ROBUSTNESS AND ROBUSTNESS OF DESIGN

Doug Wiens*

Summer 2012

*© Douglas P. Wiens, Department of Mathematical & Statistical Sciences, Faculty of Science, University of Alberta (2012).
Contents

1 General Concepts of Robustness 5
   1.1 Introduction to Robustness ................. 6
   1.2 Robust Regression .......................... 23
   1.3 Minimax Robust Estimation ............... 51

2 Robustness of Design 69
   2.1 Classical design theory .................... 70
   2.2 Robustness concepts in design ............ 86
Preface

For the ‘source material’ – which you are expected to familiarize yourself with before the presentations, up to the point of being able to answer the questions which are posed in the presentations – I use the following abbreviations:


• [W578] My Stat 578 notes on the course web site.


The papers of which I am an author can be downloaded from my web site.
Part 1

General Concepts of Robustness
1.1. Introduction to Robustness

Source material:
[Hb], ch. 1-3; [Hm], ch. 1-3; [MMY], ch. 1-3

1.1.1. Why robustness?

• What can go wrong?

  – Outliers (≡ sample values that are surprising in relation to the majority of the sample). Even quite ‘clean’ data sets typically contain 5-10% erroneous values – transcription errors, machine errors (electronic recording → interference), sampling an unintended population, etc.

  – Model errors (in design, for instance).

  – Unusually influential \(x\)-values in a regression.
• Undetected errors can completely ruin a classically ‘optimal’ method. ‘Optimal’ usually refers to behaviour at a particular model – commonly Normal. The behaviour can deteriorate very quickly – perhaps instantaneously – as one moves away from this model.

  – Example: sample mean vs. sample median.
    ‘Breakdown point’ = 0% vs. about 50%.

    *Someone will give examples illustrating each of these statements (using ‘breakdown’ in an informal sense).*

• Can’t we just ‘clean’ the data first’ so as to remove outliers?

  – Users typically don’t, even if we want them to.

  – Downweighting of dubious observations is typically more efficient that an accept/reject decision.
- Multivariate data – very tough to spot outliers.

- Cleaning requires the distribution theory to be adjusted; in particular the variances will be underestimated from the cleaned data.

- Example: The ‘optimal’ (at the Normal) estimate of a population variance \( \sigma^2 \) is the sample variance \( S^2 \). A competing estimate is the square of the mean absolute deviation

\[
d = \frac{1}{n} \sum |X_i - \bar{X}|.
\]

At the Normal \( S^2 \xrightarrow{pr} \sigma^2 \) but \( d^2 \xrightarrow{pr} \frac{2}{\pi} \sigma^2 \) [someone will explain why - briefly, informally and without a lot of details], so we consider \( \tilde{\sigma}^2 = \frac{\pi}{2} d^2 \). The Asymptotic Relative Efficiency of \( \tilde{\sigma}^2 \) relative to \( S^2 \) is

\[
ARE = \lim_{n} \frac{\text{var} \left[ S^2 \right]}{\text{var} \left[ \tilde{\sigma}^2 \right]},
\]

and at the Normal this is .876. (So \( \tilde{\sigma}^2 \) is almost as efficient as \( S^2 \) even at the Normal.) Suppose
now that we are in fact sampling from an ‘$\varepsilon-$contaminated Normal distribution’, with d.f.

$$F(x) = (1 - \varepsilon) \Phi \left( \frac{x}{\sigma} \right) + \varepsilon \Phi \left( \frac{x}{3\sigma} \right).$$

Interpretation?

\[
\begin{array}{c|ccccc}
\varepsilon & 0 & 0.001 & 0.002 & 0.01 & 0.05 \\
\text{ARE} & 0.876 & 0.948 & 1.016 & 1.44 & 2.04 \\
\end{array}
\]

Would you know if 1% of your data came from a more highly varied population?

• Even $d$ is very non-robust (due to its dependence on simple averaging). Better choices are

$$IQR = X_{(.75n)} - X_{(.25n)},$$

$$MAD = \text{med}_i \left\{ \left| X_i - \text{med}_j \{ X_j \} \right| \right\}.$$  

Each is often normalized for consistency at $\Phi$, for instance ‘MAD’ often refers to that above divided by $0.6745 = \Phi^{-1}(0.75)$ \textit{[someone will explain why - what does $\Phi^{-1}(0.75)$ represent?]}.  

• The ‘location model’ is the simplest and most widely studied; it is one for which the sample arises from a population with density $f(x - \theta)$ for some parameter $\theta$. For the moment we will also assume that $f(x)$ is symmetric around 0; thus $\theta = \text{median} = \text{mean}$ [if this exists – what is an example in which it doesn't?].

- Possible estimates: sample mean (completely non-robust), median (robust but inefficient), trimmed mean (a class containing both the mean and median).

1.1.2. Some common robust estimates

• **M-estimation of location.** This generalizes maximum likelihood estimation. The mle of a location parameter is the maximizer of the log-likelihood $\sum \log f(x_i - \theta)$, hence satisfies the likelihood equation

$$\sum \psi(x_i - \theta) = 0, \quad (1.1.1)$$
for $\psi = -f'/f$. An M-estimate of location is a solution to (1.1.1) for a user-chosen ‘score’ function $\psi$; more generally it is a minimizer of $\sum \rho(x_i - \theta)$ (and then $\psi = \rho'$).

- **What is $\psi$ for the mean? median?**

- To make the estimate scale equivariant we include a (robust) estimate of scale as well — how?

- Huber’s $\psi_c$ — a compromise between the mean and median, also is minimax in the class

  \[ \{ F = (1 - \varepsilon) \Phi + \varepsilon G; G \text{ symmetric} \} \]

  of contaminated Normal distributions.

- To compute — Newton-Raphson? IRLS? [What are these? How are they implemented?]

- To analyze such estimates we represent them as functions of the empirical distribution function:

  \[ \hat{F}_n(x) = \frac{\# \text{ of observations } \leq x}{n}. \]
Note that for fixed $x$, $n\hat{F}_n(x)$ is a $\text{bin}(n, F(x))$ r.v., and so $\hat{F}_n(x) \overset{pr}{\to} F(x)$. Also

$$\int g(x)d\hat{F}_n(x) = \frac{1}{n} \sum g(x_i) \overset{pr}{\to} E_F[g(X)] = \int g(x)dF(x).$$

In particular an M-estimate satisfies

$$\int \psi(x - \hat{\theta})d\hat{F}_n(x) = 0.$$ 

We write $\hat{\theta} = T(\hat{F}_n)$, where $T = T(\hat{F}_n)$ is defined implicitly by $\int \psi(x - T)d\hat{F}_n(x) = 0$. Under some regularity, $T(\hat{F}_n) \overset{pr}{\to} T(F)$ as $n \to \infty$, and $T(F)$ is defined here by

$$\int \psi(x - T(F))dF(x) = 0.$$ 

It is this form which is analyzed in order to ascertain the asymptotic properties of the M-estimate.

- **L-estimation.** An L-estimate is a Linear combination of (a function of) order statistics:

$$\hat{\theta} = \sum a_{ni} h(X(i)),$$
where $X_{(i)}$ is the $i^{th}$ order statistic:

$$X(1) \leq \cdots \leq X(n).$$

We define the quantile function by

$$F^{-1}(t) = \min \{ x \mid F(x) \geq t \},$$

for $0 < t < 1$. Then [why?] $X(i) = \hat{F}_n^{-1}(\frac{i}{n})$ and so if the weights are generated by $a_{ni} = M \left\{ \left( \frac{i-1}{n}, \frac{i}{n} \right) \right\}$ for some measure on $(0, 1)$ with total mass of 1, we have [why?]

$$\hat{\theta} = \int h \left( \hat{F}_n^{-1}(t) \right) M(dt).$$

[This defines $\hat{\theta} = T \left( \hat{F}_n \right)$, then what is $T \left( F \right)$?]

- Example: If $a_{ni*} = M \left\{ \left( \frac{i^*-1}{n}, \frac{i^*}{n} \right) \right\} = 1$, all others = 0 then (with $h(x) = x$) we have $\hat{\theta} = X(i^*)$. In particular $i^* = n/2$ gives the median.

- Example: The $\alpha$-trimmed mean derives from a measure $M$ with density

$$m(t) = \frac{1}{1 - 2\alpha} I (\alpha < t < 1 - \alpha),$$
and then
\[ T(F) = \int F^{-1}(t)m(t)dt = \frac{1}{1 - 2\alpha} \int_{\alpha}^{1-\alpha} F^{-1}(t)dt. \]

- **R-estimation.** An R-estimate is obtained by inverting a rank test. With two samples \(X_1, \ldots, X_m\) and \(Y_1, \ldots, Y_n\) from \(F(x)\) and \(G(x) = F(x - \theta)\) respectively we test \(\theta = 0\) against \(\theta > 0\) by rejecting for small values of

\[ S_{m,n} = \frac{1}{m} \sum_{i=1}^{m} a(R_i), \]

where \(R_i\) is the rank of \(X_i\) in the combined sample. The scores are generated by, e.g.,

\[ a_i = J \left( \frac{i}{m + n + 1} \right), \text{ or} \]

\[ a_i = (m + n) \int_{(i-1)/(m+n)}^{i/(m+n)} J(t)dt. \]

e.g. for the Wilcoxon test, \(J(t) = t - \frac{1}{2}\). To estimate location, take \(m = n\) and \(Y_i = 2\tilde{\theta} - X_i\); adjust \(\theta\) so that \(S_{n,n} \approx 0\). This leads to the Hodges-Lehmann estimate \(\hat{\theta} = med \left\{ \frac{X_i + X_j}{2} \right\} \) (over all \(n^2\))
pairs, or only those with $i < j$; these are asymptotically equivalent).

1.1.3. Robustness measures: Influence function, breakdown point

- Suppose that we sample from a population with d.f. $(1 - \varepsilon) F + \varepsilon G$, resulting in an estimate with representation $\hat{\theta} = T ((1 - \varepsilon) F + \varepsilon G)$. The *breakdown point* $\varepsilon^* = \varepsilon^* (\hat{\theta}, F)$ of $\hat{\theta}$ at $F$ is the largest $\varepsilon^* \in (0, 1)$ such that for $\varepsilon < \varepsilon^*$ the estimate $\hat{\theta}$, viewed as a function of $G$, remains bounded (and also bounded away from the boundary of the parameter space). It is often evaluated as $\lim_{n \to \infty} \varepsilon^* (\hat{\theta}, F_n)$.

- Example: $\hat{\theta} = \bar{X}$ has $\varepsilon^* (\hat{\theta}, \Phi) = 0$.

- Example: $\hat{\theta} = med$ has $\varepsilon^* (\hat{\theta}, F) = .5$. 
• **Definitions:** The *Kolmogorov distance* between d.f.s $F$ and $G$ is

$$d(F, G) = \sup_x |F(x) - G(x)| \quad (\leq 1).$$

A functional $T$ is *continuous* at $F$ if

$$d(F_n, F) \to 0 \Rightarrow T(F_n) \to T(F).$$

• **Theorem:** If $X_1, \ldots, X_n \overset{i.i.d.}{\sim} F$ and $T(\cdot)$ is continuous at $F$, then $T(\hat{F}_n)$ is consistent for $T(F)$.

