Contents

13

Super-saturated designs

Sometimes the number p of regression variables is larger than the number n of observations we can take. Just as ***saturated*** models have one parameter per obervation, ***supersaturated*** models have $p > n$ or even $p \gg n$.

In this chapter we look at how do design for such cases. This leads us to consider Hadamard matrices, some history of 'random balance designs', compressed sensing, and the Johnson-Lindenstrauss lemma. There is a survey of supersaturated designs in Georgiou (2014). Krahmer and Ward (2011) discusses experimental design for compressed sensing.

We have already seen supersaturated designs defined via fractional factorials where there are fewer observations than parameters. Here we focus on problems where there are fewer observations than main effects (plus intercept). A plain least squares fit will then interpolate the data, both signal and noise, assuming as is reasonable that no two predictors $\boldsymbol{x}_i$ are equal. There is no obvious way to estimate the error variance and now obvious way to check for lack of fit.

The designs we consider are also called ***screening designs*** because their goal is to identify the perhaps small number of relatively important predictors. They are well suited to settings where the regression model is thought to be sparse with the majority of regression coefficients either zero or at least negligible.

13.1 Hadamard matrices

We begin with Hadamard matrices that are suitable for saturated settings. The matrix $H \in \{-1, 1\}^{n\times n}$ is a ***Hadamard matrix*** if $H^\mathsf{T}H = HH^\mathsf{T} = nI$. For

instance, with $n = 4$,

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

is a Hadamard matrix. We could use it as a saturated design to fit an intercept and 3 binary variables. We have seen it before as a 2^{3-1} factorial. There is a good account of Hadamard matrices in (Hedayat et al., 1999, Chapter 7). It is definitive apart from a few recent contributions that one can find online either at Wikipedia or Neil Sloane's web site.

The ***Sylvester construction***, which actually pre-dates Hadamard's interest in these matrices is as follows:

$$H_1 = \begin{pmatrix} 1 \end{pmatrix}, \quad H_2 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = \begin{pmatrix} H_1 & H_1 \\ H_1 & -H_1 \end{pmatrix}, \quad H_4 = \begin{pmatrix} H_2 & H_2 \\ H_2 & -H_2 \end{pmatrix}$$

and so on with

$$H_{2^{k+1}} = \begin{pmatrix} H_{2^k} & H_{2^k} \\ H_{2^k} & -H_{2^k} \end{pmatrix}$$

for $k \geqslant 1$ in general.

Sylvester's construction is a special case of a ***Kronecker construction*** that works as follows. If $A \in \{-1, 1\}^{n\times n}$ and $B \in \{-1, 1\}^{m\times m}$ then their Kronecker product is

$$A \otimes B = \begin{pmatrix} A_{11}B & A_{12}B & \cdots & A_{1n}B \\ A_{21}B & A_{22}B & \cdots & A_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1}B & A_{n2}B & \cdots & A_{nn}B \end{pmatrix} \in \{-1, 1\}^{nm\times nm}$$

If A is a H_n and B is an H_m then $A \otimes B$ is an $H_{n\times m}$

Now if A is an H_n and B is an H_m then $A \otimes B$ is an $H_{n\times m}$. The proof is simple and it illustrates some basic rules for manipulating Kronecker products:

$$\begin{aligned} (A \otimes B)^{\mathsf{T}}(A \otimes B) &= (A^{\mathsf{T}} \otimes B^{\mathsf{T}})(A \otimes B) \\ &= (A^{\mathsf{T}}A) \otimes (B^{\mathsf{T}}B) \\ &= (nI_n) \otimes (mI_m) \\ &= (nm)I_n \otimes I_m \\ &= nmI_{nm}. \end{aligned}$$

Every step follows directly from the definition of the Hadamard product, so readers seeing Kronecker products for the first time should take a moment to check each step.

