by

$$\mathbf{t}_1 = \mathbf{X}_1 \mathbf{w}_1$$

and the corresponding loading vector, $\mathbf{p}_1$, of the X-block is computed as the projection of $\mathbf{X}_1$ onto $\mathbf{t}_1$

$$\mathbf{p}_1 = \mathbf{X}_1^T \mathbf{t}_1 / \mathbf{t}_1^T \mathbf{t}_1$$

The corresponding loading vector, $\mathbf{q}_1$ of the Y-block (in the present case the response vector $\mathbf{y}$) is computed as

$$\mathbf{q}_1 = \mathbf{Y}^T \mathbf{t}_1 / \mathbf{t}_1^T \mathbf{t}_1$$

This gives the following approximations of the X-bock and the Y-block:

$$\mathbf{X}_1 = \mathbf{t}_1 \mathbf{p}_1^T + \mathbf{E}_1$$

$$\mathbf{Y}_1 = \mathbf{t}_1 \mathbf{q}_1^T + \mathbf{F}_1$$

where $\mathbf{E}_1$ and $\mathbf{F}_1$ are the matrices of the residuals of the X-block and the Y-block, repspectively, i.e the variation that is not described by the first component..
To find the next PLS component, the procedure is repeated using $\mathbf{E}_1$ and $\mathbf{F}_1$ as starting matrices. This is repeated until the significant number, *A*, of components has been found. This can be determined by cross validation.[8] The scores and the loadings of the X- and the Y-blocks are stored as columns in the matrices

$$\mathbf{T} = [\mathbf{t}_1 \ \mathbf{t}_2 \ ... \ \mathbf{t}_A]$$

$\mathbf{P} = [\mathbf{p}_1 \ \ \mathbf{p}_2 \ldots \ \mathbf{t}_a$

$\mathbf{Q} = [\mathbf{q}_1 \ \ \mathbf{q}_2 \ldots \ \mathbf{q}_A]$.

The weight vectors are stored in the weight matrix

$\mathbf{W} = [\mathbf{w}_1 \ \ \mathbf{w}_2 \ldots \ \mathbf{w}_A]$.

The original matrices $\mathbf{X}_1$ and $\mathbf{Y}$ can now be expressed in terms of the PLS component as:

$$\mathbf{X}_1 = \mathbf{T}\,\mathbf{P}^T + \mathbf{E}$$

and

$$\mathbf{Y} = \mathbf{T}\,\mathbf{Q}^T + \mathbf{F}$$

where $\mathbf{E}$ and $\mathbf{F}$ are the residuals from the X- and Y-block, respectively.

Given the sets of PLS components it is now possible to make predictions of $\mathbf{Y}$ from the experimental settings, $\mathbf{X}_1$, and

$$\mathbf{Y}_{\text{Predicted}} = \mathbf{X}_1\,\mathbf{W}^*\,\mathbf{Q}^T$$

where

$$\mathbf{W}^* = \mathbf{W}\,(\mathbf{P}^T\,\mathbf{W})^{-1}.$$

Estimation of the Taylor coefficients

The Taylor polynomial that relates the experimental settings to the response is

$$y = \beta_o + \Sigma\,\beta_i\,x_i + \Sigma\,\Sigma\ \beta_{ij}\,x_i\,x_j + e$$

in which β_0 is the response at the centre point. To fit the model, an experimental design is used and if the experimental points are uniformly distributed around the centre point, an estimate b_0 of β_0 will be the average of the responses in the experiments and the average response will be equal to the response in the average experimental point. Let **y** be the mean centred response vector, $\mathbf{y} = [\, \mathbf{y}_{\text{Observed}} - \mathbf{y}_{\text{Average}} \,]$. From an estimate $\mathbf{b} = [b_1 \;\; b_2 \ldots \; b_{ij} \;]^T$ of the vector of model parameters $\boldsymbol{\beta} = [\beta_1 \;\; \beta_2 \ldots \; \beta_{ij} \;]^T$, the predicted response in the experimental point will be

$$\mathbf{y}_{\text{Predicted}} = \mathbf{X}_1 \, \mathbf{b}.$$

and from PLS the predicted response is

$$\mathbf{y}_{\text{Predicted}} = \mathbf{X}_1 \, \mathbf{W}^* \, \mathbf{Q}^T.$$

