RESPONSE SURFACE METHODOLOGY (RSM)

A Brief Introduction

Statistical modeling strategies for dealing with what engineers call *processes*:

$$y = \eta(\xi_1, \xi_2, ..., \xi_f) + \epsilon$$

- y = response
- ξ's = factors, independent variables, now explicitly *continuous*
- modeling via regression and experimental design

Emphasis is on *local* approximations of η and/or estimates of its important characteristics, e.g.

- Which ξ's are most important?
- Which values of ξ's produced greatest/least response values?

Example: Chemical yield (y) as a function of process time (ξ_1) and operating temperature (ξ_2).

- Focus on estimating $E[\text{yield}]$ if process contains randomness.
- The *Region of Operability* is the $\boldsymbol{\xi}$-space in which the system *could* realistically be operated; in this case, perhaps:

$$0 \text{ hr} \leq \xi_1 \leq 7 \text{ hr} \qquad 100\ ^\circ\text{C} \leq \xi_2 \leq 800\ ^\circ\text{C}$$

- Note that in this notation, ξ's are values with *physical units*.
- η may be *complex globally*, over the entire region of operability.
- In most applications, we don't really need a model that holds for all conditions in the region of operability ... maybe just need to know (for example):
  - $\boldsymbol{\xi}$ that maximizes or minimizes η
  - $\boldsymbol{\xi}$-region for which $\eta > \eta_{min}$
  - $\boldsymbol{\xi}$-region where η is unstable (i.e. large derivatives)

Rather than focusing on the exact physical representation of the system (which is often not fully known), RSM generally relies on *local approximations* of η within smaller *Regions of Experimentation*

$$y \approx p^*(\xi_1, \xi_2, ..., \xi_f) + \epsilon$$

- p^* usually a low-order polynomial
- $\epsilon =$ additive random error/noise

$$E[y] = \eta \approx p^*(-) \qquad Var[y] = Var[\epsilon] = \sigma^2$$

Example (continued)

- Local approximation in a region of experimentation:

$$2.5 \text{ hr} \leq \xi_1 \leq 3 \text{ hr} \qquad 500\ ^\circ\text{C} \leq \xi_2 \leq 600\ ^\circ\text{C}$$

- Approximating model p^* = first-order polynomial:

$$\eta \approx p^* = \beta_0^* + \beta_1^* \xi_1 + \beta_2^* \xi_2$$

- Physical units are:
  - η, y = % yield (i.e. % of raw material converted)
  - ξ_1 = hr
  - $\xi_2 = {}^\circ$C
- Physical units must be the same in every term in the model, so:
  - β_0^* = %
  - β_1^* = %/hr
  - β_2^* = %/°C

Often, factors are re-expressed in an equivalent *scaled* or *unitless* form so that the fitted coefficients are more comparable.

- Scaled variables: $x = (\xi - \text{mid-value})/\frac{1}{2}(\text{high} - \text{low})$
  - $x_1 = (\xi_1 - 2.75)/\frac{1}{2}(3.0 - 2.5)$
  - $x_2 = (\xi_2 - 550)/\frac{1}{2}(600 - 500)$
- So: -1 $\leq x_1 \leq$ 1 -1 $\leq x_2 \leq$ 1
- Approximating model in scaled variables:

$$\eta \approx p(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \quad \text{(no "*" now)}$$

- Units are:
  - η, y = % yield $\beta_0 = \beta_0^* =$ %
  - x_1 = unitless, so β_1 = %
  - x_2 = unitless, so β_2 = %
- Can also standardize each ξ by std. dev. of the values actually used in the experimental runs, rather than half-range.

Some elements of RSM:

- Design for 1st- and 2nd-order polynomial regression models
- Analysis of fitted surfaces, e.g.
  - lack-of-fit
  - gradients
  - "ridge" characterization
- Strategies for sequential experiments
  - usually for finding values of ξ's that maximize or minimize the expected response

Designs for First-Order Polynomial Models

Good news! You've already seen most of them:

- full 2^f designs
- regular 2^{f-s} fractions of resolution at least III
- Placket-Burman irregular fractions

In each case, inputs tested at only two levels are used to make *interpolative* inference across the interval of values for each variable.

Even though *these* designs are also used for 2^f factorial studies, the modeling problem is fundamentally different here, e.g.:

- experiment with $x_i = \pm 1$
- predict y or estimate $E(y)$ at $x_1 = 0.25, x_2 = 0.60, ...$

And *some* first-order RSM designs aren't just "on the corners":

Saturated First-Order Designs: Simplex Designs

- A design is "saturated" if the number of unique runs is equal to the number of model parameters, and $\mathbf{X}$ is of full column rank.
- Minimal P-B designs are saturated when $f + 1$ is a multiple of 4, and "near-saturated" for other values of f; the number of distinct treatments is never more than 3 larger than $f + 1$.
- Simplex designs are saturated for first-order models for *every* value of f:
  - the unique design points are the $f + 1$ corners of a f-dimensional *simplex*, a geometric figure with all edges of equal length.
  - examples:
    - $f = 2$, the 3 corners of an equilateral triangle
    - $f = 3$, the 4 corners of a tetrahedron

- The "design matrix" $\mathbf{D}$ is the N-row by f-column matrix for which

$$\{\mathbf{D}\}_{i,j} = x_{i,j}$$

the value of the independent variable x_i in the jth experimental run.

- *One* way to generate a design matrix for a simplex design in f variables is:

$$\mathbf{D} = \sqrt{f+1} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{12}} & \cdots & \frac{1}{\sqrt{f^2+f}} \\ \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{12}} & \cdots & \frac{1}{\sqrt{f^2+f}} \\ 0 & \frac{-2}{\sqrt{6}} & \frac{1}{\sqrt{12}} & \cdots & \frac{1}{\sqrt{f^2+f}} \\ 0 & 0 & \frac{-3}{\sqrt{12}} & \cdots & \frac{1}{\sqrt{f^2+f}} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & \frac{-f}{\sqrt{f^2+f}} \end{pmatrix}$$

- Examples:

$$f = 2: \begin{pmatrix} \sqrt{\frac{3}{2}} & \sqrt{\frac{3}{2}} \\ -\sqrt{\frac{3}{2}} & \sqrt{\frac{3}{2}} \\ 0 & -\sqrt{2} \end{pmatrix} \quad f = 3: \begin{pmatrix} \sqrt{2} & \sqrt{\frac{2}{3}} & \frac{1}{\sqrt{3}} \\ -\sqrt{2} & \sqrt{\frac{2}{3}} & \frac{1}{\sqrt{3}} \\ 0 & -2\sqrt{\frac{2}{3}} & \frac{1}{\sqrt{3}} \\ 0 & 0 & -\frac{3}{\sqrt{3}} \end{pmatrix}$$

- Notes:
  - not 2-level
  - scaling here is by SD rather than Range/2, but each x is still centered on zero
  - can rotate the simplex in f-space to get a different design matrix
- Saturation $\rightarrow \hat{y} = y$, 0 error d.f., no L.O.F. check ...