It is known that if a matrix H_n exists then $n = 1$ or 2 or $4m$ for some integer $m \geqslant 1$. The ***Hadamard conjecture*** is that H_n exists whenever $n = 4m$ for $m \geqslant 1$. There is no known counter example, but matrices for $n \in \{668, 716, 892\}$

n	# distinct H_n
1,2,4,8,12	1
16	5
20	3
24	60
28	487
32	13,710,027

Table 13.1: Number of distinct Hadamard matrices of sizes up to 32. From `https://oeis.org/A007299` as of October 2020.

have not (as of October 2020) been found and there are 10 more missing cases for $n \leqslant 2000$. These missing values are not a problem for experimental design. If we want H_{668} but have to use H_{772} instead, it would not be a costly increase in sample size. The most plausible uses for such large matrices are in software and four additional function evaluations are unlikely to be problematic.

Two Hadamard matrices are ***equivalent*** if one can be turned into the other by permuting its rows, or by permuting its columns, or by flipping the signs of an entire row or by flipping the signs of an entire column. The number of distinct (non-equivalent) Hadamard matrices that exist for some small values of n are in Table 14.1.

Given a Hadamard matrix we can always find an equivalent one whose first row and column are all $+1$. Hadamard matrices are often depicted in this form. In an experiment we would then use the first column to represent the intercept and the next $n-1$ columns for $n-1$ predictor variables. What we get is a Resolution III design (main effects clear) in $n-1$ binary variables.

In addition to the Sylvester construction there is a construction of Paley (1933) that is worth noting. If $n = 4m$ and $s = n-1 = p^r$ for a prime number p and exponent $r \geqslant 1$, then Paley's first construction provides H_n. The construction is available whenever $p^r \equiv 3 \bmod 4$. Figure 14.1 shows one of these matrices for $n = 252$ and prime number $p = 251 \equiv 3 \bmod 4$. Apart from a border of $+1$ at the left and top, each row of this matrix is a cyclic shift of the row above it. That means we can construct the matrix 'on the fly' and need not store it. That feature would be very useful for $n > 10^6$. There is a construction in (Hedayat et al., 1999, Theorem 7.1) that is quite simple to use for a prime number $p \bmod 4$. For $n-1 = p^r$ it would require finite field arithmetic that when $r = 1$ reduces to arithmetic modulo p. Note that the theorem gives a first row of H_n equal to $\begin{pmatrix}1 & -1 & -1 & \cdots & -1\end{pmatrix}$ instead of all 1s. However, we would randomize all the columns once constructed. (Be sure to pick one randomization for each of the $n-1$ columns and use it for all n rows.) Paley (1933) has a second construction, but it does not have quite the same simply implemented striping pattern.