This gives an estimate of the coefficients in the Taylor polynomial that will be obtained by

$$\mathbf{b} = \mathbf{W}^* \, \mathbf{Q}^T.$$

We have used this relation to obtain estimates of the Taylor polynomial from the selected experiments. It is not possible, however, to obtain accurate estimates when the number of experiments is lower than the number of parameters to be estimated, the estimates will be biassed. If the average response is different from the constant in the Taylor polynomial, i.e. the response in everage experimental point and the difference is $d = \beta_0 - y_{\text{Average}}$ we have to adjust for this bias prior to fitting the model, and

$$\text{E}\,[\mathbf{b}] = \boldsymbol{\beta} + (\mathbf{X}^T \, \mathbf{X})^{-1} \, \mathbf{X}^T \, \mathbf{d}$$

in which **d** is the bias vector added to the observed responses.

The objective of the present invesitgation is to evaluate if a limited number of experiments can be used to identify the important variables. In this context, even a biassed estimate will be indicative.

Results and discussion

The results of the PLS modelling are summarised in Table 5.

Bromination of the acetal: We have analysed the roles played by the variables by the cumulative normal probability distribution plots of their coefficients.[9] Figure 2 shows the plots obtained when five, six, and all experiments, respectively, were used to establish the the PLS model. It is clearly seen that one variable, x_1, is visible in all plots as an outlier from the noise line. This indicates that the reaction temperature is an important variable to control. The other variables have only a minor importance. This was also the conclusion presented in Ref. 5.

Fig 2 is placed here.

Enamine synthesis: This is a more complicated system. Figure 3 shows the cumulative normal probability distributions of the estimated coefficients obtained when eight, ten, sixteen, twenty, and all experiments, respectively, were used to establish the PLS model.

Fig 3 is placed here.

With all experiments included, two variables, x_2 (reaction temperature) and x_3 (type of molecular sieve) are clearly significant. With fewer experiments included in the X-block in the PLS models, the same variables were also indicated as significant. With fewer experiments, also x_1 (the type of acid) and x_7 (the molar concentration of the ketone) show up as possibly significant variables. Some interactions also show up as possibly important.

A screening experiment is carried out to identify which variables among many possible are likely to have a significant influence on the result. These variables should then be more carefully studied in subsequent experiments. If some variables found to be significant in the

first screening should turn out to have only minor influences in the follow-up experiments, no harm is done and these variables can then be safely removed from further considerations. It is much worse if important variables are overlooked.
The coefficients estimated with PLS are biassed, see above. However, even a biassed estimate contains information. The objective is not to estimate a response surface model with high precision, but to discern which variables are likely to be important. And for this the PLS estimates will be sufficient.

To verify the results obtained, by PLS, the variables in the enamine example were also studied by a 2^{7-4} fractional factorial design (**I** = **12** = **13** = **23** = **123**) complemented with a fold-over design ((**I** = −**12** = −**13** = −**23** = **123**) to estimate the linear coefficients free of confounding with the two-factor interactions. The design in shown in Table 6 and the estimated coefficients in Figure 4.

Table 6 is placed here

Figure 4 is placed here.

It is seen in Fig. 4 that variables x_2 and x_3 are clearly significant and that x_1 is on the border to be significant. This is actually what is seen in Figure 3.

Conclusions

If the objective is to fit a response model with high precision then, of course, it is necessary to use a statistical design that permits accurate estimations of the model parameters. However, such designs often contain a fairly large number of individual experimental runs and sometimes this precludes their use. Under such circumstances, we have suggested that a design based on near-orthogonal experiments can be useful and in this paper we have shown that sets of near-orthogonal experiment make it possible to discern the important experimental

variables from only a few runs. As the experiments are run sequentially, one by one, it is possible to run the number of experiments necessary to obtain a clear picture. We assume that this will be an appealing technique in the realm of process chemistry.