Center Points and First-Order Diagnostics

- All first-order designs we've discussed in n_f points are such that:

$$\sum_{u=1}^{n_f} x_{iu} = 0 \quad \sum_{u=1}^{n_f} x_{iu}x_{ju} = 0, i \neq j \quad \sum_{u=1}^{n_f} x_{iu}^2 = n_f$$

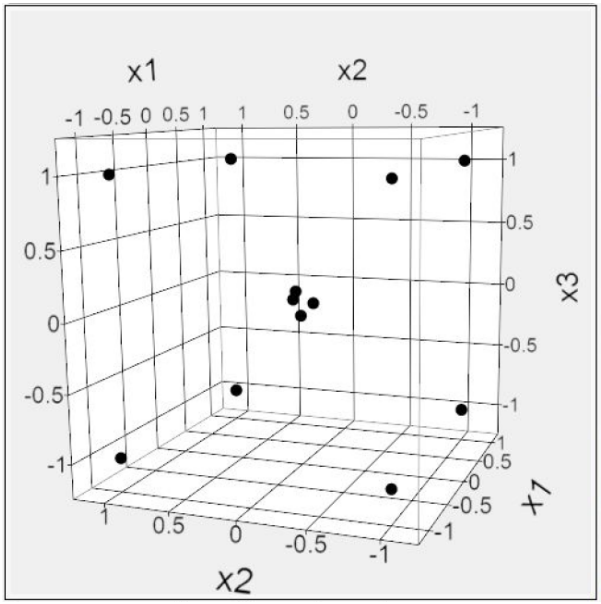
- If, as we assume, a first-order model is correct, the average of responses at these points has expectation $E(\bar{y}_f) = \beta_0$.

- But suppose the model is actually second-order:

$$E(y) = \beta_0 + \sum_{i=1}^{f} \beta_i x_i + \sum_{i<j}^{f} \beta_{ij} x_i x_j + \sum_{i=1}^{f} \beta_{ii} x_i^2$$

- Then $E(\bar{y}_f) = \beta_0 + \sum_{i=1}^{f} \beta_{ii}$

- Now, suppose we also add n_c center-points $(x_1, x_2, x_3, ..., x_f)' = \mathbf{0}$.



- Regardless of the appropriate polynomial model, $E(\bar{y}_c) = \beta_0$.
- So, $\text{Hyp}_0 : \sum_{i=1}^{f} \beta_{ii} = 0$ can be tested with:

$$\frac{\bar{y}_f - \bar{y}_c}{s_c \sqrt{{n_f}^{-1} + n_c^{-1}}} : t(n_c - 1)$$

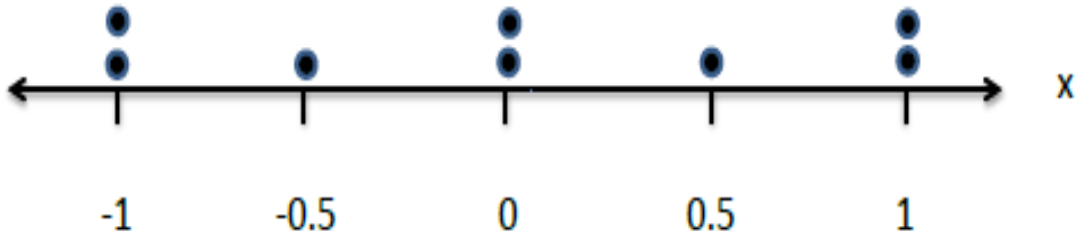
where s_c is the sample s.d. of center-point responses

Center point test is a "one-degree-of-freedom" special case of lack-of-fit testing. Quick overview/review:

- Use, as an example, simple linear regression:

$$y_{ij} = \beta_0 + \beta_1 x_i + \epsilon_{ij}$$

with $N = 8$, 5 unique values of x, and $y_{ij} = j$th value in the ith group, e.g.:



- Variance decomposition, with M for "model" and E for "error":

source	df	sum-of-squares
model	1	$SS_M(\beta_1\|\beta_0) = SS_M(\beta_0,\beta_1) - SS_M(\beta_0)$
		$= SS_E(\beta_0) - SS_E(\beta_0,\beta_1)$
residual	6	$SS_E(\beta_0,\beta_1)$
LOF	3	$SS_M(\text{groups}\|\beta_0,\beta_1) = SS_M(\text{groups}) - SS_M(\beta_0,\beta_1)$
		$= SS_E(\beta_0,\beta_1) - SS_E(\text{groups})$
replication	3	$SS_E(\text{groups}) = \sum_{i,j}(y_{ij} - \bar{y}_i)^2$
c.t.	7	$SS_E(\beta_0)$

First-Order Analysis: Steepest Ascent/Decent

- Used when the goal is to find a "path" of experimental conditions that lead toward a maximum(minimum) response, and this is clearly outside of the current experimental region.
- Most often used where the response function is at least approximately locally linear; suppose:

$$E(y) = \beta_0 + \sum_i \beta_i x_i$$

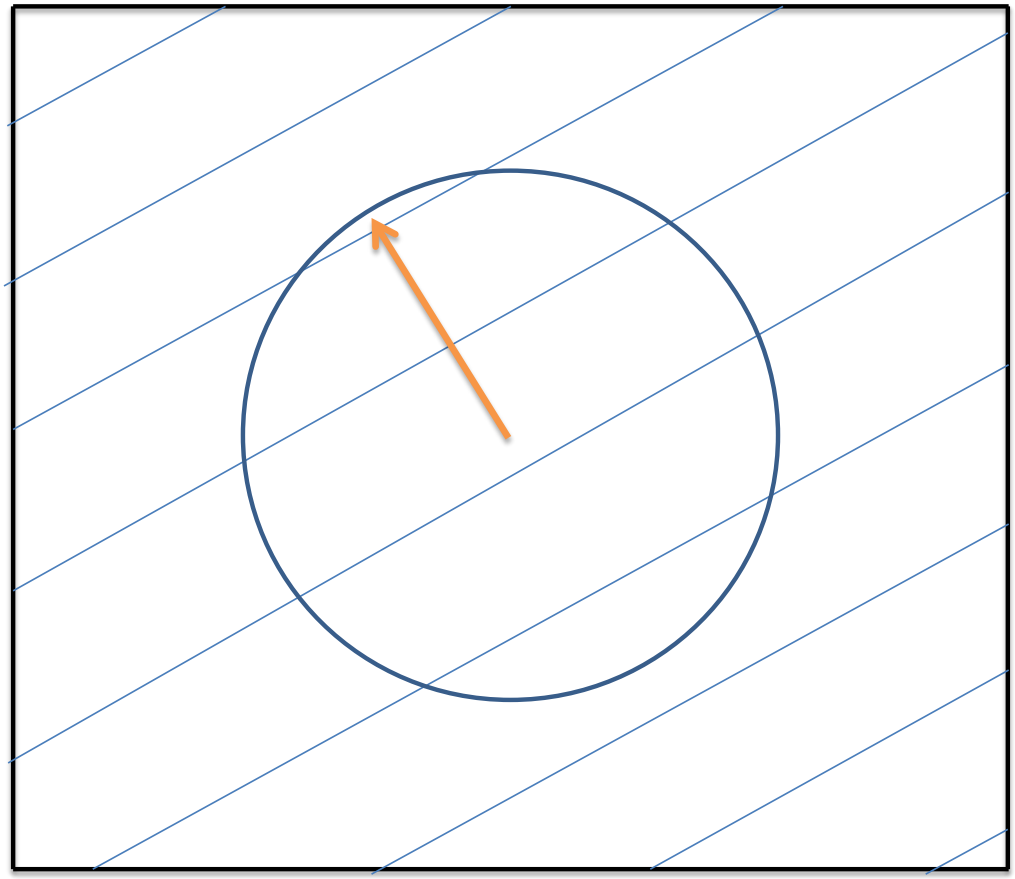
- New notation:
  - $\underline{x}' = (x_1, x_2, ..., x_f)$, e.g. a row from $\mathbf{D}$
  - $\mathbf{x}'$, e.g. a row from the model matrix ... "extended" $\underline{x}'$
- Starting at the center of the current experimental region, $\underline{x}_0' = (0, 0, ..., 0)$, which direction should we go in the design space to maximize the resulting increase/decrease in $E(y)$ for a fixed stepsize r?

- More formally, what $\underline{x}_r' = (x_1, x_2, ..., x_f)$ satisfies:

  maximize $[\beta_0 + \sum_i \beta_i x_i] - [\beta_0] = \sum_i \beta_i x_i$

  subject to $\sum_i x_i^2 = r^2$

- Solve this with Method of Lagrangian Multipliers ...
- Define $\phi(\underline{x}, \lambda) = \sum_i \beta_i x_i - \lambda[\sum_i x_i^2 - r^2]$
- Take derivatives, set equal to zero, solve:
  - $\frac{\partial}{\partial x_i}\phi = \beta_i - 2\lambda x_i = 0$
  - $\frac{\partial}{\partial \lambda}\phi = \sum_i x_i^2 - r^2 = 0$
- Derivatives w.r.t. $\underline{x}_r \rightarrow x_i$ must be proportional to β_i
- Then derivative w.r.t. $\lambda \rightarrow x_i = \frac{\beta_i}{\sqrt{\sum_i \beta_i^2}} r$
- For steepest *descent*, the sign on the objective function is reversed, and so is the sign on the solution.



- The value of r used in the above argument is arbitrary – in fact, "The Method of Steepest Ascent" is generally based on taking a series of new observations along the "path", usually at evenly spaced steps.
- Some practitioners advocate scaling the actual step so that the incremental change in one of the independent variables is ± 1. If this is done for the variable with the largest (absolute) β, it assures that the first step will not be outside of the present region of experimentation:
  - Let $M = \max_i |\beta_i|$
  - Then define $\Delta_i = \beta_i / M, i = 1, 2, 3, ..., f$
  - New design points (in current coding of variables) are then

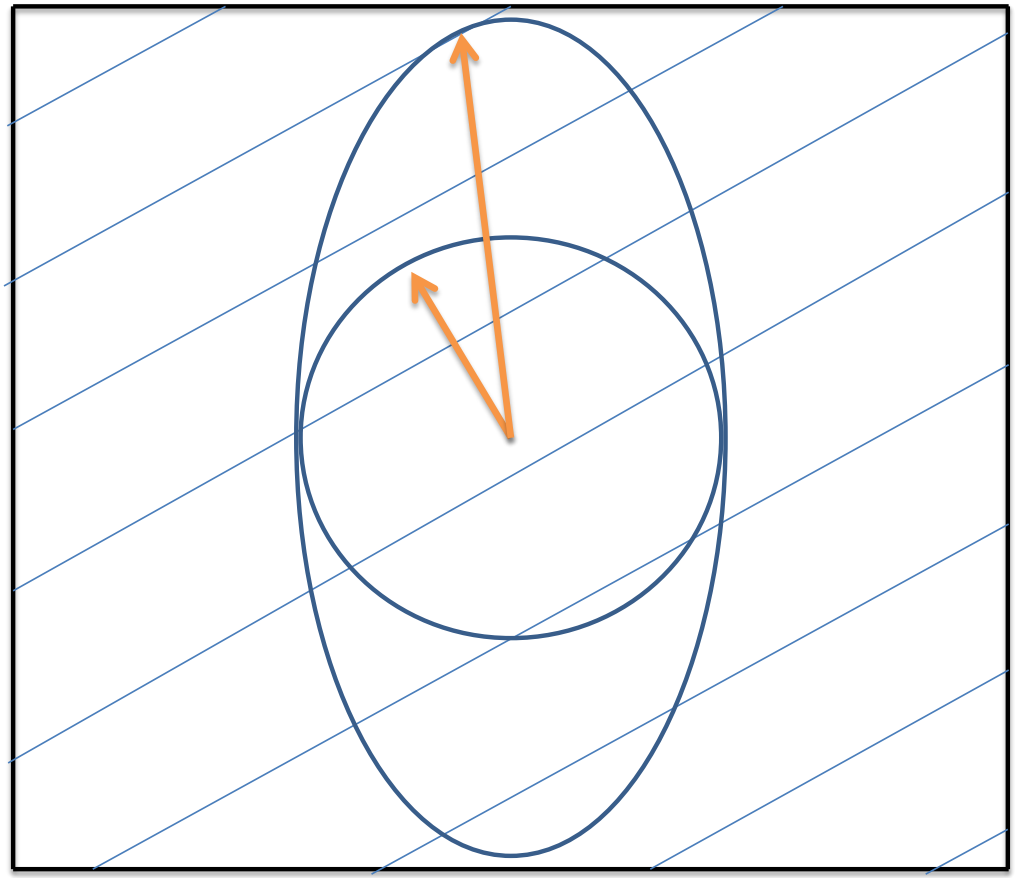
$$\underline{\Delta},\ 2\underline{\Delta},\ 3\underline{\Delta},\ ...$$