We can use foldovers of Hadamard matrices. First split the intercept column

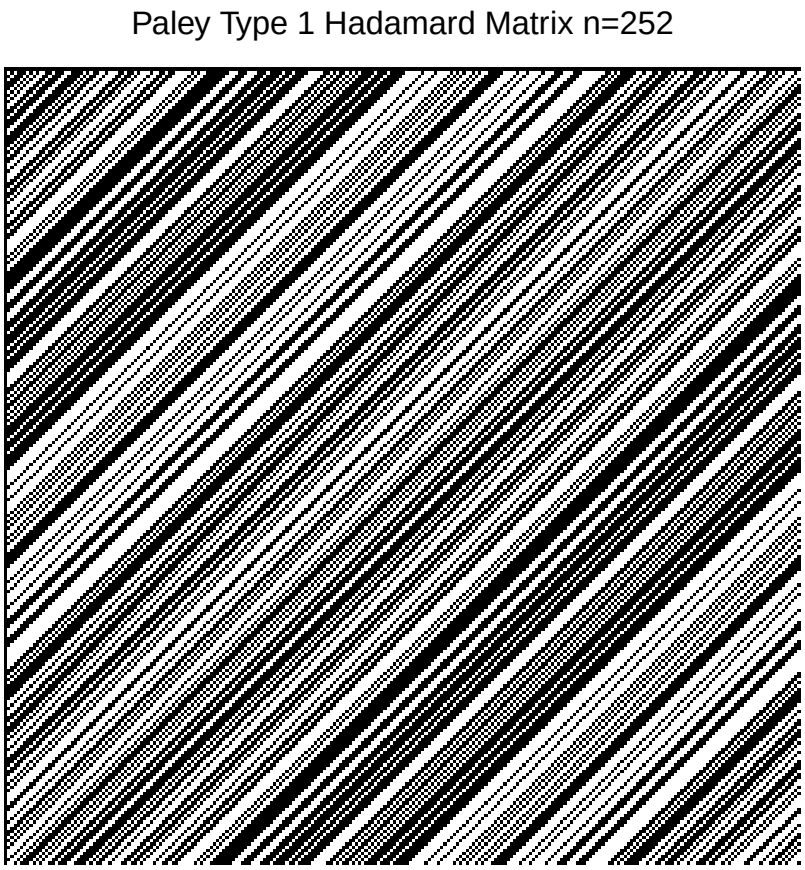

Figure 13.1: Image of a Hadamard matrix constructed using Paley's first construction with prime $p = 251$. Raw data from `http://neilsloane.com/hadamard/had.252.pal.txt` downloaded October 2020.

off producing $\tilde{H}_{4m} \in \{-1,1\}^{4m\times 4m-1}$ as follows

$$H_{4m} = \begin{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \tilde{H}_{4m} \end{pmatrix} \in \{-1,1\}^{4m\times 4m}.$$

Then flip all the non-intercept columns yielding

$$\begin{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} & \begin{pmatrix} \tilde{H}_{4m} \\ \\ -\tilde{H}_{4m} \end{pmatrix} \end{pmatrix} \in \{-1,1\}^{8m\times 4m}$$

Any three columns of this matrix have all eight of $\{-1,1\}^3$ m times each.

13.2 Group testing and puzzles

Suppose that

$$Y = X\beta + \varepsilon \in \mathbb{R}^n \qquad X \in \mathbb{R}^{n\times p} \quad p > n.$$

Finding β is doable if β is sparse, having only a few nonzero entries.

As a familiar example of recent interest, suppose that one is doing group testing of blood samples. For convenience we might be testing $p = 1000$ people and we give them labels $000, 001, \ldots, 999$. Now suppose that we know that with high probability either none of them have covid or just one of them has it. Suppose further that if we pool samples from 30 people that we can still get a valid indication of whether at least one of those 30 has covid.

We could then use group testing. We take three samples from each person. We pool one sample from everybody whose first digit was 0. If that comes back positive then we have narrowed the set of candidates to 100 people. If it does not we can do nine more tests for those with first digits 1 through 9. Next we test groups of people based on their second digits and third digits. If all 30 tests come back negative then we have learned that none of those 1000 people have it. If somebody does have it, then three of the tests will come back positive and we will have identified the person.

As an exercise, formulate this group testing into the linear regreassion model above. Figure out what X_{ij} would be and what β is and even what is assumed about ε.

Group testing is faster and less expensive when the phenomenon is sparse.

There are some closely related ideas in old puzzles about finding which coin in a set has the wrong weight in a small number of weighings. When the person who has to figure this out has a equal arm balance then putting a coin on one sides corresponds to $X_{ij} = 1$, the other side corresponds to $X_{ij} = -1$ and leaving the coin off the balance corresponds to $X_{ij} = 0$. Those puzzles are usually sequential where the outcome of one test informs the following ones.

13.3 Random balance

Random balance is an idea proposed by Satterthwaite (1959) and supported by industry experience of Budne (1959) with a discussion by Youden et al. (1959). The idea is to simply take $x_{ij} \overset{\text{iid}}{\sim} F_j$ for $i = 1, \ldots, n$. There are versions of random balance where all variables are sampled this way and in other settings perhaps only some of them are sampled randomly while others are balanced carefully in Latin squares, factorial experiments or other such designs. In a two level design we might take all $x_{ij} \overset{\text{iid}}{\sim} \mathbb{U}\{-1, 1\}$ after rescaling the predictors. The author Satterthwaite is now best known for using the method of moments to approximate a sum of weighted χ^2 random variables by a random variable having a weighted χ^2 distribution (Satterthwaite, 1946).