To prove this, one shows that $d(\hat{F}_n, F) \overset{pr}{\to} 0$; this is (a weak version of) the celebrated Glivenko-Cantelli Theorem. (Pointwise convergence is merely the WLLN.)

• **Example.** $T(F) = E_F[X]$. This ‘mean functional’ is not continuous [why not?]. A consequence is that $T(\hat{F}_n) = \bar{X}$ need not be consistent for $E_F[X]$ in the presence of outliers.
• **Influence functions.** Let $T(F)$ be a functional defined for $F \in \mathcal{F}$, a convex class of d.f.s:

$$F_0, F_1 \in \mathcal{F} \Rightarrow F_\varepsilon \overset{def}{=} (1 - \varepsilon)F_0 + \varepsilon F_1 \in \mathcal{F},$$

for $0 \leq \varepsilon \leq 1$. Consider

$$\dot{T}(F_0; F_1) = \lim_{\varepsilon \to 0} \frac{T((1 - \varepsilon)F_0 + \varepsilon F_1) - T(F_0)}{\varepsilon} = \frac{d}{d\varepsilon} T(F_\varepsilon)_{|\varepsilon=0}.$$

When $F_1 = \delta_x$ (point mass at $x$) this represents the limiting, normalized influence of a new observation, with value $x$, on the statistic $T(F_0)$. We call

$$\dot{T}(F_0; \delta_x) = IF(x) \ (or \ IF(x; T, F_0))$$

the **Influence Function.** It can be used as a measure of the robustness of a procedure against outliers (ideally we would like it to be bounded); it can also be used to give a quick asymptotic normality proof.
- e.g. let \( T(F) = E_F[X] \). The IF [what is it?] is unbounded – further evidence of the lack of robustness of the sample average. Indeed, a single arbitrarily large outlier can push \( \bar{X} \) beyond all bounds.

- This derivation highlights that

\[
\dot{T}(F_0; F_1) = E_{F_1}[IF(X)]
\]

and

\[
E_{F_0}[IF(X)] = 0;
\]

these turn out to hold quite generally.

- Derive the IF of an M-estimate \( \hat{\theta} = T(\hat{F}_n) \), with \( \theta = T(F) \) defined by

\[
\int \psi(x; \theta) dF(x) = 0;
\]

it is

\[
IF(x; T, F_0) = \frac{\psi(x; \theta_0)}{E[\psi'(x; \theta_0)]},
\]

(where \( \psi' \) means what?)
Similarly, under appropriate regularity conditions the IF of an L-estimate of location is

\[ IF(x; T, F) = \int_{-\infty}^{x} h'(y) m(F(y)) dy \]  

\[ - \int_{-\infty}^{\infty} (1 - F(y)) h'(y) m(F(y)) dy. \]

[The tricky part here is to differentiate the relationship \( t = F_\varepsilon \left( F_{\varepsilon}^{-1}(t) \right) \) w.r.t. \( \varepsilon \)- how is this done?]

- **Asymptotic normality.** Suppose \( X_1, \ldots, X_n \overset{i.i.d.}{\sim} F_0 \). By Taylor’s Theorem, expanding \( T(F_\varepsilon) \) around \( \varepsilon = 0 \) gives

\[ T(F_\varepsilon) = T(F_0) + \dot{T}(F_0; F_1)\varepsilon + o(\varepsilon), \text{ whence } \]

\[ T(F_1) = T(F_0) + \dot{T}(F_0; F_1) + \text{Remainder}. \]

Typically

\[ \dot{T}(F_0; F_1) = E_{F_1}[\psi(X)] \]  

(1.1.3)

for some function \( \psi \). With \( F_1 = \delta_x \) we obtain

\( IF(x) = \psi(x) \). With \( F_1 = F_0 \) we obtain [how?]

\[ E_{F_0}[IF(X)] = 0. \]
Thus

\[ T(F_1) = T(F_0) + E_{F_1}[IF(X)] + \text{Remainder} \]

and then, with \( F_1 = \hat{F}_n \), we have (a ‘Mean Value Theorem’)

\[
\sqrt{n} \left( T(\hat{F}_n) - T(F_0) \right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} IF(X_i) + \sqrt{n} R_n,
\]

where the \( IF(X_i) \) are i.i.d. r.v.s with mean 0 and variance

\[ \gamma^2(F_0) = E_{F_0}[IF^2(X)]. \]

By the CLT and Slutsky’s theorem’ if

\[ \sqrt{n} R_n \xrightarrow{pr} 0 \quad (1.1.4) \]

where

\[ R_n = T(\hat{F}_n) - T(F_0) - \frac{1}{n} \sum_{i=1}^{n} IF(X_i), \]

we have

\[ \sqrt{n} \left( T(\hat{F}_n) - T(F_0) \right) \xrightarrow{L} N \left( 0, \gamma^2(F_0) \right). \quad (1.1.5) \]
• Sometimes we just use (1.1.5) as a guide to what to expect, and then prove it using another technique. In other cases we can verify (1.1.3) and (1.1.4) so as to infer (1.1.5).

• Evaluate the asymptotic variance of an $M$-estimate. What then is it for the median?

• A further problem: Under appropriate regularity conditions, if $\hat{\theta}$ is an asymptotically normal estimate of location computed on the basis of a sample from $F$:

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{L} N(0,V(F)),$$

then [why? – what is being referred to here?]

$$V(F) \geq 1/I(F),$$

where, with $\psi_F(x) \overset{def}{=} -f'(x)/f(x)$, $I(F) = \int \psi_F^2(x)f(x)dx$ is Fisher information. If equality is attained, we say that $\hat{\theta}$ is asymptotically efficient. Suppose that $F$ is known. Show that
then the M-estimate of location is asymptotically efficient if (when $\theta = 0$)

$$IF(x; T, F) = \psi_F(x) / I(F).$$

(1.1.6)

Show further that condition (1.1.6) is attained for M-estimation if $\psi(x)$ is any multiple of $\psi_F(x)$. 
1.2. Robust Regression

Source material: [MMY] ch. 4, 5; [RL] ch. 2, 6; [W578] §IV

1.2.1. The need for robustness; M-estimation

• Robustness deals with the behaviour of statistical methods under violations of the assumptions, and with the derivation of methods which work ‘almost’ as well when these assumptions are violated as when they hold.

• Under what assumptions is Least Squares an optimal estimation method? The Gauss-Markov Theorem states that LS is optimal (= MVU), in the linear model $Y = X\theta + \varepsilon$, if (i) the errors are uncorrelated and homoscedastic, and (ii) we insist on estimates which are linear in the $y_i$. [What is the precise statement of the Gauss-Markov Theorem?] If these assumptions hold then to improve on LS we should look among non-linear estimates.
• Large sample inferences impose a further requirement. We typically carry out inferences about $a^T \theta$ by using the normal approximation

$$a^T \hat{\theta}_{OLS} \overset{d}{\approx} N \left( a^T \theta, \sigma^2 \varepsilon a^T (X^T X)^{-1} a \right),$$

valid asymptotically even for non-normal errors under a condition that states, roughly, that no observations can have too large an influence on the fit. More precisely, in order that all LSEs $a^T \hat{\theta}_{OLS}$ be asymptotically normal, it is necessary and sufficient that ‘Huber’s condition’ hold:

$$\max_i h_{ii} \to 0 \text{ as } n \to \infty.$$ 

In regression diagnostics we tend to be wary of observations with $h_{ii} > 2\bar{h} = 2p/n$.

• Observations which dominate the LS fit due to unusual x-values are ‘leverage’ points (and the $h_{ii}$ are sometimes called leverage values). Observations with unusually large (in absolute value) y-values are ‘outliers’.
Simulated data: \( Y = x + \varepsilon \) with, on the left, one additional observation which is both highly influential (extreme \( x \)– value) and has an outlying \( y \)– value. Plot on right is after removal of this point.

- Since LS = ML for Normal errors, in looking for robust alternatives we might start with ML estimation for other distributions. If \( Y_i = x_i^T \theta + \varepsilon_i \), with

\[
\varepsilon_i \sim F \left( \frac{\varepsilon}{\sigma} \right), \text{ density } \frac{1}{\sigma} f \left( \frac{\varepsilon}{\sigma} \right),
\]
then $Y_i$ has density $\frac{1}{\sigma} f \left( \frac{y_i - x_i^T \theta}{\sigma} \right)$. Represent the MLE as the solution to a certain minimization problem (in terms of $\rho(r) = -\log f(r)$); go on to obtain the likelihood equations

$$\frac{1}{n} \sum_i \psi \left( \frac{y_i - x_i^T \theta}{\sigma} \right) x_i = 0,$$

$$\frac{1}{n} \sum_i \psi \left( \frac{y_i - x_i^T \theta}{\sigma} \right) \left( \frac{y_i - x_i^T \theta}{\sigma} \right) - 1 = 0,$$

(1.2.1)

with ‘score function’ $\psi = \rho' = -f'/f$. Show that each LHS $\xrightarrow{pr} 0$.

• An M-estimate of regression is a solution to

$$\frac{1}{n} \sum_i \rho \left( \frac{r_i (\hat{\theta})}{\hat{\sigma}} \right) = \text{min}, \text{ or of }$$

$$\frac{1}{n} \sum_i \psi \left( \frac{r_i (\hat{\theta})}{\hat{\sigma}} \right) x_i = 0,$$

where $r_i (\hat{\theta}) = y_i - x_i^T \hat{\theta}$ is the residual, and $\hat{\sigma}$ is an estimate of scale, perhaps determined by a version of (1.2.1).
1.2.2. Huber’s $\psi_c$; Computing M-estimates

- The LS estimate ($\psi(r) = \text{what?}$) allows large residuals to have a large influence on the fit, and is non-robust for this reason. The L1 estimate ($\psi(r) = \text{what?}$) gives all residuals the same influence; for this reason it is highly robust but not very efficient if the errors are Normal. A compromise is Huber’s $\psi$:

$$
\psi_c(r) = \begin{cases} 
  r, & |r| \leq c, \\
  c \cdot \text{sgn}(r), & |r| \geq c.
\end{cases}
$$

- In practice, one typically takes $c \in (1, 2)$.

- Computing. Suppose first that scale $\sigma$ is known, and so we wish only to compute $\hat{\theta}$ by solving

$$
\frac{1}{n} \sum_i \rho \left( \frac{r_i(\hat{\theta})}{\sigma} \right) = \min, \text{ or }
$$

$$
\frac{1}{n} \sum_i \psi \left( \frac{r_i(\hat{\theta})}{\sigma} \right) x_i = 0. \quad (1.2.2)
$$
Suppose as well that $\psi$ is monotone, i.e. $\psi' \geq 0$. If $\psi$ is strictly increasing then we are guaranteed a unique solution, since the function being minimized is convex [Why? Prove this. Show also that for Huber’s $\psi$ we have only that the Hessian is positive semi-definite, and that indeed $\theta$ can be chosen so badly that all $|r_i(\theta)| > c\sigma$ and so the Hessian is the zero matrix. But the objective function is still convex in a neighbourhood of a solution $\hat{\theta}$ for which most of the residuals satisfy $|r_i(\hat{\theta})| \leq c\sigma$.]