- Basic strategy:
  - Use a first-order design; augment with center-points.
  - Test for lack-of-fit; if no evidence of second-order terms ...
  - Fit a first-order model.
  - *Estimate* a path of steepest ascent, using $\hat{\boldsymbol{\beta}}$ in place of of $\boldsymbol{\beta}$.
  - Conduct experimental runs along the path until the response no longer increases. (Unless the global response surface has no maximum, this will eventually happen ...)
  - Use this point (or one nearby) as the centerpoint of a new experimental region, and plan and execute a new experiment.

Effect of Scaling: Suppose in truth, $E(y) = 100 + 2\xi_1 + 1\xi_2$

- Experimenter #1:
  - E.R.: $10 \leq \xi_1 \leq 20$, $10 \leq \xi_2 \leq 50$
  - $E(y) = [-] + [2 \times 5]\frac{\xi_1 - 15}{5} + [1 \times 20]\frac{\xi_2 - 30}{20}$
  - $\underline{\Delta}$ is proportional to $(10, 20)$, say $(\frac{1}{2}, 1)$
  - First step in *uncoded* variables is $(\Delta_1 \times 5 + 15, \Delta_2 \times 20 + 30) = (17.5, 50)$
- Experimenter #2:
  - E.R.: $5 \leq \xi_1 \leq 25$, $10 \leq \xi_2 \leq 50$
  - $E(y) = [-] + [2 \times 10]\frac{\xi_1 - 15}{10} + [1 \times 20]\frac{\xi_2 - 30}{20}$
  - $\underline{\Delta}$ is proportional to $(20, 20)$, say $(1, 1)$
  - First step in *uncoded* variables is $(\Delta_1 \times 10 + 15, \Delta_2 \times 20 + 30) = (25, 50)$

- So the path of steepest ascent is *not* scale-invariant ... it depends on the range of values set for each ξ in the experiment
- Why? Recall, the path is defined by the direction in which $E(y)$ is most increased for a step of fixed size *in the coded variables*.
- If scaling of ξ_1 is changed so that $x_1 \in [-1, +1]$ is *larger* on the physical scale, this increases β_1 relative to β_2 ...
- Even so, the steepest ascent path in either coding *does* lead to increased responses (at least within the region where the first-order approximation is accurate).



How would Steepest Ascent work if you wanted to base it on a model with main effects and two-factor interactions (sometimes called "bilinear" terms)?

- Lagrangian Multipliers ...

  maximize $(\beta_0 + \sum_i x_i\beta_i + \sum_{i<j} x_i x_j \beta_{ij}) - (\beta_0)$

  subject to $\sum_i x_i^2 = r^2$

- $\phi = \sum_i x_i\beta_i + \sum_{i<j} x_i x_j \beta_{ij} - \lambda[\sum_i x_i^2 - r^2]$
- $\frac{\partial}{\partial x_i} = \beta_i + \sum_{j\neq i} x_j\beta_{ij} - 2\lambda x_i = 0$

$$\begin{pmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_k \end{pmatrix} = \begin{pmatrix} 2\lambda & -\beta_{12} & -\beta_{13} & \dots & -\beta_{1k} \\ -\beta_{12} & 2\lambda & -\beta_{23} & \dots & -\beta_{2k} \\ \dots & \dots & \dots & \dots & \dots \\ -\beta_{1k} & -\beta_{2k} & -\beta_{3k} & \dots & 2\lambda \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_k \end{pmatrix}$$

- λ controls the step size ... large λ = small step

Example:

- $E(y) = 10 + 2x_1 + 3x_2 - 1x_1x_2$:

$$\begin{pmatrix} 2 \\ 3 \end{pmatrix} = \begin{pmatrix} 2\lambda & 1 \\ 1 & 2\lambda \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

λ	∞	5	2	1	0.75	0.65
x_1	0	0.17	0.33	0.33	0	−0.58
x_2	0	0.28	0.67	1.33	2	2.75

Central Composite Designs for Second-Order Models

- As optimization nears the "top of the hill," first-order and bilinear models become ineffective ... need full quadratic:

$$E(y) = \beta_0 + \sum_{i=1}^{f} \beta_i x_i + \sum_{i<j}^{f} \beta_{ij} x_i x_j + \sum_{i=1}^{f} \beta_{ii} x_i^2$$

- Most parameters are estimable using designs we've already discussed, even in the presence of the β_{ii} terms:
  - β_0 : e.g. any design with center points
  - β_i : e.g. Resolution IV fraction
  - β_{ij} : e.g. Resolution V fraction
- Q: What is the *simplest* additional collection of design points that would allow estimation of the β_{ii}, given the presence of the lower-order parameters in the model?