Random balance was a provocative proposal and it evoked a strong response. For instance, Box wrote in his discussion that "The only thing wrong with

random balance is random balance". Tukey was more supportive.

Satterthwaite defined ***exact balance*** between two variables as what we now call orthogonality: under the empirical distribution on that pair of variables, they are independent. Then random balance is just that they are sampled from a distribution where they're independent with observed values that could violate orthogonality. A variable has random balance with respect to a set of other variables if it is independently sampled conditionally on them. When all variables are sampled randomly and independently the design is one of "pure random balance". If a bad randomly drawn experiment is discarded in favor of trying again, the design is "restricted pure random balance". This process is of renewed interest in the causal inference literature, where it is known as ***rerandomization***. See Morgan and Rubin (2012) and Li and Ding (2020) for more.

Much of the controversy around random balance is about its efficiency or lack of same. Satterthwaite mentions several settings that favor random balance. Sometimes the data can be collected very cheaply. Sometimes random balance simplifies the administration of an experiment. A related point is that people with very little statistical training might be more able to run random balance experiments than others.

The reason to include random balance in this section is that one of the use cases was for high dimensional input spaces and the random balance proposal was instrumental in raising this issue. Satterthwaite claims that they're nearly efficient as exactly balanced designs, becoming more so as the number of data sets increases (page 121). However the reasoning behind his evaluations is not given.

Budne (1959) writes in favor of random balance for ***screening experiments***. Those settings have a large number of variables but only a small number of them affect the mean response, or, only a small number affect the variance of the response. He then shows some example experiments with a graphical analyiss that identifies the few important variables.

It is interesting to see the issues brought up in the paper and discussion. The discussion reveals things that the experimenters knew about but might not have emphasized in many of their other writings. Multiple comparisons were well known (Tukey had worked on them a few years earlier). Pooling bias refers to selecting the small effects to create an error estimate (without adequate adjusments). Budne refers to screening experiments for both the mean and the variance. Youden reports running experiments with "dummy treatments", where the same treatment is given two labels, like the A/A tests now used in industry. Youden notes that an experiment to identify the large effects without balancing the smaller ones adds those smaller ones to the noise variance, reducing power. It will then be difficult to find the moderately sized effects. Kempthorne makes a similar comment with more detail about power. Kempthorne also notes that random balance is being proposed without regard to what we now call selective inference issues: the multiple comparisons underlying screening many variables and the bias in forming a variance estimate from the ones found to be smaller. Box makes some analyses of the efficiency of ran-

dom balance and in particular a stepwise approach of Satterthwaite's. He finds random balance mostly inefficient but one exception arises when there is only one nonzero coefficient (extreme sparsity) and the stepwise approach is used.

Tukey was more supportive than the other authors. He mentioned that broadening the base of inference (which is a form of external validity) is a worthy tradeoff for lower efficiency. He also remarked that he learned nearly as much from Budne's scatterplots as from the complete analysis of variance. There was also some discussion about random balance being so easy to do that it gets used more than classical designs that are hard to understand.

In a stepwise analysis, we might begin by estimating a slope for each x_{ij} individually, via

$$\tilde{\beta}_j \equiv \frac{1}{n}\sum_{i=1}^{n}(x_{ij}-\bar{x}_{\bullet j})Y_{ij} \Big/ \frac{1}{n}\sum_{i=1}^{n}(x_{ij}-\bar{x}_{\bullet j})^2.$$

For a binary x_{ij} the most efficient allocation would have half of the observations at each of the two levels. With random balance they would be somewhat unequally split. The second inefficiency in random balance is that the terms $x_{ij'}\beta_{j'}$ would raise the variance of Y_{ij} in a regression model to $\sigma^2+\sum_{j'\neq j}\beta_{j'}^2\mathrm{var}(x_{ij'})$. On the other hand, in a full regression model the matrix $X^{\mathsf{T}}X/n$ is far from identity and then $\det((X^{\mathsf{T}}X)^{-1})$ is infinite for $p>n$ no matter what design is used.