- We will always assume that $\psi$ is an odd function (and so $\psi(0) = 0$ if $\psi$ is continuous), and $\psi(r) \geq 0$ for positive $r$. Introduce ‘weights’ $w(x) = \psi(x)/x$ ($= \psi'(0)$ at $x = 0$); thus $w(x)$ is even and everywhere non-negative. Then with $w_i = w\left(r_i(\hat{\theta})/\sigma\right)$, (1.2.2) can be written

$$\frac{1}{n} \sum_{i} \left( y_i - x_i^T \hat{\theta} \right) w_i x_i = 0,$$
with ‘solution’

$$\hat{\theta} = \left(X^T W X\right)^{-1} X^T W y.$$ 

This is only a ‘solution’ because the weights depend on $\hat{\theta}$. But we can do IRLS:

1. Start with $\theta(0)$; compute residuals $y_i - x_i^T \theta(0)$ and weights $w_{i,(0)} = w \left( r_i \left( \theta(0) \right) / \sigma \right)$.

2. Do a WLS regression of $y$ on $X$ with weights $w_{i,(0)}$ to obtain $\theta(1)$.

3. Iterate to convergence.

• Typically scale is to be estimated as well, and we replace $\sigma$ by $\hat{\sigma}$ in these expressions. Commonly one uses $\hat{\sigma} = \text{MAD}$. Another common approach is to update $\hat{\sigma}$ along with $\hat{\theta}$: after $\theta(k)$ and $r_i \left( \theta(k) \right)$ have been obtained, update $\hat{\sigma}$ to $\sigma(k+1)$ by solving (in analogy with LS)

$$\frac{1}{n - p} \sum_i \psi^2 \left( \frac{y_i - x_i^T \theta}{\sigma} \right) = E_\Phi \left[ \psi^2 \left( \frac{\varepsilon}{\sigma} \right) \right] \overset{\text{def}}{=} \delta$$
through (how? – write in terms of \( w \left( r_i \left( \theta(k) \right) / \sigma(k) \right) \))

\[
\sigma^2_{(k+1)} = \frac{\sigma^2(k)}{\delta(n - p)} \sum_i \psi^2 \left( \frac{r_i^2(\theta(k))}{\sigma^2(k)} \right).
\]

### 1.2.3. Asymptotics; Inferences

- Making inferences requires an approximate distribution of \( \hat{\theta} \). Techniques as used in STAT665 to establish the asymptotic normality of the MLE work here, with minor changes. We must make the key assumption:

\[
E_F \left[ \psi \left( \frac{\varepsilon}{\sigma} \right) \right] = 0
\]

(where \( \varepsilon \sim F \)) and then we obtain

\[
\hat{\theta} \approx N \left( \theta_0, \sigma^2 \frac{E_F \left[ \psi^2 \left( \frac{\varepsilon}{\sigma} \right) \right]}{\left( E_F \left[ \psi' \left( \frac{\varepsilon}{\sigma} \right) \right] \right)^2} \left[ \sum_i x_i x_i^T \right]^{-1} \right).
\]

- Or, we can use the IF, which for an ordinary M-
estimate defined by

\[ \sum \psi(z_i; \theta) = 0 \text{ for } \psi(z; \theta) = \psi \left( \frac{y - x^T \theta}{\sigma} \right) x_i \]

can be derived just as earlier, and is

\[ IF(z) = \left\{ E_F \left[ -\frac{\partial}{\partial \theta} \psi(Z; \theta) \right] \right\}^{-1} \psi(z; \theta). \]

The AN result from §1.1 generalizes to

\[ \sqrt{n} \left( \hat{\theta} - \theta \right) \xrightarrow{L} N \left( 0, E \left[ IF \cdot IF^T \right] \right), \]

where

\[ E \left[ IF \cdot IF^T \right] = M^{-1}QM^{-1} \]

with

\[ M = E_F \left[ -\frac{\partial}{\partial \theta} \psi(Z; \theta) \right], \]
\[ Q = E_F \left[ \psi(Z; \theta) \psi^T(Z; \theta) \right]. \]

- When scale is estimated as well, we replace \( \sigma \) by \( \hat{\sigma} \) in order to apply the approximation, which we can also write as

\[ \hat{\theta} \xrightarrow{d} N \left( \theta_0, V(\psi, F') \left( X^T X \right)^{-1} \right) \]
where
\[
V (\psi, F) = \sigma^2 \frac{E_F \left[ \psi^2 \left( \frac{\varepsilon}{\sigma} \right) \right]}{\left( E_F \left[ \psi' \left( \frac{\varepsilon}{\sigma} \right) \right] \right)^2}
\]
is estimated by
\[
v_{\psi} = \hat{\sigma}^2 \frac{1}{n-p} \sum_i \psi^2 \left( \frac{r_i(\hat{\theta})}{\hat{\sigma}} \right) \left[ \frac{1}{n} \sum_i \psi' \left( \frac{r_i(\hat{\theta})}{\hat{\sigma}} \right) \right]^2.
\]
– It follows, exactly as it will in §1.3, that Huber’s \( \psi_c \) yields a minimax estimate.

- Inferences can be made in much the same way as when least squares estimates are used, after making appropriate modifications for the revised covariance structure of \( \hat{\theta} \). For instance, tests and confidence intervals on \( a^T \theta \) use the approximation
\[
a^T \hat{\theta} \overset{d}{\approx} N \left( a^T \theta, V (\psi, F) a^T (X^T X)^{-1} a \right),
\]
with

\[
\frac{a^T (\hat{\theta} - \theta)}{s_\psi \sqrt{a^T (X^T X)^{-1} a}} \approx t_{n-p}
\]

for \( s_\psi \overset{def}{=} \sqrt{v_\psi} \).

– Note that the only change here (and in F-tests, etc.) is that the LS-based \( S \) is replaced by \( s_\psi \).

1.2.4. Random regressors; Breakdown

• To now we have implicitly treated the regressors \( x_i \) as fixed, i.e. non-random. In practice they are often observed values of random variables. In the model \( Y_i = x_i^T \theta + \varepsilon_i \) we now assume that \( x_i \) and \( \varepsilon_i \) are independently (but perhaps not Normally) distributed.

• The case of random regressors raises new robustness issues, since now there may be highly influential values of the \( x_i \). Example: the ‘mineral’
data set in [MMY] gives values of zinc vs. copper in 53 rock samples from Western Australia. One observation (\#15) is a clear outlier. Both the LS line and a robust fit using Huber’s $\psi_{1.5}$ are strongly influenced by this point. Also shown is the LS fit after removing this point.
• Breakdown Point (BP). Roughly speaking, this is the largest fraction of data values which can be corrupted (made arbitrarily bad) with the estimates remaining bounded. Formally, for a data set \( Z = \{x_i, y_i\}_{i=1}^n \), let \( Z_m \) denote any data set with at least \( n - m \) elements in common with \( Z \) (so at most \( m \) can be corrupted). Define

\[
m^* = \max \left\{ m \mid \hat{\theta}(Z_m) \text{ is bounded for all } Z_m \right\}.
\]

Then \( \varepsilon^* = m^*/n \) is called the ‘finite-sample breakdown point’, and \( \lim_{n \to \infty} \varepsilon^* \) is the breakdown point.

• In the regression model \( Y_i = x_i^T \theta + \varepsilon_i \) with random \( x_i \), an M-estimate of regression with monotone (i.e. non-decreasing) \( \psi \) has \( BP = 0 \).

• In view of this last result, it is important to find robust estimates of regression with positive BPs (when contaminated regressors are a possibility). In fact \( BP = .5 \) is attainable. We will look
at two possibilities – (i) expand the class of M-estimates to allow for the effect of influential x’s to be bounded, or (ii) drop the requirement of a monotone $\psi$. The first of these leads to ‘Bounded Influence’ or ‘Generalized’ M-estimation, the second to ‘MM-estimation’.

1.2.5. Generalized M-estimation

• High breakdown estimators. An early attempt at finding a regression estimate with high BP led to the ‘Least Median of Squares’ estimate. This is defined by

$$\text{med} \left\{ (y_i - x_i^T \hat{\theta})^2 \right\} = \text{min}.$$  

The LMS estimate is in general very difficult to compute, does not have a limiting Normal distribution, and in fact converges to a non-Normal distribution at the rate $n^{-1/3}$, i.e. more slowly than the usual $n^{-1/2}$. But $BP = 1/2$. 
A more recent proposal is ‘Least Trimmed Squares’. The LTS regression method minimizes the sum of the \( h \) smallest squared residuals, where \( h \) must be at least half the number of observations and is typically taken to be slightly greater than \( n/2 \). Formally, if the absolute values of the residuals are ordered:

\[
|r|_1(1) \leq \cdots \leq |r|_1(n),
\]

then

\[
\sum_{i=1}^{h} |r|_i^2 = \min.
\]

Again difficult to compute, but it converges at the standard rate of \( n^{-1/2} \) and has a BP of .5. A drawback is that it is very inefficient if the errors are in fact Normal.

A proposal to modify the definition of an M-estimate, so as to bound the influence of outlying x-values,
resulted in ‘Generalized M-estimation’. A GM-estimate is a solution to
\[
\frac{1}{n} \sum \eta \left( x_i, \frac{y_i - x_i^T \hat{\theta}}{\hat{\sigma}} \right) x_i = 0,
\]
where
\[
\eta \left( x_i, \frac{r_i(\hat{\theta})}{\hat{\sigma}} \right) = w(x_i) \psi \left( \frac{r_i(\hat{\theta})}{\hat{\sigma}} \right)
\]
(there are other variations of this in the literature). The weights \( w(x_i) \) are to be chosen for robustness against outlying \( x \)-values. As with (ordinary) M-estimates, scale is estimated by solving an auxiliary equation. A GM-estimate can be computed just as an M-estimate was, by IRLS. Alternatively, use Newton-Raphson: define
\[
F(\hat{\theta}) = \frac{1}{n} \sum \eta \left( x_i, \frac{y_i - x_i^T \hat{\theta}}{\hat{\sigma}} \right) x_i
\]
and solve \( F(\hat{\theta}) = 0 \) through the iteration scheme
\[
\theta_{(k+1)} = \theta_{(k)} - \left[ \hat{F}(\theta_{(k)}) \right]^{-1} F(\theta_{(k)})
\]
with

\[ \hat{F}(\theta_{(k)}) = \frac{-1}{n\hat{\sigma}} \sum \eta' \left( x_i, \frac{y_i - x_i^T \theta_{(k)}}{\hat{\sigma}} \right) x_i x_i^T \]

(note \( \eta' (x, r) = (d/dr) \eta (x, r) = w(x) \psi' (r) \)).