Experimental

Computations

The experimental designs were generated in MATLAB.[10] The PLS models were obtained with the SIMCA P-11 software.[11] The fold-over design was evaluated using the MODDE-8 software.[12]

Chemicals

Morpholine (puriss.) was obtained from FLUKA. 4-Methyl-2-pentanone (HPLC grade). Cyclohexane (99.5%), phenylcyclohexane (puriss.) internal standard for GC were obtained from Aldrich.. They were used as delivered. Molecular sieves, 5A powder and pellets were obtained from FLUKA. They were activated at 300 ^{o}C for 24 h prior to uses and stored in a desiccator over phosphorus pentoxide. A reference sample of the morpholine enamine from 4-methyl-2-pentanone used for GC-calibration was prepared according to Ref. 13.

GC Analyses

Enamine synthesis. A Varian 3400 gas chromatograph with a flame ionisation detector coupled to a Varian 4400 integrator was used. The columns was SPB-5, 30m, 0.35 mm i.d. operated with the following temperature program: 70 ^{o}C, 5 min; 10 ^{o}C min^{-1}; 180 ^{o}C.. The yields were determined from the integrated peak areas using phenylcyclohexane as internal standard.

General procedure for the screening experiments, bromoacetal synthesis.

The experimental procedure for the bromination of the acetal is given in Ref. 5 and it is not reproduced here.

General procedure for the screening experiments, enamine synthesis.

The settings of the variables are shown in Table 3. The experiments were run in 50 mL test tubes using a heating bloch reactor system Bohdane 2080 Miniblock™ from Mettler Toledo. In the experiments, 5 mmol of the ketone, 4-methyl-2-pentanone was used. The test tube was charged with the ketone, the amount x_5 of the molecular sieves of type x_3, the amount x_5 of morpholine, a carefully weighed amount, ca.200 mg, of phenylcyclohexane (internal standard), and 20 mg of the acid x_1. The calculated amount of cyclohexane solvent to give the concentation x_7 was the added to the test tube The temperature, x_2, and the stirring, x_4 were adjusted. The reaction was monitored by gas chromatography. Samples, 0.1 mL, were withdrawn, filtered through a plug of cotton, diluted with 2 mL of pentane, and analysed by GC. Integrated peak areas were used for quantification. The yields obtained after 24 h are shown in Table 4.

Acknowledgement

We thank the American Chemical Society for the kind permission to reproduce details from Ref. 5. We also thank the Norwegian Research Council for financial support via the KOSK-II program.

References

1. (a) Box, G. E. P.; Hunter, J. S. *Technometrics* **1961**, *3*, 311–351;
(b) Box, G. E. P.; Hunter, J. S. *Technometrics* **1961**, *3*, 449–458;
(c) Box, G. E. P.; Hunter, J. S. ; Hunter W. G. *Statistics for Experimenters. Design, Innovation, and Discovery, 2nd Edition*; Wiley, Hoboken NJ, 2005.

2. (a) Nalimov, V. V., Golikova, T. I.; Mikeshina, N. G. *Technometrics* **1970**, *12*, 799–812;
(b) Fedorov, V. V. *Theory of Optimal Experiments*, Academic Press, New York 1972.

3. Plackett, R. L.; Burman, J. P. *Biometrika* **1946**, *33*, 305–325.

4. Carlson, R.; Carlson, J. E. *Design and Optimization in Organic Synthesis, Second Revised and Enlarged Edition*; Elsevier, Amsterdam, 2005.

5. Carlson, R.; Simonsen; G.; Descomps, A.; Carlson, J. E. *Organic Process Research & Development* **2009**, *13*, 798–803.