13.4 Quasi-regression

Suppose that we know the distribution of the feature vector $f(\boldsymbol{x}_i)\in\mathbb{R}^p$. Then the regression parameter that minimizes squared error in predicting a finite variance value Y_i with a linear combination of these features is

$$\beta = (\mathbb{E}(f(\boldsymbol{x}_i)f(\boldsymbol{x}_i)^{\mathsf{T}}))^{-1}\mathbb{E}(f(\boldsymbol{x}_i)Y_i). \tag{13.1}$$

This β minimizes $\mathbb{E}((Y-f(\boldsymbol{x})^{\mathsf{T}}\beta)^2)$ whether or not $\mathbb{E}(Y\,|\,\boldsymbol{x})=f(\boldsymbol{x})^{\mathsf{T}}\beta$. In linear regression with $n>p$ we estimate β by

$$\hat{\beta} = \Big(\frac{1}{n}F^{\mathsf{T}}F\Big)^{-1}\Big(\frac{1}{n}F^{\mathsf{T}}Y\Big) \tag{13.2}$$

where $F\in\mathbb{R}^{n\times p}$ has i'th row $f(\boldsymbol{x}_i)$ and $Y\in\mathbb{R}^n$ has i'th component Y_i. The expectations in (14.1) have been replaced by corresponding sample averages in (14.2) to get $\hat{\beta}$.

A very popular choice is to take $f(\boldsymbol{x})$ to be products of orthogonal polynomials. The resulting expansion approximating $\mathbb{E}(Y\,|\,\boldsymbol{x})$ is known as ***polynomial chaos***.

Now suppose that $\boldsymbol{x}_i$ have been sampled at random. When we choose the sampling distribution we may well know $\mathbb{E}(f(\boldsymbol{x}_i)f(\boldsymbol{x}_i)^{\mathsf{T}})$ because we also chose the features. For instance, if the features are all polynomials and $\boldsymbol{x}_i$ have a

convenient distribution such as uniform or Gaussian we would have easily computable moments of f. In this case, we can estimate β by

$$\tilde{\beta} = (\mathbb{E}(f(\boldsymbol{x}_i)f(\boldsymbol{x}_i)^{\mathsf{T}}))^{-1}\Big(\frac{1}{n}F^{\mathsf{T}}Y\Big), \tag{13.3}$$

replacing the estimate $(F^{\mathsf{T}}F)/n$ by its expected value.

The estimate (14.3) is known as ***quasi-regression***. See An and Owen (2001) and Jiang and Owen (2003) who used quasi-regression to get interpretable approximations to black box functions. Maybe $\mathbb{E}(f(\boldsymbol{x}_i)f(\boldsymbol{x}_i)^{\mathsf{T}})$ is diagonal or block structured. If $\mathbb{E}(f(\boldsymbol{x}_i)f(\boldsymbol{x}_i)^{\mathsf{T}}) = nI$ then

$$\tilde{\beta} = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{x}_i)Y_i$$

which can be computed at $O(np)$ cost instead of $O(np^2)$ that least squares costs. Quasi-regression can be used when $p > n$ and it avoids the $O(p^2)$ space required for linear regression.

When $p > n$ then shrinkage estimators as in Jiang and Owen (2003) are advised. Ordinarily $\mathrm{var}(\tilde{\beta}) \geqslant \mathrm{var}(\hat{\beta})$ when both are possible. This is a counterexample to the usual rule in Monte Carlo sampling where plugging in a known expectation in place of a sampled quantity ordinarily helps. Blatman and Sudret (2010) report that sparse regression methods outperform methods based on numerically estimating $\mathbb{E}(ff^{\mathsf{T}})$ and $\mathbb{E}(fY)$.