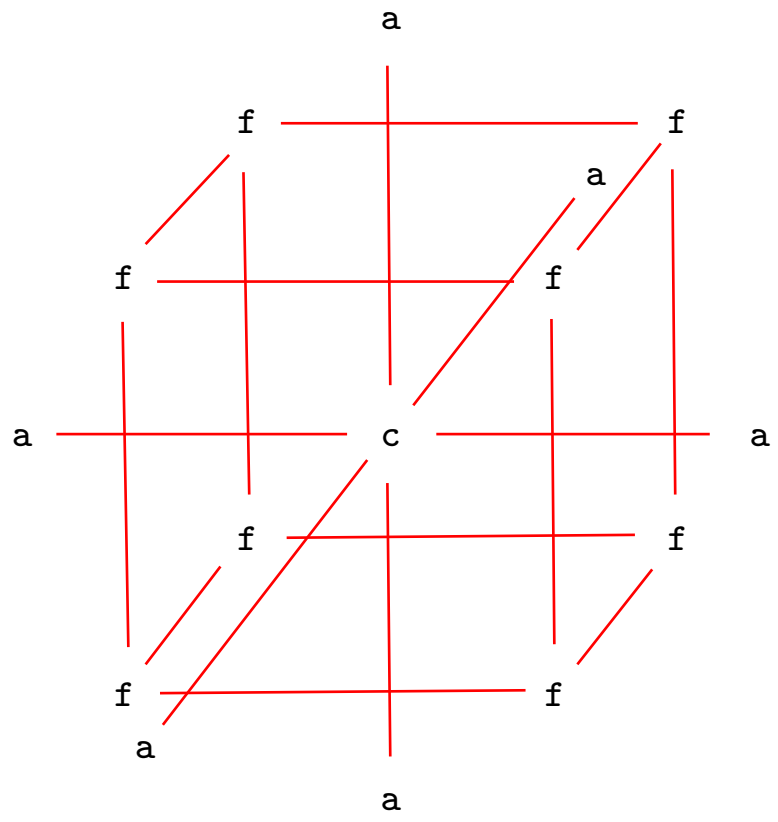
- A: Set of $2f$ "star" or "axial" points, in conjunction with center points (already in the design):

$$\begin{pmatrix} +\alpha & 0 & ... & 0 \\ -\alpha & 0 & ... & 0 \\ 0 & +\alpha & ... & 0 \\ 0 & -\alpha & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & +\alpha \\ 0 & 0 & ... & -\alpha \end{pmatrix}$$

- This gives 3 *collinear* points in the direction of each of the f design axes, allowing assessment of "curvature" (i.e. estimate β_{ii}) e.g. in x_1,

$$(-\alpha, 0, 0, ..., 0), \quad (0, 0, 0, ..., 0), \quad (+\alpha, 0, 0, ..., 0)$$

- Center points + 2-level fraction (Res $\geq$ V) + axials
  = Central Composite Design (CCD),
  Box and Wilson (1951) *JRSSB*.



- "Design parameters" generally refer to the values of
  - α = 1-dim distance from origin to each axial point (in scaled units)
  - n_c = number of center points (assuming other points are unreplicated)
- Complete specification also requires selection of a particular Resolution V 2-level design; variance properties depend only on n_f, the number of points in the factorial portion of the design, not the specific fraction. (But bias properties are a different matter ...)
- One nice feature of CCD's is the potential for sequential construction based on need, e.g.:
  - begin with Res III fraction, but if 1-st order fit isn't acceptable,
  - augment to Res V fraction, but if bi-linear fit isn't acceptable,
  - add axial points for CCD.

Dividing columns into intercept, first-order, bi-linear, and pure quadratic terms:

$$\mathbf{x'x} =$$

$$\left(\begin{array}{c|ccc|ccc|ccc}
N & 0 & \dots & 0 & 0 & \dots & 0 & n_f + 2\alpha^2 & \dots & n_f + 2\alpha^2 \\
\hline
0 & n_f + 2\alpha^2 & \dots & 0 & 0 & \dots & 0 & 0 & \dots & 0 \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
0 & 0 & \dots & n_f + 2\alpha^2 & 0 & \dots & 0 & 0 & \dots & 0 \\
\hline
0 & 0 & \dots & 0 & n_f & \dots & 0 & 0 & \dots & 0 \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
0 & 0 & \dots & 0 & 0 & \dots & n_f & 0 & \dots & 0 \\
\hline
n_f + 2\alpha^2 & 0 & \dots & 0 & 0 & \dots & 0 & n_f + 2\alpha^4 & \dots & n_f \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
n_f + 2\alpha^2 & 0 & \dots & 0 & 0 & \dots & 0 & n_f & \dots & n_f + 2\alpha^4
\end{array}\right)$$

Non-zeros are diagonal elements, (β_0, β_{ii}) pairs, and (β_{ii}, β_{jj}) pairs.

Design Moments are often used to characterize estimation properties of CCD's. A design moment is a normalized (by N) element of $\mathbf{X}'\mathbf{X}$, denoted using the subscripting system we use for elements of $\boldsymbol{\beta}$. Let x_{iu} be the value of x_i in run u ... then, e.g.:

- $[i] = \frac{1}{N}\sum_u x_{iu}$, "first moments"
- $[ij] = \frac{1}{N}\sum_u x_{iu}x_{ju}$, "second mixed moments"
- $[iiii] = \frac{1}{N}\sum_u x_{iu}^4$, "fourth pure moments"

Moments through order 4 are needed to describe all the precision characteristics of a design used to fit a quadratic model. For a CCD, the non-zero moments through order 4 are:

- $[ii] = (n_f + 2\alpha^2)/N$
- $[iiii] = (n_f + 2\alpha^4)/N$
- $[iijj] = n_f/N$

Rotatability:

- A design is said to be *rotatable* if $Var(\hat{y}(\underline{x}))$ is the same for all $\underline{x}$ satisfying $\underline{x}'\underline{x} = r^2$ for any radius r.
- For a simple start, suppose we are just fitting a first-order model using our central composite design:

$$\mathbf{X}'\mathbf{X} = N \left(\begin{array}{c|ccc} 1 & 0 & ... & 0 \\ \hline 0 & [11] & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & [ff] \end{array} \right)$$

- Since all $[ii]$ are equal for this design, and because $\mathbf{X}'\mathbf{X}$ is diagonal:

  $Var(\hat{y}(\underline{x})) = \sigma^2 \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x} = \frac{\sigma^2}{N}(1 + \underline{x}'\underline{x}/[ii]) = \frac{\sigma^2}{N}(1 + r^2 \frac{N}{n_f + 2\alpha^2})$

  so any CCD is rotatable for a first-order model. (Also true of all the other 1st order orthogonal designs we've talked about.)