- The asymptotic properties of an estimate obtained in this way are not radically different than those of an ordinary M-estimate. The estimate is asymptotically normal:

\[ \sqrt{n} \left( \hat{\theta} - \theta \right) \xrightarrow{L} N \left( 0, M^{-1}QM^{-1} \right) \tag{1.2.3} \]

with

\[ M = \frac{1}{\sigma} E \left[ \eta' \left( x, \frac{\varepsilon}{\sigma} \right) xx^T \right] \]
\[ = \frac{1}{\sigma} E \left[ \psi' \left( \frac{\varepsilon}{\sigma} \right) \right] E \left[ w(x) xx^T \right], \]
\[ Q = E \left[ \eta^2 \left( x, \frac{\varepsilon}{\sigma} \right) xx^T \right] \]
\[ = E \left[ \psi^2 \left( \frac{\varepsilon}{\sigma} \right) \right] E \left[ w^2 (x) xx^T \right]. \]

These are estimated by replacing the expectations by averages over the sample.
1.2.6. One-step GM-estimation

- A drawback of GM-estimation is that the BP, while positive, is only about $1/p$. A way out of the problem is to compute a ‘one-step’ GM-estimate:

1. Take a high breakdown initial estimate of $\theta$, such as the LTS estimate, and a corresponding scale estimate $\hat{\sigma} = \sqrt{\frac{1}{h} \sum_{i=1}^{h} |r_i|^2}$. (This is multiplied by a correction factor – see \texttt{ltsReg(robustbase)} or the Pison, Van Aelst & Willems paper on the STAT578 web site for details.)

2. Compute as well highly robust weights $w(x_i)$ (discussed later).

3. Perform just one iteration of Newton-Raphson (\textbf{not} IRLS – this results in the wrong asymptotic properties when only one iteration is performed. Update $\hat{\sigma}$ to $\hat{\sigma} = MAD$.}
• It can be shown – see the Simpson, Ruppert & Carroll (‘SR&C’) paper on the STAT578 web site for details – that \( \hat{\theta} \) computed in this way inherits the high BP of the initial estimate, while gaining the high efficiency of the M-estimate. In particular (1.2.3) continues to hold.

• One need not stop at one step; one can use \( \theta_{(1)} \) in place of \( \hat{\theta}_{LTS} \) and do one more iteration, obtaining a two-step GM estimate \( \theta_{(2)} \), etc. In fact SR&C recommend a three-step. BUT the number of iterations \( k \) must be decided on in advance [why?].

• How can the robust weights \( w(x) \) be computed? These should decrease as \( x \) moves away from the rest of the sample. An obvious possibility is something like \( w(x_i) = 1/h_{ii} \), but these are very non-robust – outlying \( x_i \) can determine the measure (the ‘masking’ effect). This is most clear in
straight line regression, where

\[
h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_j - \bar{x})^2}.
\]

- To get more robust weights, we first look for robust estimates \( \mathbf{t} \) and \( \mathbf{V} \) of the location and scatter of the \( x_i \). These are obtained from the Minimum Covariance Determinant (MCD) method [what is this?] resulting in

\[
\mathbf{V} = \frac{1}{h} \sum_i \left( \mathbf{x}(i) - \mathbf{t} \right) \left( \mathbf{x}(i) - \mathbf{t} \right)^T
\]

(and \( \mathbf{t} = \bar{x} \), the average of these \( h > n/2 \) points). Then a reweighting step is carried out to improve the efficiency — see covMcd(robustbase) or the Pison, Van Aelst & Willems paper for details. This results in robust estimates \( \hat{\mu} \) (a reweighted average of these \( h \) points) and \( \hat{\Sigma} \) (a reweighted covariance matrix). Finally, weights are computed:

\[
w(x_i) = \min \left( 1, \frac{\chi_{p-1}^2 (.95)}{(x_i - \hat{\mu})^T \hat{\Sigma}^{-1} (x_i - \hat{\mu})} \right)^{1/2}.
\]
• The original proposal of SR&C was to use Minimum Volume Ellipsoid (MVE) weights. [*What are these?*] This suffers from the same problems as the LMS estimate, however.

• An R function to compute K-step GM-estimates is on the STAT578 web site. It is applied here to the ‘mineral’ data set. See plots below.
• The remaining lines, very close to the LS line (after removing point 15) are the GM line using a (redescending) ‘bisquare’ ψ-function

$$
\psi_{bi} (r; c) = r \left( 1 - \left( \frac{r}{c} \right)^2 \right) \cdot I(|r| \leq c)
$$

with \( c = 4.5 \), and the MM line, to be considered next.

1.2.7. MM-estimation

• MM estimation (so named because it uses two M-estimates) is the high breakdown regression method currently in vogue. Like GM-estimation it starts with a high breakdown initial estimate. But rather than LTS or LMS another method – ‘S estimation’ – is used and is discussed below.
The method depends on two bounded ‘\( \rho \)-functions’ \( \rho_0 \) and \( \rho_1 \). Such functions must be nondecreasing in \( |r| \), with \( \rho(0) = 0 \), \( \rho(\infty) = 1 \) and \( \rho \) strictly increasing in \( |r| \) where \( \rho(r) < 1 \). [So how must \( \psi \) look?]. Recommended is the bisquare \( \rho \)-function

\[
\rho_{bi}(r; c) = \min \left\{ 1, 1 - \left( 1 - \left( \frac{r}{c} \right)^2 \right)^3 \right\},
\]

with derivative

\[
\rho'_{bi}(r; c) = \frac{6}{c^2} \psi_{bi}(r; c).
\]

The S-estimate \( \hat{\theta}_S \), and scale estimate \( \hat{\sigma}_S \), are defined as follows. For any \( \hat{\theta} \), with residuals \( r_i(\hat{\theta}) \), define a scale estimate \( \hat{\sigma} = \hat{\sigma}(\hat{\theta}) \) by

\[
\frac{1}{n} \sum \rho_0 \left( \frac{r_i(\hat{\theta})}{\hat{\sigma}} \right) = .5.
\]

(Non-robust example: \( \rho_0(r) = r^2/2 \) gives \( \hat{\sigma}^2 = \sum r_i^2/n. \)) The S-estimate of regression is the
solution to
\[
\hat{\sigma} (\hat{\theta}_S) = \min_{\hat{\theta}} \hat{\sigma} (\hat{\theta}),
\]
and then
\[
\hat{\sigma}_S = \hat{\sigma} (\hat{\theta}_S).
\]
This is the major computational challenge. Theoretical details are in the paper by Victor Yohai, on the STAT578 web site.

- Regression is then estimated by solving
\[
L (\hat{\theta}) \overset{\text{def}}{=} \frac{1}{n} \sum \rho_1 \left( \frac{r_i (\hat{\theta})}{\hat{\sigma}_S} \right) = \min,
\]
starting with \( \hat{\theta}_S \). It is required that
\[
\rho_1 \leq \rho_0, \quad (1.2.4a)
\]
\[
L (\hat{\theta}) \leq L (\hat{\theta}_S); \quad (1.2.4b)
\]
these ensure the high BP (\( \rightarrow 0.5 \)).

- The recommended \( \rho \)-functions are
\[
\rho_0 (r) = \rho_{bi} (r; c_0) \text{ and } \rho_1 (r) = \rho_{bi} (r; c_1),
\]
where:

1. $c_0 = 1.56$ so that, asymptotically for Normal errors, $\hat{\sigma}$ will correspond to the standard deviation;

2. $c_1$ must be $\geq c_0$ to satisfy (1.2.4a), and is chosen for a prescribed efficiency at the Normal, e.g. for 95% efficiency $c_1 = 4.68$; larger values give greater efficiency but allow large residuals to have a greater influence on the fit.

- Computation of $\hat{\theta}_S$ and $\hat{\sigma}_S$: all approaches rely on ‘subsampling’ schemes; these are also used in the computation of LMS, MCD, etc. For instance for LMS, consider a subsample

$$\{(x_i, y_i) \mid i \in J\},$$

where $J$ is any one of the $\binom{n}{p}$ sets of $p$ indices chosen from $\{1, 2, \cdots, n\}$. Assume that the corresponding design matrix $X_J$, with rows $\{x_i^T \mid i \in J\}$
has full rank (if not, drop this subsample and take another). Then the corresponding regression coefficients are

$$\hat{\theta}_J = \left( X_J^T X_J \right)^{-1} X_J^T y_J = X_J^{-1} y_J$$

and this estimated model fits these $p$ datapoints exactly:

$$\hat{y}_J = X_J \hat{\theta}_J = y_J.$$  

In practice we can’t consider all $\binom{n}{p}$ subsamples. Instead, a large number $N$ of them are randomly chosen, and the ‘best’, i.e. the one for which the median of squared residuals is the least, is taken as the solution.

- Despite its apparent complexity, an MM-estimate is just an ordinary M-estimate corresponding to a redescending $\psi$. The complexity comes in only through the initial estimates $\left( \hat{\sigma}_S, \hat{\theta}_S \right)$; then their high BP is inherited by virtue of (1.2.4).
1.2.8. Additional problems

1. Describe the use of ‘pseudovalues’ in robust regression, using Ordinary M-estimation. These are observations $\tilde{y}_i$ which are constructed, from the data and from estimates which have already been computed, in such a way that a final, least squares regression on the pseudo-values, together with the computer printout generated from a least squares regression, yields p-values which are asymptotically correct for the robust estimates. [W578 Lec. 17]

2. Show that, in the regression model $Y_i = x_i^T \theta + \varepsilon_i$ with random $x_i$, an M-estimate of regression with monotone (i.e. non-decreasing) $\psi$ has $BP = 0$. [MMY]

3. [RL] Consider a straight line regression model. Suppose that one of the regressors, say $x_1$, is an outlier which is so far away from the rest of the sample that it is larger than $\bar{x}$, whereas all other $x_i$ are smaller than $\bar{x}$. 
(a) Show that the L1 line necessarily passes through $(x_1, y_1)$.