6 For comments on this, see: (a) Box, G. E. P.; Hunter, J. S. ; Hunter W. G. *Statistics for Experimenters. Design. Innovation, and Discovery*, 2nd Edition; Wiley, Hoboken NJ, 2005, pp. 243–244; (b) Box, G. E. P.; Draper, N. R. *Response Surfaces, Mixtures, and Ridge Analysis, 2nd Edition*; Wiley, Hoboken NJ, 2007, pp 148–150.

7. (a) Wold, S.; Sjöström, M.; Eriksson, L. *Chemometrics Intell. Lab. Syst.* **2001**, *58*, 109–130:
(b) Carlson, R.; Carlson, J. E. *Design and Optimization in Organic Synthesis, Second revised and Enlarged Edition*; Elsevier, Amsterdam, 2005, Chapter 18.

8. Eriksson, L.; Johansson, E.; Kettaneh-Wold. N.; Wold, S. *Multi- and Megavariate Data Analysis: Principles and Applications*, Umetrics, Umeå, Sweden 2001.

9. Daniel, C. *Application of Statistics to Industrial Experimentation*, Wiley New York 1076.

10. (a) *MATLAB*, The MathWorks, Inc.: 3 Apple Hill Drive, Natick, MA 01760, U.S.A. 2007; (b) For the generation of designs based on near-orthogonal experiments, see NO_Designs: http://jec-solutions.mine.nu.

11 *SIMCA-P-11*, Umetrics Inc./ MKS Instuments, 70 W. Rio Robles Drive, San José, CA 95134, U.S.A.

12 *MODDE-8*, Umetrics Inc./ MKS Instuments, 70 W. Rio Robles Drive, San José, CA 95134, U.S.A.

13. Carlson, R.; Nilsson, Å.; Strömqvist. M. *Acta Chem. Scand. Ser. B* **1983**, *37*, 7–13.

Tables

Table 1: Experimental variables in the bromination of the acetal and the levels of their settings.[a]

Variables	Levels of the settings		
	−1	0	+1
x_1: Reaction temperature/ °C	0	1 5	30
x_2: Concentration of acetal/M	0.2	0.3	0.4
x_3: Stirring rate/rpm	250	325	400
$x_{4:}$ Rate of bromine addition/meq min^{-1}	20	50	70

[a] Reproduced with permission from the American Chemical Society.

Table 2: Experimental design and yields obtained in the bromination of the acetal.[a]

	Design				Yield
Exp #	x_1	x_2	x_3	x_4	y
1	1.0	1.0	1.0	1.0	87.4
2	1.0	−1.0	−1.0	−1.0	95,8
3	−1.0	1.0	−1.0	1.0	79.5
4	−1.0	−1.0	1.0	−1.0	63.7
5	−1.0	−1.0	0.2	1.0	53.9
6	−1.0	1.0	1.0	−1.0	68.7
7	1.0	1.0	1.0	−1.0	58.8
8	1.0	−1.0	1.0	−1.0	93.5
9	1.0	−1.0	−1.0	1.0	94.0
10	−1.0	−1.0	−1.0	−1.0	77.1
11	1.0	−1.0	1.0	1.0	80.9
12	0	0	0	0	88.6

[a] Reproduced with permission from the American Chemical Society.

Table 3: Experimental variables and their settings in the enamine synthesis.

Variables	Settings				
	-1	-0.5	0	0.5	1
x_1: Type of acid	Nafion®				TFA
$x_{2:}$ Temperature / ^{O}C	0	10	20	30	40
x_3: Type of moleculae sieve 5A	Powder				Pellets
x_4: Stirring	None				300 rpm
x_5: Ratio morpholine/ketone / mol/ml	1.0	1.5	2.0	2.5	3.0
x_6: Ratio molecular sives/ketone / g/mol	200	300	400	500	600
x_7: Molar concentration of ketone	2.5	2.9	3.3	4.0	5.0

Table 4: Experimental design and the yields obtained in the enamine synthesis.