13.5 Supersaturated criteria and designs

We can see in some of the discussions of Satterthwaite (1959) the beginnings of criteria to improve upon random balance for super saturated settings. We clearly cannot use D-optimality because we are sure to have the regression matrix X satisfy $\mathrm{rank}(X^{\mathsf{T}}X) \leqslant n < p$. Then $X^{\mathsf{T}}X$ is singular with $\det(X^{\mathsf{T}}X) = 0$ and then effectively "$\det((X^{\mathsf{T}}X)^{-1}) = \infty$". Georgiou (2014) is a survey of criteria and algorithms for supersaturated designs.

Another thing that changes in the supersaturated setting is that we will need adaptive methods that decide which β_j to estimate and which to leave at a default value like 0. These adaptive methods will have to use the observed Y_i values. In the end the estimate of β is not ordinarily a linear combination of Y_i like it is in least squares. The variance of the estimated β becomes more complicated and will depend on what the true β is. For instance if the true β is all zeros except the intercept then methods based on sparsity could be very accurate and even get most of the coefficients exactly right. If instead β has all nonzero elements of roughly equal size, perhaps described by $\beta_j \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ then no algorithm can find them. In other words, the accuracy with which β can be estimated now depends on the true value of β, because even though the model relating $\mathbb{E}(Y)$ to X is linear in X, the estimator of β is not linear in Y_i.

Booth and Cox (1962) citing Box's discussion of random balance introduce some criteria. The consider the matrix $X \in \{-1, 1\}^{n\times p}$ with $p > n-1$ and each column of X containing $n/2$ values for each of ± 1. Then they consider the matrix $S = X^{\mathsf{T}}X$ which is proportional to sample correlations among the predictors. Their criterion is $\max_{1\leqslant j<j'\leqslant p} |S_{jj'}|$. They break ties by preferring designs where a smaller number of column pairs attain the maximum of $|S_{jj'}|$. They give some small examples with n and p of a few dozens. The examples were found by computer search. When compared to random balance the designs they obtain have much better values of S. They also report the variance of $|S_{jj'}|$ values.

Georgiou (2014) considers the criterion

$$\mathbb{E}(S^2) \equiv \frac{1}{\binom{p}{2}} \sum_{1\leqslant j<j'\leqslant p} S_{jj'}^2$$

and remarks that algorithms that try to optimize it may possibly yield identical columns. Incidentally, Georgiou (2014) gives this as the first criterion that Booth and Cox (1962) consider but in preparing these notes I do not find that in their article.

Georgiou (2014) reports some lower bounds on $\mathbb{E}(s^2)$. The precise bounds depend on things like the value of n modulo 4. In all cases

$$\mathbb{E}(S^2) \geqslant \frac{n^2(p-n+1)}{(n-1)(p-1)}$$

where his definition of $\mathbb{E}(S^2)$ includes the intercept column of all ones. If we normalize each $S_{jj'}$ to $S_{jj'}/n$ then

$$\mathbb{E}((S/n)^2) \geqslant \frac{p-n+1}{(n-1)(p-1)}.$$

Let's suppose that $n \gg 1$ and $p - n \gg 1$. Then ignoring the ± 1s above, we get

$$\mathbb{E}((S/n)^2) \geqslant \frac{p-n+1}{(n-1)(p-1)} \approx \frac{p-n}{np} = \frac{1-n/p}{n}.$$

Exercise: what would we get for $X_{ij} \overset{\text{iid}}{\sim} \mathbb{U}\{-1, 1\}$? What would we get if half of column j were randomly chosen to be each of ± 1 and column j' were chosen that way too but independently of column j?

Georgiou (2014) presents numerous design strategies for the supersaturated setting. Of special interest is the proposal of Lin (1993). This design works with a Hadamard matrix H_n. It picks one of the columns and keeps only the $n/2$ runs with $+1$ in that column. This is appealing because Hadamard matrices are so abundant and both the Sylvester and first Paley constructions are easy to use. Lin's approach provides $n/2$ experimental runs in $p = n-2$ variables. One of the n colunns is lost to the intercept term and one more is lost because of the column chosen to define the selected runs. The specific column chosen

makes makes little difference. Exercise: compare the $\mathbb{E}(S^2)$ criterion that Lin gets to what he would get choosing runs with -1 in the given column.