(b) Use (a) to give an alternate proof (i.e. different from the one above for monotone M-estimators) that the breakdown point of the estimate is 0.
1.3. Minimax Robust Estimation

Source material: [Hb] ch. 4; [H64]

Asymptotically normal estimates: We estimate

\[ \theta = T(F) \]

(location M-parameter: \( E_F [\psi (X - T(F))] = 0 \)) by

\[ \hat{\theta}_n = T(\hat{F}_n) \]

(location M-estimate: \( \frac{1}{n} \sum_{i=1}^{n} \psi (x_i - \hat{\theta}_n) = 0 \)); typically

\[ \sqrt{n} (T(\hat{F}_n) - T(F)) \xrightarrow{L} N(0, V(F,T)) \]

where

\[ V(F,T) = E_F [IF^2 (F,T)] . \]
For an M-estimate of location,

\[ V(F, T) = \frac{E_F \left[ \psi^2(X - T(F)) \right]}{(E_F [\psi'(X - T(F))])^2}. \]

The bias in estimating a target value \( \theta_0 \) (e.g. \( \theta_0 = T(G) \) for a central distribution \( G \)) is \( T(F) - \theta_0 \). Assume ‘Fisher consistency’:

\[ T(F) = \theta_0 \] for all \( F \) being considered.

e.g. all distributions symmetric around \( \theta_0 \) (= 0, say) and \( \psi \) an odd function. Then for an M-estimate

\[ \sqrt{n} (\hat{\theta}_n - \theta_0) \xrightarrow{L} N \left( 0, \frac{E_F [\psi^2(X)]}{(E_F [\psi'(X)])^2} \right). \]
The estimate $T_0(F_n)$ is asymptotically minimax if, as $F$ varies over a class $\mathbb{F}$ of distributions, we have

$$\sup_{F \in \mathbb{F}} V(F, T_0) \leq \sup_{F \in \mathbb{F}} V(F, T)$$

for any other $T$. (Depends on $\mathbb{F}$ and on the class of $T$'s being considered.)

The easy way (if it works!) to check this is to verify the saddlepoint property: We look for a pair $(F_0, T_0)$ such that, for all other $(F, T)$,

$$V(F, T_0) \leq V(F_0, T_0) \leq V(F_0, T).$$

[Show that this implies that $T_0$ is minimax.]
First step: $T_0$ must minimize the variance at $F_0$. Consider an M-estimate, with

$$V(F, T) = \frac{E_F [\psi^2(X)]}{(E_F [\psi'(X)])^2}.$$ 

Assume $F$ has a differentiable density $f$ and put $\psi_F(x) = -f'(x)/f(x)$. Note

$$E_F [\psi_F(X)] = 0, \quad VAR_F [\psi_F(X)] = I(F).$$

The denominator in $V(F, T)$ is (integration by parts)

$$\int \psi'(x) f(x)dx = COV_F [\psi(X), \psi_F(X)]$$

so

$$V(F, T) = \frac{VAR_F [\psi(X)]}{(COV_F [\psi(X), \psi_F(X)])^2} \geq \frac{1}{I(F)},$$

with equality iff $\psi = \psi_F$.

Summary: Among M-estimates, for any fixed $F_0$ the best $T_0$ is the MLE based on $F_0$. A particular $F_0$, and the corresponding MLE $T_0$, possess the saddlepoint property if

$$\sup_{F \in \mathcal{F}} V(F, T_0) = V(F_0, T_0) = 1/I(F_0).$$
Similarly for an L-estimate: $X_{1:n} \leq ... \leq X_{n:n}$ the order statistics, $m(\cdot)$ a weights-generating function:

$$T_L = \frac{1}{n} \sum m \left( \frac{i}{n+1} \right) X_{i:n}.$$  

From (1.1.2), the IF is

$$IF(x; F) = \int_{-\infty}^{x} m(F(y)) dy - c,$$

where $c$ is determined by $E_F[IF(X; F)] = 0$, and so the choice [how?]

$$m_F(t) = \frac{\psi'_F \left( F^{-1}(t) \right)}{I(F)}$$

yields $V(F, T_L) = 1/I(F) = \min$. Thus the saddle-point property holds if we can find $F_0$ such that the L-estimate $T_0$, using weights $m_{F_0}$, has

$$\sup_{F \in F} V(F, T_0) = V(F_0, T_0) = 1/I(F_0).$$

R-estimates can be considered analogously.
Theorem 1: For an M-estimate, with variance

\[ V(\psi, F) = \frac{E_F[\psi^2(X)]}{(E_F[\psi'(X)])^2} \]

the saddlepoint property holds. It is attained by finding \( F_0 \) which minimizes Fisher information in \( F \), and taking \( \psi_0 = -f'_0/f_0 \).

Proof: We use the following repeatedly. If \( a(t), b(t) \) are linear functions of \( t \in [0, 1] \), with \( b(t) > 0 \), then \( r(t) = \frac{a^2(t)}{b(t)} \) is convex.

[How is this proven?]

Thus

\[ r(t) \text{ is minimized at } 0 \iff r'(0) \geq 0. \]
When is \( V(\psi_0, F) \) maximized at \( F_0 \)? With

\[
F_t \overset{\text{def}}{=} (1 - t) F_0 + t F_1
\]

(\( \in \mathbb{F} \) if \( \mathbb{F} \) is a convex class; assumed) we require that

\[
r(t) = \frac{1}{V(\psi_0, F_t)} = \frac{\left( E_{F_t} \left[ \psi'_0(X) \right] \right)^2}{E_{F_t} \left[ \psi^2_0(X) \right]}
\]

be minimized at 0 for every \( F_1 \in \mathbb{F} \); calculating \( r'(0) \) leads to

\[
\int (2\psi'_0 - \psi^2_0)(x)(f_1 - f_0)(x)dx \geq 0 \ \forall f_1.
\]

(1.3.1)

(Can restrict to \( F_1 \) with \( I(F_1) < \infty \); Huber showed that these have differentiable densities.)
Similarly

\[ I(F_t) = \int \frac{[f'_t(x)]^2}{f_t(x)} \, dx \]

is convex; differentiating w.r.t. \( t \) shows that it is minimized at \( t = 0 \) iff (1.3.1) holds. Thus

saddlepoint \iff \[ V\left(\psi_0 = -\frac{f'_0}{f_0}, F\right) \] is maximized at \( F_0 \)

\iff (1.3.1) holds

\iff \[ I(F') \] is minimized at \( F_0 \). \( \square \)

Summary: We are to find the least informative \( F_0 \in \mathcal{F} \); the MLE corresponding to it is minimax.

Huber (1964) found \( F_0 \) in the gross errors neighbourhood

\[ \mathcal{G}_\varepsilon = \{ f = (1 - \varepsilon)g + \varepsilon h, \, h \text{ arbitrary but symmetric} \} \]

for \( g \) symmetric and ‘strongly unimodal’;

we study this case next.
Minimizing $I(F')$ in $\mathcal{G}_\varepsilon$

Recall the necessary and sufficient condition (1.3.1):

$$\int (2\psi_0' - \psi_0^2)(x)(f_1 - f_0)(x)dx \geq 0 \forall f_1.$$ 

With $f_i = (1 - \varepsilon)g + \varepsilon h_i$ and $J(\psi) \overset{\text{def}}{=} 2\psi' - \psi^2$ we require

$$\int J(\psi_0)(x)(h_1 - h_0)(x)dx \geq 0 \forall h_1.$$ 

We must place the mass of $h_0$ in such a way that $\int J(\psi_0)hdx$ is minimized by $h_0$. Define $\xi = -g'/g$. One conjectures that the mass should then be placed where $J(\xi)$ is small, and on this region the resulting $\psi_0$ should satisfy $J(\psi_0) \equiv -\lambda^2$ for some constant $\lambda$.

This yields the following set of sufficient conditions; Collins and Wiens (1985) showed them to be necessary as well.
Theorem 2: In order that \( f_0 = (1 - \varepsilon) g + \varepsilon h_0 \) minimize \( I(F) \) in \( G_\varepsilon \) it is sufficient that \( f_0 \) and \( \psi_0 = -f'_0/f_0 \) be continuous, that \( \int f_0 \, dx = 1 \) and that with \( S \overset{\text{def}}{=} \{ x \mid h_0(x) > 0 \} \) we have

(i) \( J(\psi_0) \equiv -\lambda^2, \ x \in S \)

(ii) \( J(\xi) \geq -\lambda^2, \ x \notin S. \)

Then

\[
f_0(x) = \begin{cases} 
(1 - \varepsilon) g(x), & x \notin S, \\
(1 - \varepsilon) g(x) + \varepsilon h_0(x), & x \in S;
\end{cases}
\]

\[
\psi_0(x) = \begin{cases} 
\xi(x), & x \notin S, \\
\text{sol’n to } J(\psi_0) \equiv -\lambda^2 & x \in S.
\end{cases}
\]

[Prove that (1.3.1) holds, if (i) and (ii) do.]
Example 1: \( g = \phi \).

Here

\[
\xi(x) = -\frac{\phi'}{\phi} = -\frac{d}{dx} \log \phi(x) = \frac{d}{dx} \frac{x^2}{2} = x,
\]

so \( J(\xi)(x) = 2 - x^2 \) is decreasing on \((0, \infty)\) and the support \( S \) of \( h_0 \) in \((0, \infty)\) is an interval \([c, \infty)\). The solution (there are 3) to \( J(\psi_0) \equiv -\lambda^2 \) which works for us is \( \psi_0 \equiv \lambda \), so that here

\[
f_0(x) = e^{-\int_a^x \psi_0(t)dt} = e^{-\lambda(x-a)} \text{ for some } a.
\]

Choosing \((a, c)\) for continuity of \((f_0, \psi_0)\) gives

\[
f_0(x) = \begin{cases} 
(1 - \varepsilon) \phi(x), & |x| \leq \lambda, \\
(1 - \varepsilon) \phi(\lambda) e^{-\lambda(|x| - \lambda)}, & |x| \geq \lambda;
\end{cases}
\]

\[
\psi_0(x) = \begin{cases} 
x, & |x| \leq \lambda, \\
\text{sign}(x) \cdot \lambda & |x| \geq \lambda.
\end{cases}
\]

Now \( \lambda = \lambda(\varepsilon) \) is chosen so that \( \int f_0(x)dx = 1 \).
Check that \( h_0(x) \geq 0 \). Note that \( \xi(t) \geq \psi_0(t) \) on \( \{ t \geq \lambda \} \), integrating this relationship over \([\lambda, x]\) gives

\[
-\log \frac{(1 - \varepsilon) \phi(x)}{(1 - \varepsilon) \phi(\lambda)} \geq -\log \frac{f_0(x)}{f_0(\lambda)}.
\]

Since

\[
(1 - \varepsilon) \phi(\lambda) = f_0(\lambda)
\]

we have

\[
f_0(x) \geq (1 - \varepsilon) \phi(x) \text{ for } x \geq \lambda,
\]

i.e. \( h_0(x) \geq 0 \) there. See plots (Appendix page A1).
Example 2: \( g = \text{Cauchy} = \left( \pi \left( 1 + x^2 \right) \right)^{-1} \).

Here

\[
\xi(x) = \frac{2x}{1 + x^2}, \quad J(\xi)(x) = \frac{4 \left( 1 - x^2 \right)}{(1 + x^2)^2}.
\]

See plots (Appendix page A2): \( \xi \) is non-monotone (\( g \) not strongly unimodal) and the region where \( J(\xi) \) is small is an interval \([a, b]\). On this interval \( J(\psi_0) \equiv -\lambda^2 \) and we require a decreasing solution. This gives

\[
f_0(x) = \begin{cases} 
(1 - \varepsilon) g(x), & |x| \notin [a, b], \\
(1 - \varepsilon) s \cosh^2 \left( -\frac{\lambda}{2} (x - \omega) \right), & |x| \in [a, b];
\end{cases}
\]

\[
\psi_0(x) = \begin{cases} 
\xi(x), & |x| \notin [a, b], \\
\lambda \tanh \left( -\frac{\lambda}{2} (x - \omega) \right), & |x| \in [a, b].
\end{cases}
\]

Now \((s, a, b, \lambda, \omega)\) are chosen so that \( f_0 \) and \( \psi_0 \) are continuous at \( a \) and \( b \), and \( \int f_0(x) dx = 1 \).
Example 3: Kolmogorov and Lévy neighbourhoods.