- Rotatability for 2nd order models is a bit messier, since the algebra requires inversion of a non-diagonal matrix, but the (general) bottom line is:
- A design is rotatable w.r.t. full 2nd order polynomial models iff:
  - all odd moments* through 4th order are zero, and
  - $[iiii]/[iijj] = 3$ for all $i \neq j$

  *: Any moment with an odd number of letters, whether distinct or not, e.g. through 4th order,

$$[i] = [ijk] = [iij] = [iii] = 0.$$

- For central composite designs, this is satisfied by setting:

$$[iiii]/[iijj] = \frac{n_f + 2\alpha^4}{n_f} = 3$$

$$\alpha = n_f^{1/4}$$

A Weakness in the Rotatability Argument:

- Consider $f = 2$, $n_f = 4$, $\alpha = \sqrt{2}$ for rotatability.
- Suppose we include no center points in the design:

$$\mathbf{X} = \left(\begin{array}{c|cc|c|cc} 1 & +1 & +1 & +1 & 1 & 1 \\ 1 & +1 & -1 & -1 & 1 & 1 \\ 1 & -1 & +1 & -1 & 1 & 1 \\ 1 & -1 & -1 & +1 & 1 & 1 \\ 1 & +\sqrt{2} & 0 & 0 & 2 & 0 \\ 1 & -\sqrt{2} & 0 & 0 & 2 & 0 \\ 1 & 0 & +\sqrt{2} & 0 & 0 & 2 \\ 1 & 0 & -\sqrt{2} & 0 & 0 & 2 \end{array}\right)$$

- The design is *singular* because $x_1^2 + x_2^2 = 2$ in each run.
- Same thing happens for $f = 4$, and "almost" for $f = 3$.
- This can be fixed by adding center points, but it may take several of them to produce good predictive variances near the origin.

Practical rules of thumb for CCD's:

- For spherical experimental regions:
  - $\alpha = \sqrt{f}$
  - $n_c =$ 3-to-5 (or more if σ is relatively large)
- For cubic experimental regions:
  - $\alpha = 1$ (sometimes called a "face-centered design")
  - more freedom to set n_c

Orthogonal Blocking for CCD's

- Because CCD's are often constructed sequentially, they may need to be analyzed as *blocked* experiments.
- A goal in this case should be to avoid losing efficiency due to the need to simultaneously estimate (nuisance) block effects:

  $$E(y) = \delta_l + \beta_0 + \sum_i x_i\beta_i + \sum_{i<j} x_i x_j \beta_{ij} + \sum_i x_i^2 \beta_{ii}$$

  for an observation at $\underline{x}$ in the l-th block.
- Partitition the model as:

$$E(\mathbf{y}) = \left(\begin{array}{ccc|c} \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{X}_1 \\ \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{X}_2 \\ \ldots & \ldots & \ldots & \ldots \\ \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{X}_b \end{array}\right) \left(\begin{array}{c} \boldsymbol{\delta} \\ \boldsymbol{\beta} \end{array}\right)$$

- ... where $\mathbf{X}_l$ is the piece of the original $\mathbf{X}$ matrix executed as the l-th block, of n_l runs.
- Note the first $b+1$ columns (blocks and intercept) are linearly dependent. This can be eliminated by replacing the b block columns with $b-1$ weighted differences:

$$E(\mathbf{y}) = \left(\begin{array}{cccc|c} \frac{1}{n_1}\mathbf{1} & \mathbf{0} & ... & \mathbf{0} & \mathbf{X}_1 \\ -\frac{1}{n_2}\mathbf{1} & \frac{1}{n_2}\mathbf{1} & ... & \mathbf{0} & \mathbf{X}_2 \\ \mathbf{0} & -\frac{1}{n_3}\mathbf{1} & ... & \mathbf{0} & \mathbf{X}_3 \\ ... & ... & ... & ... & \\ \mathbf{0} & \mathbf{0} & ... & -\frac{1}{n_b}\mathbf{1} & \mathbf{X}_b \end{array}\right) \left(\begin{array}{c} \boldsymbol{\delta}^* \\ \boldsymbol{\beta} \end{array}\right) = (\mathbf{W}|\mathbf{X}) \left(\begin{array}{c} \boldsymbol{\delta}^* \\ \boldsymbol{\beta} \end{array}\right)$$

- Blocking is *orthogonal* if $\mathbf{W}'\mathbf{X} = \mathbf{0}$, so the "reduced normal equations" for $\boldsymbol{\beta}$ are $(\mathbf{X}'\mathbf{X})^{-1}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$, as if the experiment were not blocked.

Following the model just derived, orthogonal blocking requires:

1. from multiplying $\mathbf{W}'$ by the intercept column:

   $$\frac{1}{n_1}\sum_{blk\ 1} 1 = \frac{1}{n_2}\sum_{blk\ 2} 1 = ... = \frac{1}{n_b}\sum_{blk\ b} 1$$

   ... true due to my selection of "weights".

2. from multiplying $\mathbf{W}'$ by the 1-st order columns:

   $$\frac{1}{n_1}\sum_{blk\ 1} x_{iu} = \frac{1}{n_2}\sum_{blk\ 2} x_{iu} = ... = \frac{1}{n_b}\sum_{blk\ b} x_{iu}, \text{ any } i$$

3. from multiplying $\mathbf{W}'$ by the bi-linear columns:

   $$\frac{1}{n_1}\sum_{blk\ 1} x_{iu}x_{ju} = \frac{1}{n_2}\sum_{blk\ 2} x_{iu}x_{ju} = ... = \frac{1}{n_b}\sum_{blk\ b} x_{iu}x_{ju},$$
   $$\text{any } i \neq j$$

   ... both (2) and (3) are true if <u>each block</u> is an orthogonal main-effects design

4. from multiplying $\mathbf{W}'$ by the pure quadratic columns:

   $$\frac{1}{n_1}\sum_{blk\ 1} x_{iu}^2 = \frac{1}{n_2}\sum_{blk\ 2} x_{iu}^2 = ... = \frac{1}{n_b}\sum_{blk\ b} x_{iu}^2, \text{ any } i$$

- Note that the above conditions are really about design moments computed individually for each block.
- Restrict attention to blocking schemes for which each block is an orthogonal main-effects plan (to satisfy (2) and (3)), and achieve (4) by selection of α and the number of center points in each block.
- Example: $f = 2$, $b = 2$, n_{c1} and n_{c2} center points per block:

$$\text{Block 1,} \begin{pmatrix} +1 & +1 \\ +1 & -1 \\ -1 & +1 \\ -1 & -1 \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \qquad \text{Block 2,} \begin{pmatrix} +\alpha & 0 \\ -\alpha & 0 \\ 0 & +\alpha \\ 0 & -\alpha \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$$

Need: $\frac{4}{4+n_{c1}} = \frac{2\alpha^2}{4+n_{c2}} \quad \rightarrow \quad \alpha = \sqrt{\frac{16+4n_{c2}}{8+2n_{c1}}}$

- Example: $f = 4$, $b = 3$, split 2^4 into two 2^{4-1} fractions (with center points) in blocks 1 and 2, and axials (with center points) in block 3
  - Need: $\frac{8}{8+n_{c1}} = \frac{8}{8+n_{c2}} = \frac{2\alpha^2}{8+n_{c3}}$
  - Clearly, n_{c1} must be the same as n_{c2}.
  - Then: $\alpha = \sqrt{\frac{64+8n_{c3}}{16+2n_{c2}}}$
- Can also pick α and find n_c's that work, but this also usually requires some adjustment of α.