	Variables							Yield
Exp. no.	x_1	x_2	x_3	x_4	x_5	x_6	x_7	y
1	−1	−1	−1	−1	−1	−1	1	0.4
2	1	1	1	1	1	1	1	44.2
3	−1	−1	−1	−1	1	1	1	0.9
4	−1	−1	−1	1	−1	1	1	1.7
5	1	1	1	−1	1	1	−1	51.9
6	1	0.5	1	−1	−1	−1	−1	27.0
7	1	1	1	1	−1	1	−1	30.9
8	1	1	1	−1	−1	1	1	26.3
9	−1	−1	−1	−1	1	−1	−1	0.2
10	−1	−1	−1	1	−1	−1	−1	0.2
11	−1	1	1	1	−1	−1	1	22.7
12	1	−1	−1	1	−1	−1	−1	5.2
13	−1	−1	−1	1	1	−1	1	0.8
14	−1	−1	1	−1	−1	1	−1	25.6
15	1	−1	−1	1	−1	−1	1	4.5
16	1	−1	−1	1	−1	−1	1	4.5
17	−1	1	−1	−1	−1	1	−1	4.8
18	1	−1	−1	−1	−1	1	−1	4.5
19	1	−1	−1	−1	1	−1	1	6.2
20	1	1	−1	1	1	−1	−1	27.4
21	−1	1	−1	1	−1	1	1	11.1
22	1	−1	−1	1	1	1	−1	5.6
23	−1	1	−1	1	1	1	−1	13.3
24	−1	1	1	−1	1	−1	−1	25.2
25	−1	−1	1	1	−1	−1	1	11.5

Table 4: (continued)

26	−1	1	1	−1	1	−1	−1	25.2
27	1	1	−1	−1	−1	−1	1	17.0
28	1	−1	1	1	1	1	1	25.4
29	−1	1	1	1	1	1	−1	43.6
30	−1	−1	−1	−1	−1	1	−1	0.4

Table 5: Summary of PLS results

Reaction system	Experiments	PLS components	R^2	Q^2
Acetal	5	2	1.00	0.726
	6	l	0.898	0.382
	12	1	0.808	0.269
Enamine	8	2	0.894	0.572
	10	1	0.933	0.449
	16	1	0.841	0.545
	20	1	0.858	0.565
	30	1	0.890	0.560

Table 6: Fractional factorial design and fold-over in the enamine synthesis.

	Variables							Yield
Exp. no.	x_1	x_2	x_3	x_4	x_5	x_6	x_7	y
1	-1	-1	-1	1	1	1	-1	0.1
2	1	-1	-1	-1	-1	1	1	10.9
3	-1	1	-1	-1	1	-1	1	17.6
4	1	1	-1	1	-1	-1	-1	40.5
5	-1	-1	1	1	-1	-1	1	11.9
6	1	-1	1	-1	1	-1	-1	10.9
7	-1	1	1	-1	-1	1	-1	50.3
8	1	1	1	1	1	1	1	94.4
Fold-over								
9	-1	-1	-1	-1	-1	-1	-1	0.1
10	1	-1	-1	1	1	-1	1	8.9
11	-1	1	-1	1	-1	1	1	23.6
12	1	1	-1	-1	1	1	-1	50.3
13	-1	-1	1	-1	1	1	1	14.9
14	1	-1	1	1	-1	1	-1	15.1
15	-1	1	1	1	1	-1	-1	43.5
16	1	1	1	-1	-1	-1	1	61.4