Huber (1964) also found $F_0$ in the Kolmogorov neighbourhood

$$K_\varepsilon = \{ F \mid |F(x) - G(x)| \leq \varepsilon \text{ for all } x \}$$

for $G = \Phi$ and $\varepsilon \leq .0303$.

Sacks and Ylvisaker (1972) for $\varepsilon > .0303$; Wiens (1986) for more general $G$ (e.g. $G = \text{Cauchy}$). See Appendix page 3.

Collins and Wiens (1985) found $F_0$ in $G_\varepsilon$ without the assumption of strong unimodality (e.g. $G = \text{Cauchy}$) and (1989) in Lévy neighbourhoods of strongly unimodals. See Appendix pages A3, A4.
Minimax L-estimation

$1/V(T, F')$ is no longer convex, so

$$r'(0) \geq 0 \text{ for all } F_1$$

becomes a necessary but not sufficient condition.

One still aims to verify the saddlepoint condition

$$V(T_0, F') \leq 1/I(F_0)$$

where $T_0$ is the efficient L-estimator for $F_0$ and $I(F_0) = \min.$
$G_\varepsilon$: \( g \) strongly unimodal; saddlepoint property holds (Jaeckel 1971)

\( g \) not strongly unimodal \( \rightarrow \) quite generally, the saddlepoint property fails (Wiens 1990).

$K_\varepsilon$: \( G = \Phi \): Fails for \( \varepsilon > .07 \) (Sacks & Ylvisaker 1972)

More generally, if \( G \) is strongly unimodal the saddlepoint property fails in \( K_\varepsilon \) for all \( \varepsilon > 0 \); also in all Lévy neighbourhoods (Collins & Wiens 1989).

In fact if \( G \) is merely strictly increasing it fails for all \( \varepsilon > 0 \). Wiens (1990) proved, for this class \( K_\varepsilon \), that if

\[
\mathbb{F}_0 = \left\{ F \in K_\varepsilon \mid \int \left( 2\psi'_0 - \psi^2_0 \right) (x) d(F_1 - F_0)(x) = 0 \right\}
\]

then for \( F \in \mathbb{F}_0 \) we have that \( V(T_0, F) \) is non-constant and is minimized at \( F_0 \).
In $\mathbb{K}_\varepsilon$ with $G = \Phi$ the saddlepoint property fails, but is $T_0$ minimax nonetheless?

No – if $T_1$ is an $\alpha$-trimmed mean, with $\alpha$ chosen to minimize the maximum variance, then

$$\inf_{T} \sup_{\mathbb{K}_\varepsilon} V(T, F) \leq \sup_{\mathbb{K}_\varepsilon} V(T_1, F) < \sup_{\mathbb{K}_\varepsilon} V(T_0, F).$$


$T_1$ is much simpler than $T_0$. Although it is also not a saddlepoint solution, it might still be minimax – the first inequality might be an equality. (Open)
1.3.1. Additional problems


2. Show that the sample mean is the minimax variance M-estimate in the class $F_v$ of symmetric distributions with $\text{var}_F [X] \leq v^2$, for a fixed constant $v$.

3. Show that the sample median minimizes the maximum bias, among all translation invariant estimates of location, in a gross errors neighbourhood of a symmetric unimodal distribution.
Part 2

Robustness of Design
2.1. Classical design theory

Source material: [CR, Pb]

- Notation: An experimenter anticipates making observations at various values of certain predictor variables (vectors \( x \)) and then observing the corresponding values of a dependent variable \( Y \). The predictors affect \( E [Y] \) through a vector \( f (x) \) of regressors. For instance in bivariate regression with an intercept and two interacting predictors,

\[
x = (x_1, x_2)', \quad f (x) = (1, x_1, x_2, x_1x_2)'.
\]

We must distinguish between \( x \) and \( f \) because we are interested in choosing \( x \), not \( f \). We also specify a ‘design space’ \( \chi \) from which the \( x \)'s are to be chosen.

- After making the observations the experimenter intends to analyze the data \( \{Y_i, x_i\}_{i=1}^n \) by fitting a regression model. For linear regression he fits

\[
Y = f' (x) \theta + \varepsilon.
\]
• The classical assumptions are that $E[Y] = f'(x) \theta$
  for some $\theta$, and that the errors are i.i.d. Normal-
  ity plays only a limited role here – the measures
  of performance typically depend only on the first
two moments of the estimates.

• Example: Water purity ($y$) measured by experi-
  menter, as a function of input variable $x = chlo-
  rine; set by the experimenter and coded to $\chi =
  [-1, 1]$.

• For straight line regression

  $$y_i = \theta_0 + \theta_1 x_i + \epsilon_i, -1 \leq x_i \leq 1,$$
  and in terms of $S_X^2 = \sum (x - \bar{x})^2 / n$, one finds
  that

  $$nVAR[\hat{\theta}_1] = \frac{\sigma^2_\epsilon}{S_X^2},$$

  $$nVAR[\hat{\theta}_0] = \sigma^2_\epsilon \left(1 + \frac{\bar{x}^2}{S_X^2}\right).$$
• Both are minimized by putting 1/2 of the \( x \)'s at each of \( \pm 1 \) (so \( \bar{x}^2 = 0 = \text{min} \) and \( S_X^2 = \text{max} \)).

• Classically optimal design measure \( \xi(x) = \text{fraction of designs points placed at } 'x' \) has

\[
\xi(-1) = \xi(1) = 1/2.
\]

• In general, we aim to make \( \text{cov} \left[ \hat{\theta} \right] = \sigma_\varepsilon^2 (X'X)^{-1} \)
‘small’, i.e. choose the \( x \)'s to minimize its trace, determinant, maximum eigenvalue, etc.

• For quadratic regression

\[
y_i = \theta_0 + \theta_1 x_i + \theta_2 x_i^2 + \varepsilon_i, \quad -1 \leq x_i \leq 1,
\]

the ‘D-optimal’ design measure minimizing the determinant has \( \xi(-1) = \xi(1) = \xi(0) = 1/3 \), i.e. 1/3 of the observations on \( Y \) are made at each of 0, \( \pm 1 \).
• In both of these cases the estimates are unbiased
\[ (E[\hat{\theta}] = \theta) \] if the fitted model is correct, and
so interest naturally focusses on minimizing the
variance.

• There is a huge literature on this design problem
for linear regression, and its variations, all centred
on minimizing some function of \((X'X)^{-1}\). Much
of it concerns polynomial regression – here ex-
plicit results can be obtained, and the associated
mathematics is very interesting. A good survey is
[Pb].

2.1.1. Some optimality theory

• Let \( \xi \) be the ‘design measure’. This is sometimes
viewed as a pmf: if the design places \( n_i \) out of a
total of \( n \) observations at the point \( x_i \in \chi \) then
\( \xi (x_i) = \frac{n_i}{n} \). Sometimes as a measure on \( \chi \): for
instance

\[
X'X = \sum n_i f(x_i) f'(x_i) \\
= n \sum \xi(x_i) f(x_i) f'(x_i) \\
= n \int_X f(x) f'(x) \xi(dx).
\]

This last form makes it convenient to think of approximating a design by something with a measure \(\xi\) whose atoms are not necessarily multiples of \(n^{-1}\) — perhaps even a continuous measure.

- The covariance of the lse is \(\sigma^2/n\) times the inverse of

\[
\frac{1}{n}X'X = \int_X f(x) f'(x) \xi(dx) \overset{def}{=} M(\xi).
\]

Some common optimality principles:

- D-optimality — minimize \(\det(M^{-1}(\xi))\). \((\propto\) volume of a confidence ellipsoid on regression parameters\)

- A-optimality — minimize \(tr(M^{-1}(\xi))\). \((\text{average variance of regression parameters})\)
- E-optimality – minimize $ch_{\max}M^{-1}(\xi)$. ($\propto$ maximum variance of a normalized linear combination $c'\hat{\theta}$)

- G-optimality – minimize

$$d(\xi) = \sup_{x \in \chi} f'(x) M^{-1}(\xi) f(x).$$

(maximum variance of a prediction $\hat{Y}(x) = f'(x) \hat{\theta}$)

• **Theorem** (General Equivalence Theorem): In a $p$-parameter regression model, a design $\xi_0$ is D-optimal iff it is also G-optimal, and then $d(\xi_0) = p$.

**Proof:** For $t \in [0, 1]$ and any design $\xi_1$ define

$$\xi_t = (1 - t) \xi_0 + t \xi_1.$$

Then $\xi_0$ is D-optimal iff the function

$$\phi(t) = -\log |M(\xi_t)|$$

is minimized at $t = 0$, for every $\xi_1$. We need some preliminary results.
Lemma 1: For p.s.d. matrices $A$ and $B$ and for $t \in [0, 1]$ we have that

$$\begin{align*}
- \log |(1 - t) A + tB| & \leq (1 - t) (-\log |A|) + t (-\log |B|).
\end{align*}$$

Proof: If either of $A$ or $B$ is singular there is nothing to prove, so assume both are p.d.; assume as well that $0 < t < 1$. Define $C = A^{-1/2}BA^{-1/2}$, so that $B = A^{1/2}CA^{1/2}$; after a calculation (2.1.1) becomes

$$\begin{align*}
- \log |(1 - t) I + tC| & \leq -t \log |C|.
\end{align*}$$

(2.1.2)

In terms of the, necessarily positive, eigenvalues $\{\lambda_i\}_{i=1}^p$ of $C$,

$$\begin{align*}
|(1 - t) I + tC| &= \prod (1 - t + t\lambda_i),
|C| &= \prod \lambda_i,
\end{align*}$$

and so (2.1.2) is

$$\begin{align*}
\sum_{i} \left[ - \log ((1 - t) 1 + t\lambda_i) \right] & \leq \sum_{i} \left[ (1 - t) (-\log 1) + t (-\log \lambda_i) \right],
\end{align*}$$
which holds because the function \( f(x) = -\log x \) is convex.

- **Lemma 2**: For a p.d. matrix \( A \) the matrix of partial derivatives of \( \log |A| \) with respect to the elements of \( A \) is \( A^{-1} \).