"Small" Composite Designs

- For situations in which each run is very expensive
- Construction as with regular Central Composite, but using a Res III* 2-level design for the factorial portion (rather than Res V).
- *: Not just any Res III fraction. Must be such that no word of length 4 is used in the defining relation, e.g.

$$I = ABC = DEF (= ABCDEF)$$ is OK, but not

$$I = ABC = CDE (= ABDE).$$

- Previous logic of using Res V was to avoid confounding between mixed quadratic coefficients. So, why does this work?
  - $ABCDE$ is OK, because 2nd order terms aren't confounded
  - $ABCD$ isn't OK, because AB and CD are confounded, et cetera
  - ABC is OK, because first-order effects are confounded with bilinear terms in *only the factorial section* of the design. The first-order terms (β_i) are *also* estimable, along with the pure quadratic terms (β_{ii}) in the axial portion.
- Example: $f = 4$, use $I = ABC$ (can't use $ABCD$), 4 runs fewer than comparable CC.
- Note: Statistical performance is substantially worse for these designs than for regular Central Composite Designs, especially for larger values of f.

Box-Behnken Designs

- Based on Balanced Incomplete Block Designs (BIB's ... quick review)
- Setting:
  - compare t treatments
  - in b blocks, each of size $m < t$, in such a way that
  - every treatment appears once in each of r blocks (first-order balance), and
  - every pair of treatments appears together in each of λ blocks (second-order balance)

- Example: $t = 3$, $b = 3$, $m = 2$:

1	1	2
2	3	3

  - here $r = 2$, generally $r = \frac{bm}{t}$, and
  - here $\lambda = 1$, generally $\lambda = \frac{bm(m-1)}{t(t-1)}$
  - so, these expressions don't have to yield integers (and BIB's don't exist) for every combination of t, b, and m values.

- Now for Box-Behnken Designs:
  1. Select a BIB design with t (BIB) $= f$ (BB)
  2. For each block in the BIB, generate 2^m runs for the BB which constitute:
     - the full 2-level factorial design in *just* the factors associated with the treatments in the BIB block
     - setting all other factors to zero (coded central value) in all runs
  3. Augment with center points.

- Continuing the example:

BIBD:

1	1	2
2	3	3

$\rightarrow$ BB design matrix:

$$\begin{pmatrix} + & + & 0 \\ + & - & 0 \\ - & + & 0 \\ - & - & 0 \\ + & 0 & + \\ + & 0 & - \\ - & 0 & + \\ - & 0 & - \\ 0 & + & + \\ 0 & + & - \\ 0 & - & + \\ 0 & - & - \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}$$

- Another example: $f = 6$ is an odd case ... because of limitations on the BIB's available, a *Partially Balanced Incomplete Block Design* is used:

1	2	3	1	2	1
2	3	4	4	5	3
4	5	6	5	6	6

- First-order balance requirement is satisfied as with BIB's, but:
- Treatment pairs (1,4), (2,5), and (3,6) are called "first associates" and appear together in 2 blocks each, and all other treatment pairs are called "second associates" and appear together in only 1 block each.
- Note that in this case, each PBIB block contributes $2^3 = 8$ runs to the BB design, because the block size is 3.

BB design matrix:

$$\begin{pmatrix} \pm & \pm & 0 & \pm & 0 & 0 \\ 0 & \pm & \pm & 0 & \pm & 0 \\ 0 & 0 & \pm & \pm & 0 & \pm \\ \pm & 0 & 0 & \pm & \pm & 0 \\ 0 & \pm & 0 & 0 & \pm & \pm \\ \pm & 0 & \pm & 0 & 0 & \pm \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}$$

- Orthogonal Blocking: Recall (from CCD development) that when each block is considered as an independent design, each must have the same:
  1. $[i]$
  2. $[ij]$
  3. $[ii]$
- Call the BB runs that come from any one BIBD block a "group".
- The first and second requirements are automatically satisfied if we keep each group of runs together in a common block (... all main-effect column sums within such groups are zero ... also inner products between any two such columns ...)
- The third is satisfied if these "groups" are combined in such a way that each factor is "active" in the same number of groups, in each block.

- Example: $f = 5$

$$\text{block 1} \quad \begin{pmatrix} \pm 1 & \pm 1 & 0 & 0 & 0 \\ \pm 1 & 0 & \pm 1 & 0 & 0 \\ 0 & 0 & 0 & \pm 1 & \pm 1 \\ 0 & \pm 1 & 0 & \pm 1 & 0 \\ 0 & 0 & \pm 1 & 0 & \pm 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$\text{block 2} \quad \begin{pmatrix} \pm 1 & 0 & 0 & \pm 1 & 0 \\ \pm 1 & 0 & 0 & 0 & \pm 1 \\ 0 & \pm 1 & \pm 1 & 0 & 0 \\ 0 & \pm 1 & 0 & 0 & \pm 1 \\ 0 & 0 & \pm 1 & \pm 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

- Note, the number of center points needs to be the same in each block for this to work.

Analysis for Second-Order Models

- Now suppose we've designed and modeled assuming:

$$E(y) = \beta_0 + \sum_{i=1}^{f} \beta_i x_i + \sum_{i<j}^{f} \beta_{ij} x_i x_j + \sum_{i=1}^{f} \beta_{ii} x_i^2$$

- Continuing to focus on questions of:
  - Where is the maximum located?
  - Which direction should I go to find the maximum?
  - Within what region is the surface approximately maximized?

$$\hat{\boldsymbol{\beta}}_1 = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \cdots \\ \hat{\beta}_f \end{pmatrix} \qquad \hat{\mathbf{B}}_2 = \begin{pmatrix} \hat{\beta}_{11} & \hat{\beta}_{12}/2 & \hat{\beta}_{13}/2 & \cdots & \hat{\beta}_{1f}/2 \\ \hat{\beta}_{12}/2 & \hat{\beta}_{22} & \hat{\beta}_{23}/2 & \cdots & \hat{\beta}_{2f}/2 \\ \hat{\beta}_{13}/2 & \hat{\beta}_{23}/2 & \hat{\beta}_{33} & \cdots & \hat{\beta}_{3f}/2 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \hat{\beta}_{1f}/2 & \hat{\beta}_{2f}/2 & \hat{\beta}_{3f}/2 & \cdots & \hat{\beta}_{ff} \end{pmatrix}$$