Captions to figures

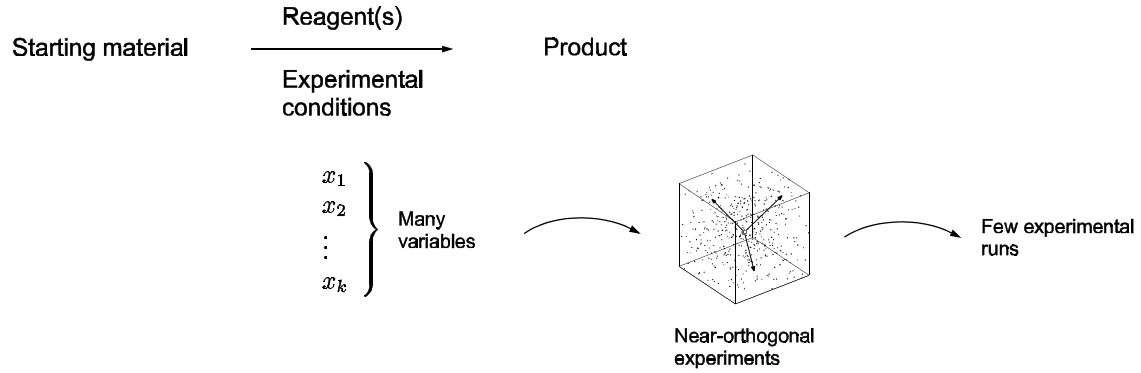
Figure 1: The direction of the weight vector, $\mathbf{w}_1$ in the model space.

Figure 2: Acetal bromination: Cumulative normal probability plots of estimated coefficients from: (a) five experiments; (b) six experiments; (c) twelve experiments.

Figure 3: Enamine synthesis: Cumulative normal probability distribution plots of estimated coefficients from: (*a*) eight experiments; (*b*) ten experiments; (c) sixteen experiments; (d) twenty experiments; (e) thirty experiments.

Figure 4: Enamine synthesis. Estimated linear coefficients from the fold-over design.

Graphical abstract



Figures

Fig.1

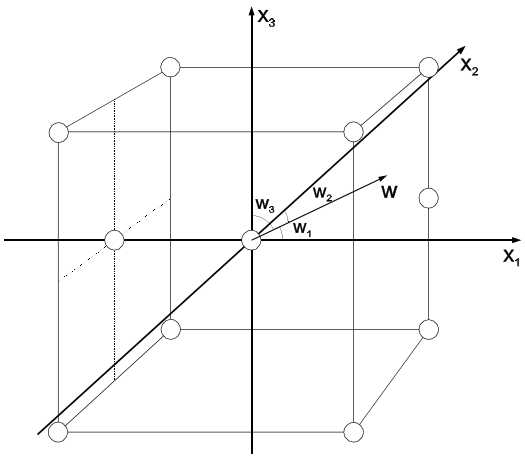
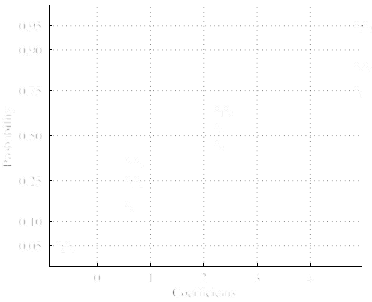
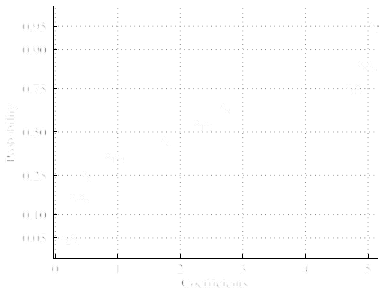