**Proof**: The \((i, j)\)th element of the matrix of partial derivatives is

\[
\lim_{\varepsilon \to 0} \frac{\log |A + \varepsilon e_ie'_j| - \log |A|}{\varepsilon} = \lim_{\varepsilon \to 0} \frac{\log |I + \varepsilon A^{-1}e_ie'_j|}{\varepsilon} = \lim_{\varepsilon \to 0} \frac{\log \left(1 + \varepsilon e'_jA^{-1}e_i\right)}{\varepsilon} = \frac{d}{d\varepsilon} \log \left(1 + \varepsilon e'_jA^{-1}e_i\right)_{\varepsilon=0} = e'_jA^{-1}e_i \overset{def}{=} a^{ji} = \left[A^{-1}\right]_{ji} = \left[A^{-1}'\right]_{ij}.
\]
By Lemma 2, for any \( M(\xi_0) = A \) and \( M(\xi_1) = B \), the function \( \phi(\tau) = -\log |M(\xi)| \) is convex, and so is minimized at \( \tau = 0 \).

We always have (how?)

\[
p = \int f(x) M^{-1}(\xi_0) f(x) \xi_0 (dx) \leq \bar{d}(\xi);
\]

\[
0 \leq \phi'(0) = -\int f(x) M^{-1}(\xi_0) f(x) (\xi_1 - \xi_0) (dx).
\]

By Lemma 1, for any \( M(\xi_0) = A \) and \( M(\xi_1) = B \), the function \( \phi(\tau) = -\log |M(\xi)| \) is convex, and so is minimized at \( \tau = 0 \).

\[
\begin{align*}
\phi(\tau) &= -\int f(x) M^{-1}(\xi_0) f(x) (\xi_1 - \xi_0) (dx) \\
&= -\int f(x) M^{-1}(\xi_0) f(x) (\xi_1 - \xi_0) (dx) \\
&= -\int f(x) M^{-1}(\xi_0) f(x) (\xi_1 - \xi_0) (dx)
\end{align*}
\]

\[
\begin{align*}
\phi(\tau) &= -\int f(x) M^{-1}(\xi_0) f(x) (\xi_1 - \xi_0) (dx) \\
&= -\int f(x) M^{-1}(\xi_0) f(x) (\xi_1 - \xi_0) (dx) \\
&= -\int f(x) M^{-1}(\xi_0) f(x) (\xi_1 - \xi_0) (dx)
\end{align*}
\]

\[
\begin{align*}
\phi(\tau) &= -\int f(x) M^{-1}(\xi_0) f(x) (\xi_1 - \xi_0) (dx) \\
&= -\int f(x) M^{-1}(\xi_0) f(x) (\xi_1 - \xi_0) (dx) \\
&= -\int f(x) M^{-1}(\xi_0) f(x) (\xi_1 - \xi_0) (dx)
\end{align*}
\]

\[
\begin{align*}
\phi(\tau) &= -\int f(x) M^{-1}(\xi_0) f(x) (\xi_1 - \xi_0) (dx) \\
&= -\int f(x) M^{-1}(\xi_0) f(x) (\xi_1 - \xi_0) (dx) \\
&= -\int f(x) M^{-1}(\xi_0) f(x) (\xi_1 - \xi_0) (dx)
\end{align*}
\]
any $\xi_1$,

$$\int_{\chi} f'(x) M^{-1}(\xi_0) f(x) \xi_1 (dx) \leq p. \tag{2.1.4}$$

If (2.1.4) holds then $f'(x) M^{-1}(\xi_0) f(x) \leq p$ for all $x \in \chi$, so that $\bar{d}(\xi_0) \leq p$; thus a D-optimal design $\xi_0$ has

$$\bar{d}(\xi_0) = p \leq \bar{d}(\xi_1),$$

for any other design $\xi_1$, and so is G-optimal as well.

Conversely, suppose that $\xi_0$ is G-optimal. Then $p \leq \bar{d}(\xi_0) \leq \bar{d}(\xi_1)$ for any other design $\xi_1$; in particular, with $\xi_1 = \xi_D$ (= the D-optimal design, which always exists — see Pb §7.13) we obtain $\bar{d}(\xi_0) = p$. Thus for any $\xi_1$,

$$\int_{\chi} f'(x) M^{-1}(\xi_0) f(x) \xi_1 (dx) \leq \bar{d}(\xi_0) = p,$$

so that (2.1.4) holds and $\xi_0$ is D-optimal. \qed
• **A sufficient condition**: This derivation brings out that a design \( \xi_0 \) is D- or G-optimal if it places all of its mass where \( f'(x) M^{-1}(\xi_0) f(x) = p \), and at all other points we have \( f'(x) M^{-1}(\xi_0) f(x) < p \). (Then \( \bar{d}(\xi_0) = p \) and so (2.1.4) holds.) This is used to check for optimality.

• **Example**: For quadratic regression on \([-1, 1]\) the design with mass \( 1/3 \) at each of 0, \( \pm 1 \) has

\[
\{f(x_i)\} = \left\{ \begin{pmatrix} 1 \\ -1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \end{pmatrix} \right\},
\]

\[
M(\xi) = \frac{1}{3} \begin{pmatrix} 3 & 0 & 2 \\ 0 & 2 & 0 \\ 2 & 0 & 2 \end{pmatrix},
\]

\[
M^{-1}(\xi) = \frac{3}{2} \begin{pmatrix} 2 & 0 & -2 \\ 0 & 1 & 0 \\ -2 & 0 & 3 \end{pmatrix},
\]

and

\[
f'(x) M^{-1}(\xi_0) f(x) - 3 = \frac{9}{2} x^2 (x^2 - 1) \leq 0,
\]

with equality only at the points of support. Thus this design is D- and G-optimal.
• Suppose that $\xi_0$ is as described in the ‘sufficient condition’, and has exactly $p$ points of support \{x_1, ..., x_p\}. Then $\xi_0(x_i) = 1/p$ for $i = 1, ..., p$.

**Proof:** For a $p$-point design we have

$$ M(\xi) = \sum_{i=1}^{p} \xi(x_i)f(x_i)f'(x_i) = F'D_\xi F, $$

where $F_{p\times p}$ has rows $f'(x_i)$ and

$$ D_\xi = diag(\xi(x_1), ..., \xi(x_p)). $$

Thus

$$ -\log |M(\xi)| = -\log |F|^2 - \sum_{i=1}^{p} \log \xi(x_i) $$

$$ = -\log |F|^2 + p \cdot aver \{ -\log \xi(x_i) \}. $$

By Jensens’ Inequality

$$ aver \{ -\log \xi(x_i) \} \geq -\log \left( aver \{ \xi(x_i) \} \right) = \log p, $$

with equality iff $\xi(x_i) = 1/p$ for $i = 1, ..., p$. \[ \square \]

• In the situation as above, is a $p$-point design desirable?
• There are similar ‘equivalence theorems’ in other design situations.

• **Computing.** Many possibilities; one simple one is as follows. Start with any design; at the \( k^{th} \) step of the algorithm, if the current design is \( \xi_k \) then find \( x(\xi_k) \) maximizing \( f'(x) \mathbf{M}^{-1}(\xi_k) f(x) \) and put

\[
\xi_{k+1} = \frac{k}{k+1} \xi_k + \frac{1}{k+1} \delta x_{(k)}.
\]

This algorithm can be shown to converge to a D-optimal design.

2.1.2. **Nonlinear design**

• For a nonlinear response with gradient

\[
f(x; \theta) = \frac{\partial E[Y|x;\theta]}{\partial \theta},
\]

(that \( f(x; \theta) \) depends on \( \theta \) can be taken as a definition of ‘nonlinear’) and Normal errors, the
asymptotic covariance matrix of the \textit{mle} \((=\textit{lse})\) \(\hat{\theta}\) is proportional to the inverse of the Information matrix

\[
\mathbf{M}(\xi; \theta) = \int_{\chi} f(x; \theta) f'(x; \theta) \xi \, (dx).
\]

The usual alphabetic optimality criteria pertain; the dependence on the parameters is usually addressed by considering only locally optimal designs, or by designing sequentially.

- See [WH2] for additional introductory material.

2.1.3. Bayesian design

- Especially in nonlinear situations an attractive alternative is to minimize a function of the information matrix, averaged over a ‘prior’ on the parameters:

\[
\xi_0 = \arg \min_{\Theta} \int_{\Theta} \phi(\mathbf{M}(\xi; \theta)) \, p(\theta) \, d\theta.
\]

This has to be carried out numerically.
2.1.4. Space-filling designs

• Consider a computer experiment, where we plan to simulate some complex system, with given values of the tuning parameters (the ‘design’ points). (Example: initial values for a system of differential equations.) There is no random error, so our objective might be to explore the response over some quite ‘uniform’ grid of points.

• In one dimension, and for the design space \( \chi = [0, 1] \), the set of equally spaced points

\[
x_i = \frac{i - .5}{n}, \quad i = 1, \ldots, n
\]

has a minimax property: the Kolmogorov distance

\[
d_K (\xi, \xi_0) = \sup_x |\xi (x) - \xi_0 (x)|
\]

between the corresponding design measure and the uniform measure \( \xi_0 (x) = x \) is a minimum.

**Proof:** For any \( n \)-point design with ordered design
points \( t_i \), the design measure is a step function:

\[
\sup_x |F_n(x) - F(x)| = \sup_x |F_n(x) - x|
\]

\[
= \max_{1 \leq i \leq n} \max \left\{ \left| t_i - \frac{i-1}{n} \right|, \left| t_i - \frac{i}{n} \right| \right\} \geq \frac{.5}{n}.
\]

Why? The inner max exceeds its average, which exceeds \(.5/n\) (equality if \( t_i \in \left[ \frac{i-1}{n}, \frac{i}{n} \right] \), strict inequality otherwise). The lower bound is attained if \( t_i = x_i \).

- It follows that an optimal (in the \( d_K\)-sense) \( n\)-point implementation of a probability measure with cdf \( \xi_0 \) is attained by placing the observations at
  \( x_i = \xi_0^{-1} \left( \frac{i-.5}{n} \right) \), \( i = 1, \ldots, n \) (follows from the fact that \( \xi_0(X) \) is uniform if \( X \sim \xi_0 \)).

- In multidimensional space this is a more difficult problem; there are asymptotic results available.
2.2. Robustness concepts in design

- Why robustness of design?
  
  - The models for which we construct designs are generally at best approximations.
  
  - The ‘best’ design for a slightly wrong model can be much more than slightly sub-optimal.
  
  - Although we will fit the assumed, ‘ideal’ model, we should design for protection against biases arising from any of a range (‘neighbourhood’) of alternate models.
  
  - We look at mse, not merely variance. (Maximum mse, as the model varies.)
- Seminal paper - Box and Draper (1959). Design for a polynomial or multinomial response when the class of alternatives consists of all such functions of a given higher degree.

- Example: Fit a straight line in an independent variable \( x \in [-1, 1] \).

- Mean response \( E[Y(x)] = \phi_0 + \phi_1 x + \phi_2 x^2 \) for parameters \( \phi \).

- Model to be fitted to the vector \( y \) of observations is \( E[y] = X\theta \) for some \( \theta = (\theta_0, \theta_1)' \); \( \hat{\theta}_{LS} = (X'X)^{-1}X'y \). Define \( \tau_k = \sum x_i^k/n \) and assume that \( \tau_1 = \tau_3 = 0 \) (symmetry?).