Lin (1993) runs forward stepwise regression on data from his design. Phoa et al. (2009) use L_1 regularization based on the Dantzig selector of Candes and Tao (2007).

Lin (1995) looks for ways to maximize the number p of binary variables in a model subject to a constraint on $\max_{j\neq j'}|S_{jj'}|$. He breaks ties based on the number of pairs at the same maximum level.

Practical investigations: how would it go to choose 1/4 or 1/8 et cetera of a Hadamard design? Would Paley or Sylvester constructions be about equally good or would one be better? This last point requires a way to make a fair comparison between Hadamard designs with different values of n.

13.6 Designs for compressed sensing

Supersaturated designs are well suited to settings where we have $p > n$ but can reasonably expect β to be sparse or nearly so. Donoho (2006) describes compressed sensing and Tibshirani (1996) introduces the ***lasso***, both of which can be used when $p > n$. Krahmer and Ward (2011) discusses experimental design for this setting. This section only surveys the issue, because a detailed discussion goes beyond prerequisites for this course.

In this context it is desireable for the matrix $X \in \mathbb{R}^{n\times p}$ (excluding intercept) to satisfy the ***restricted isometry property*** (RIP), defined in terms of a level $\delta \in (0,1)$ and an integer order $k > 0$. The vector $v \in \mathbb{R}^p$ is said to be ***k-sparse*** if $\sum_{j=1}^p 1_{v_j\neq 0} \leqslant k$. Then X satisfies the (k,δ)–RIP if

$$(1-\delta)\|v\|^2 \leqslant \|Xv\|^2 \leqslant (1+\delta)\|v\|^2$$

holds for all k-sparse v. Krahmer and Ward (2011) give RIP conditions (and references) where minimizing $\|X\beta\|_1$ subject to $X\beta = Y$ exactly recovers a sparse β. If β is sparse and there is no noise it can be recovered by L_1 penalized regression. That is, a very tractable convex optimization problem can be used to find β where searching for a sparse solution would otherwise be quite costly. There are generalizations to handle additive $Y = X\beta + \varepsilon$ (i.e., measured with noise) but sparsity remains a critical ingredient.

One of the most basic constructions of a matrix with an RIP property is to take a random subset of rows of a Hadamard matrix. The resulting design will satisfy an RIP property with very high probability. Specifically, for p predictors we can ignore the intercept column and take n randomly selected rows (without replacement) from H_{4m} where $4m \geqslant p+1$. Compared to the method of Lin (1993) this approach allows n other than $4m/2 = 2m$. When we want half of the rows it makes more sense to use Lin's design because then every column will be equally split between ± 1 values. The designs described for compressed sensing require only very small values of n, just over $\delta^{-2}k\log(p)^4$.

A related approach is to choose a random subset from a discrete Fourier matrix. That matrix has complex entries $X_{jk} = \omega^{jk}/\sqrt{n}$ for $0 \leqslant j,k < n$ (note

the zero based indexing) where $\omega = \exp(2\pi\sqrt{-1}/n)$. They split out the real and imaginary parts of X_{ij} doubling the number of columns obtained. (Those $2n$ real valued vectors in $\mathbb{R}^n$ cannot of course be mutually orthogonal.)

Another approach they describe is to take $X \in \mathbb{R}^{n\times p}$ with IID $\mathcal{N}(0,1)$ or IID $\mathbb{U}\{-1,1\}$ entries multiplied by $\sqrt{p/n}$. The amazing thing is that this, after many years, provides some justification for Satterthwaite's random balance.