- $\hat{y} = \hat{\beta}_0 + \underline{x}'\hat{\boldsymbol{\beta}}_1 + \underline{x}'\hat{\mathbf{B}}_2\underline{x}$
- $\frac{\partial}{\partial \underline{x}'} = \hat{\boldsymbol{\beta}}_1 + 2\hat{\mathbf{B}}_2\underline{x}$
- Setting the derivative vector to $\mathbf{0}$ yields the stationary point:

$$\underline{x}_s = -\tfrac{1}{2}\hat{\mathbf{B}}_2^{-1}\hat{\boldsymbol{\beta}}_1$$

- This may be a maximum, minimum, or saddle point of the fitted surface. The eigenvalues (call them λs here) and eigenvectors of $\hat{\mathbf{B}}_2$ are the key to characterizing the shape. $\underline{x}_s$ is a:
  - point of maximum if all λ's are negative
  - point of minimum if all λ's are positive
  - saddle point if λ's are of mixed sign

- More careful consideration admits that $\underline{x}_s$ is an estimate and therefore uncertain, leading to standard errors and confidence regions.
- In any case, identification of the stationary point is nearly always just a first step. "Ridges" and gradients are often even more important.

Practical point about small (absolute) eigenvalues of $\hat{\mathbf{B}}_2$, "Ridges":

- Suppose λ_1 (the largest eigenvalue) is only slightly less than zero, and $\mathbf{p}_1$ is its associated eigenvector.
- Consider points along a line: $\underline{x}^* = \underline{x}_s + c\mathbf{p}_1$
- $\hat{y}$ changes very little along this line (because λ_1 is close to zero).

- That is, coded inputs of form:

$$\underline{x} = \underline{x}_s + c_1\mathbf{p}_1$$

should all approximately maximize $\hat{y}$, for some range of positive and negative values of c_1 ... sometimes called a "ridge".

- If λ_2 and λ_1 are both close to zero:

$$\underline{x} = \underline{x}_s + c_1\mathbf{p}_1 + c_2\mathbf{p}_2$$

should all approximately maximize $\hat{y}$ ...

- Systems like this allow choice in near-optimal operating conditions ... sometimes allows simultaneous consideration of other factors (e.g. time, cost ...)

Interesting Special Case: Mixtures Experiments

- Regression problems in which the x's represent *proportions* (by weight, volume, ...) of the components in a mixture.
- e.g. components of gasoline, recipes, alloys
- Implications for modeling:
  - $0 < x_i < 1$, $i = 1, 2, 3, ..., q$ ("q" is often used for "f" in the literature on mixture designs)
  - $\sum_i^q x_i = 1$

- First-order model:
  - Usual regression form: $E(y) = \beta_0 + \sum_{i=1}^{q} \beta_i x_i$, but in this case, there is always a linear dependency among terms, so:
  - Mixture form: $E(y) = \sum_{i=1}^{q} \beta_i x_i$
- Interpretation:
  - β_i is the expected response of the ith *pure blend*
  - $E(y)$ for all other blends can be gotten by linearly interpolating these, e.g. a 50:50 blend has expected response half-way between the responses for the two pure blends.

- Second-order model:
  - Here, $\sum_{j=1}^{q} x_j x_i = (\sum_{j=1}^{q} x_j) x_i = x_i$
  - We can eliminate this dependency by removing x_i^2
  - Mixture form:

$$E(y) = \sum_{i=1}^{q} \beta_i x_i + \sum_{i<j} \beta_{ij} x_i x_j$$

- Interpretation:
  - β_i is (still) the expected response at the ith pure blend
  - β_{ij} is the nonlinear effect between pure-i and pure-j

- "Special" third-order model (models are often of higher order than in unconstrained RSM ...):
  - Now, besides earlier relationships, $\sum_k x_i x_j x_k = x_i x_j$
  - We *can* eliminate this dependency by removing $x_i x_k^2$ and $x_j x_k^2$
  - "Special" mixture form:

    $$E(y) = \sum_{i=1}^{q} \beta_i x_i + \sum_{i<j} \beta_{ij} x_i x_j + \sum_{i<j<k} \beta_{ijk} x_i x_j x_k$$

  - Note: Response function along any $x_i : x_j$ "edge" is *quadratic*, since all 3rd-order terms are zero here

- The above model is called "special" because it's overconstrained; we dropped two terms to eliminate each linear dependency, when one would have sufficed. So in the *full* 3rd order mixture model:
  - Drop both $x_i^2 x_j$ and $x_i x_j^2$, but put back $x_i x_j (x_i - x_j)$
  - The following form looks a little odd, but allows us to maintain symmetry in all the x's

$$E(y) = \sum_{i=1}^{q} \beta_i x_i + \sum_{i<j} \beta_{ij} x_i x_j + \sum_{i<j} \delta_{ij} x_i x_j (x_i - x_j) + \sum_{i<j<k} \beta_{ijk} x_i x_j x_k$$

- Analysis to find response maxima or minima often rely on graphical representations of the fitted model, but Lagrangian multipliers *can* also be used.

Experimental Designs for Mixture Experiments:

- Simplex Lattice: $\{q, m\}$ refers to the the collection of points for which:
  - q proportions are represented
  - all combinations of $x_i = \frac{0,1,2,\ldots}{m}$ are included such that
  - $\sum_{i=1}^{q} x_i = 1$
- e.g. $\{3, 2\}$, $\{3, 3\}$, $\{4, 2\}$, $\{4, 3\}$
- $\{q, m\}$ allows the fit of a full (mixture) polynomial of order m

- Simplex-Centroid: Centroid (center) point of the full experimental region and each lower-dimensional simplex:

$(\frac{1}{q}, \frac{1}{q}, ..., \frac{1}{q})$	q-dimensional simplex
$(0, \frac{1}{q-1}, ..., \frac{1}{q-1})$ ... $(\frac{1}{q-1}, \frac{1}{q-1}, ..., 0)$	$(q-1)$-dimensional simplices
...	...
$(0, 0, ..., 1)$ ... $(1, 0, ..., 0)$	1-dimensional simplices

- allows the fit of "special" polynomial of order q ... all possible monomials made up of a product of a subset of x_i's