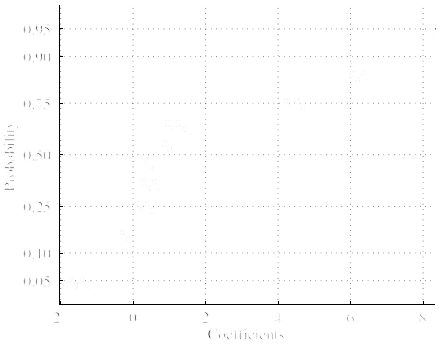
Fig 2



a

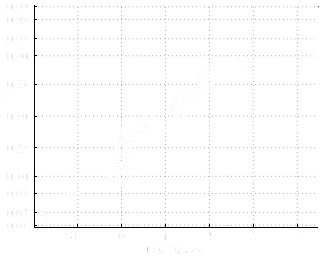


b

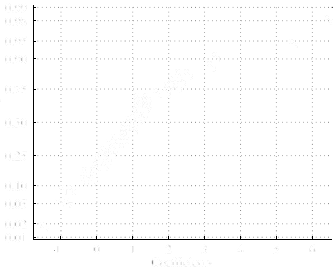


c

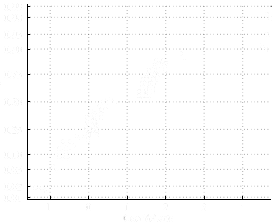
Fig 3



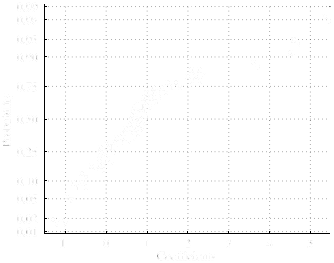
a



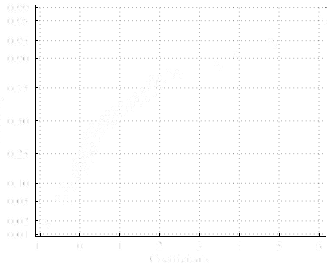
e



b

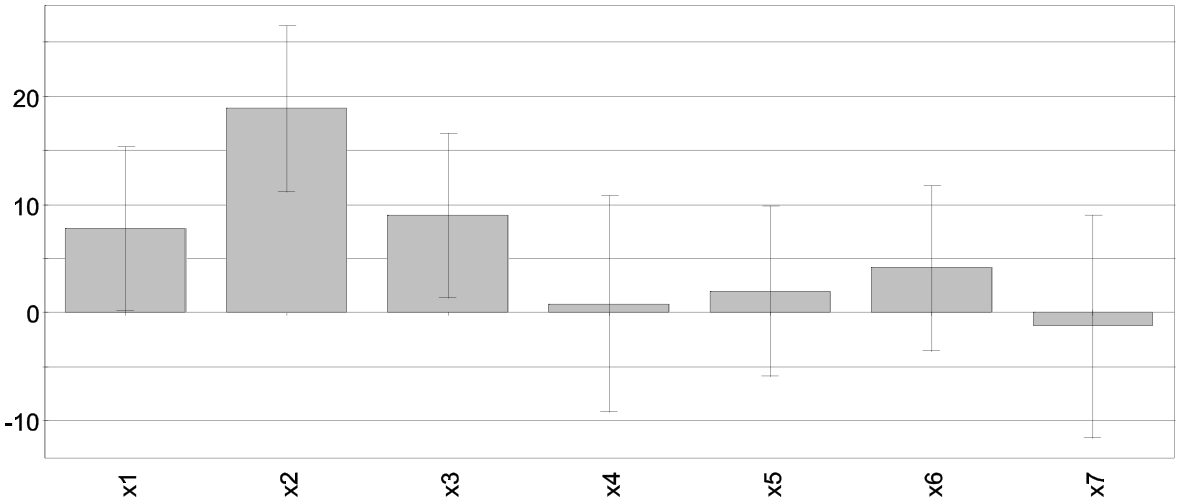


c

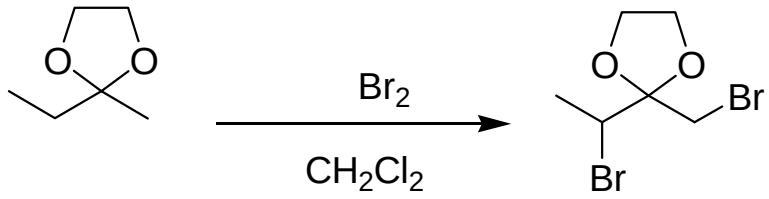


d

Figure 4



Scheme 1



Scheme 2