Part of the argument in Krahmer and Ward (2011) is based on the ***Johnson-Lindenstrauss*** lemma (Johnson and Lindenstrauss, 1984). Think of a saturated design as $p-1$ column vectors in $\mathbb{R}^p$, orthogonal to each other and to the vector of ones. Now we project those columns $v_1,\dots,v_{p-1}$ into a lower dimensional space by multiplying them by a matrix Φ. That is we get $u_i = \Phi v_i \in \mathbb{R}^n$ for $i = 1,\dots,n$ for a matrix $\Phi \in \mathbb{R}^{n\times p}$. The Johnson-Lindenstrauss lemma shows that there is a mapping from $\mathbb{R}^p$ to $\mathbb{R}^n$ that preserves interpoint distances to within ϵ. When that mapping is the linear one described above this means that

$$(1-\epsilon)\|v_j - v_{j'}\|^2 \leqslant \|u_j - u_{j'}\|^2 \leqslant (1+\epsilon)\|v_j - v_{j'}\|^2.$$

The original Johnson-Lindenstrauss Lemma allowed a nonlinear but Lipschitz continuous function $u_j = g(v_j)$ instead of the $u_j = \Phi v_j$ that we use here. If the interpoint distances are nearly preserved in this mapping then so are angles. Think of a triangle defined by v_j, $v_{j'}$ and $v_{j''}$. If all three interpoint distances are nearly preserved then so are all three angles defined by the points. Now when v_j are orthogonal (right angles) then the u_j are nearly orthogonal.

In the Johnson-Lindenstrauss lemma the number n does not have to be very large compared to p. It only needs to be $O(\epsilon^{-2}\log(p))$.

Bibliography

An, J. and Owen, A. (2001). Quasi-regression. *Journal of complexity*, 17(4):588–607.

Blatman, G. and Sudret, B. (2010). Efficient computation of global sensitivity indices using sparse polynomial chaos expansions. *Reliability Engineering & System Safety*, 95(11):1216–1229.

Booth, K. H. and Cox, D. R. (1962). Some systematic supersaturated designs. *Technometrics*, 4(4):489–495.

Budne, T. A. (1959). The application of random balance designs. *Technometrics*, 1(2):139–155.

Candes, E. and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The annals of Statistics*, 35(6):2313–2351.

Donoho, D. L. (2006). Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306.

Georgiou, S. D. (2014). Supersaturated designs: A review of their construction and analysis. *Journal of Statistical Planning and Inference*, 144:92–109.

Hedayat, A. S., Sloane, N. J. A., and Stufken, J. (1999). *Orthogonal arrays: theory and applications*. Springer Science & Business Media, New York.

Jiang, T. and Owen, A. B. (2003). Quasi-regression with shrinkage. *Mathematics and Computers in Simulation*, 62(3-6):231–241.

Johnson, W. B. and Lindenstrauss, J. (1984). Extensions of Lipschitz mappings into a Hilbert space. *Contemporary mathematics*, 26(189–206):1.

Krahmer, F. and Ward, R. (2011). New and improved Johnson–Lindenstrauss embeddings via the restricted isometry property. *SIAM Journal on Mathematical Analysis*, 43(3):1269–1281.

Li, X. and Ding, P. (2020). Rerandomization and regression adjustment. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Lin, D. K. (1993). A new class of supersaturated designs. *Technometrics*, 35(1):28–31.

Lin, D. K. (1995). Generating systematic supersaturated designs. *Technometrics*, 37(2):213–225.

Morgan, K. L. and Rubin, D. B. (2012). Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, 40(2):1263–1282.

Paley, R. E. (1933). On orthogonal matrices. *Journal of Mathematics and Physics*, 12(1-4):311–320.

Phoa, F. K., Pan, Y.-H., and Xu, H. (2009). Analysis of supersaturated designs via the Dantzig selector. *Journal of Statistical Planning and Inference*, 139(7):2362–2372.

Satterthwaite, F. (1959). Random balance experimentation. *Technometrics*, 1(2):111–137.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics bulletin*, 2(6):110–114.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288.

Youden, W., Kempthorne, O., Tukey, J. W., Box, G., and Hunter, J. (1959). Discussion of the papers of messrs. satterthwaite and budne. *Technometrics*, 1(2):157–184.