- Mean vector and covariance matrix of \( \hat{\theta} \) are \( (\phi_0 + \tau_2\phi_2, \phi_1)' \) and \( (\sigma_{\varepsilon}^2/n) \text{diag}(1, \tau_2^{-1}) \) respectively, so that the predictions \( \hat{Y}(x) = (1, x)\hat{\theta} \) have \( \text{mse} \)

\[
\text{mse} \left[ \hat{Y}(x) \right] = E \left[ \left\{ \hat{Y}(x) - E[Y(x)] \right\}^2 \right] \\
= \frac{\sigma_{\varepsilon}^2}{n} \left( 1 + \frac{x^2}{\tau_2} \right) + \left( \phi_2 \left( \tau_2 - x^2 \right) \right)^2.
\]
A common measure of performance is the integrated mean squared error (imse) of the predictors, which in this instance is

\[ \text{imse} = \int_{-1}^{1} \text{mse} \left[ \hat{Y}(x) \right] dx \]  \hspace{1cm} (2.2.1)

\[ = \left\{ \frac{2\sigma_x^2}{n} \left( 1 + \frac{1}{3\tau_2} \right) \right\} + \left\{ 2\phi_2^2 \left( \left( \tau_2 - \frac{1}{3} \right)^2 + \frac{4}{45} \right) \right\}. \]

The first term in braces in (2.2.1) is the integrated variance, and is minimized by the classically optimal design with half of the observations made at each of \( x = \pm 1 \). The second – the integrated bias, dominating the first once \( n \) is sufficiently large – is minimized if \( \tau_2 = 1/3 \), which is the second moment of the (continuous) uniform distribution on \([-1, 1]\).

- If higher order alternatives are considered then minimization of this second term requires higher order moments to agree with those of the uniform.
A design approximating the uniform is the equally spaced design \( x_i = -1 + 2(i - 1)/n \), with \( \tau_2 = (1/3) + 2/(3(n - 1)) \).

Classically 'optimal' design with misspecified response

\[
E(Y) = 1 + x + 0.3(1/3 - x^2)
\]

error.std.dev = 0.3

Assumed response linear in \( x \); in fact

\[
E[Y] = 1 + x + 0.3 \left(1/3 - x^2 \right).
\]
• “... the optimal design in typical situations in which both variance and bias occur is very nearly the same as would be obtained if variance were ignored completely and the experiment designed so as to minimize the bias alone.” (Box and Draper 1959, p. 622.)

• Generalize: Fit the model $E[Y(x)] = f'(x)\theta$, for a $p$-vector $f(x)$ of regressors, each element of which is a function of $q$ independent variables $x = (x_1, ..., x_q)'$, with $x$ to be chosen from a design space $\chi$. Entertain alternatives

$$E[Y(x)] = f'(x)\theta + \psi(x),$$

(2.2.2)

for some function $\psi$. But now what does $\theta$ mean? It is not identifiable: in the alternate models one might equally well write $E[Y(x)] = f'(x)(\theta + \phi) + (\psi(x) - f'(x)\phi)$ for arbitrary $\phi$. So define the target parameter, for instance by

$$\theta = \arg\min_\eta \int_\chi (E[Y(x)] - f'(x)\eta)^2 \, dx,$$

(2.2.3)
and then define $\psi(x) = E[Y(x)] - f'(x) \theta$. This leads \((how?)\) to the orthogonality requirement

$$\int_{\chi} f(x) \psi(x) \, dx = 0. \quad (2.2.4)$$

- We identify a design, denoted $\xi$, with its design measure – a probability measure $\xi(dx)$ on $\chi$. If $n_i$ of the $n$ observations are to be made at $x_i$ we also write $\xi_i = \xi(x_i) = n_i/n$. Define

$$M_\xi = \int_{\chi} f(x) f'(x) \xi(dx), \quad (2.2.5a)$$

$$b_{\psi,\xi} = \int_{\chi} f(x) \psi(x) \xi(dx), \quad (2.2.5b)$$

and assume that $M_\xi$ is invertible. The covariance matrix of the lse $\hat{\theta}$ is $(\sigma^2_\varepsilon/n) M_\xi^{-1}$, the bias is \((how?)\) $E[\hat{\theta} - \theta] = M_\xi^{-1} b_{\psi,\xi}$, and the general
version of (2.2.1) is found to be

\[
\text{imse} = \int_{\chi} \text{mse} \left[ \hat{Y}(x) \right] \, dx
\]

\[
= \frac{\sigma^2}{n} \text{trace} \left( AM^{-1} \right)
\]

\[
+ b'_{\psi, \xi} M^{-1}_\xi AM^{-1}_\xi b_{\psi, \xi} + \int_\chi \psi^2(x) \, dx.
\]

(2.2.6)

• We must bound the influence of \( \psi(\cdot) \), to complete the definition of a class \( \Psi \) of alternatives in (2.2.2). Huber (1975; SLR) defines \( \Psi \) by (2.2.4) together with

\[
\int_{\chi} \psi^2(x) \, dx \leq \tau^2 / n,
\]

(2.2.7)

for a given constant \( \tau \). Marcus and Sacks (1976), Pesotchinsky (1982) and Li and Notz (1982) instead take

\[
|\psi(x)| \leq \phi(x),
\]

(2.2.8)

for a specified function \( \phi(\cdot) \). The resulting optimal ‘minimax’ designs (maximize over \( \psi \), then
minimize over $\xi$) depend on the form of $\phi$, but commonly the design mass is concentrated on a small number of extreme points of $\chi$. This precludes an investigation of the response in the interior of $\chi$, thus incurring criticism from ‘robust-niks’.

- The approach of Huber (1975) is itself not immune from criticism – his class $\Psi$ is so rich that any design with finite maximum loss is necessarily absolutely continuous, hence must be approximated in order to be implemented. That this is so is intuitively clear: the Lebesgue integrals in (2.2.4) and (2.2.7) may be modified on sets of Lebesgue measure zero; thus if $\xi$ places mass on any such set – as does any discrete measure – then one can choose $\psi$ to be arbitrarily large there, thus exploding the elements of $b_{\psi,\xi}$. A formal proof may be found in Heo, Schmuland and Wiens (2001). However, as in Wiens (1992), “Our attitude is that an approximation to a design which is robust against more realistic alternatives is preferable to an exact solution in a neighbourhood which is unrealistically sparse.”
• To obtain a design, robust with respect to (2.2.4) and (2.2.7), one first maximizes (2.2.6) under these constraints. In contrast to the next step – minimization over the class of designs – this can be done in complete generality. Let \( m(x) \) be the density of \( \xi \) and define \( H_\xi = M_\xi A^{-1} M_\xi \), \( K_\xi = \int_\chi f(x) f'(x) m^2(x) \, dx \) and

\[
G_\xi = K_\xi - H_\xi
\]

\[
= \int_\chi \left\{ \left[ (m(x) I_p - M_\xi A^{-1}) f(x) \right] \cdot \right\} \, dx.
\]

The matrix \( G_\xi \) is clearly positive semi-definite; assume for the moment that it is positive definite and define a function

\[
r(x) = \left( \tau / \sqrt{n} \right) G_\xi^{-1/2} \left( m(x) I_p - M_\xi A^{-1} \right) f(x).
\]

It is shown in Wiens (1992) that the class \( \Psi_0 = \{ \psi_\beta(x) = r'(x) \beta \mid \| \beta \| = 1 \} \) is a sub-class of \( \Psi \) (with equality in (2.2.7)) and is least favourable in that the supremum, over \( \Psi \), of (2.2.6) is attained by a member of \( \Psi_0 \).
• Proof of this: First verify that

\[ (1) \int r(x) r'(x) dx = \left( \frac{\tau^2}{n} \right) I_p; \]
\[ (2) \int f(x) r'(x) dx = 0; \]
\[ (3) \int r(x) \psi(x) dx = \left( \frac{\tau}{\sqrt{n}} \right) G_{\xi}^{-1/2} b_{\psi, \xi}; \]
\[ (4) \int f(x) r'(x) m(x) dx = \left( \frac{\tau}{\sqrt{n}} \right) G_{\xi}^{1/2}. \]

From (1) and (2), \( \Psi_0 \subset \Psi \). Then from (4),

\[ b_{\psi \beta, \xi} = \int f(x) \psi_\beta(x) m(x) dx = \frac{\tau}{\sqrt{n}} G_{\xi}^{1/2} \beta, \]

so that

\[ \text{imse}_{\psi_\beta} = \frac{\sigma_\varepsilon^2}{n} \text{trace} \left( AM_{\xi}^{-1} \right) \]
\[ + b'_{\psi \beta, \xi} M_{\xi}^{-1} AM_{\xi}^{-1} b_{\psi \beta, \xi} + \int \psi_{\beta}^2(x) dx \]
\[ = \frac{\sigma_\varepsilon^2}{n} \text{trace} \left( AM_{\xi}^{-1} \right) + \frac{\tau^2}{n} \beta' G_{\xi}^{1/2} H_{\xi}^{-1} G_{\xi}^{1/2} \beta + \frac{\tau^2}{n}. \]

Choose any \( \psi \); define \( \beta = G_{\xi}^{-1/2} b_{\psi, \xi} / \| G_{\xi}^{-1/2} b_{\psi, \xi} \|. \)
Then for this $\beta$,

$$\text{imse}_{|\psi_\beta} = \frac{\sigma^2}{n} \text{trace} \left( AM^{-1}_\xi \right) + \frac{\tau^2}{n} \frac{b'_{\psi,\xi} H_{\xi}^{-1} b_{\psi,\xi}}{\| G_{\xi}^{-1/2} b_{\psi,\xi} \|}$$

$$\geq \frac{\sigma^2}{n} \text{trace} \left( AM^{-1}_\xi \right) + b'_{\psi,\xi} H_{\xi}^{-1} b_{\psi,\xi} + \int_\chi \psi^2 (x) \, dx$$

$$= \text{imse}_{|\psi},$$

as long as

$$\frac{\tau^2}{n} \geq \left\| G_{\xi}^{-1/2} b_{\psi,\xi} \right\|^2.$$

(2.2.10)

To prove (2.2.10), note that by C-S and then (3),

$$\frac{\tau^2}{n} \geq \left| \int_\chi \psi (x) \psi_\beta (x) \, dx \right| = \left| \int_\chi \psi (x) r'(x) \, dx \beta \right|$$

$$= \left| \frac{\tau}{\sqrt{n}} b'_{\psi,\xi} G_{\xi}^{-1/2} \beta \right| = \frac{\tau}{\sqrt{n}} \left\| G_{\xi}^{-1/2} b_{\psi,\xi} \right\|. \square$$
• Evaluating (2.2.6) at an arbitrary $\psi_{\beta}$ (recall (2.2.9)) gives

$$imse_{\psi_{\beta}} = \frac{\sigma_{\varepsilon}^2}{n} \text{trace} \left( AM_{\xi}^{-1} \right)$$

$$+ \frac{\tau^2}{n} \beta' \left[ G_{\xi}^{1/2} H_{\xi}^{-1} G_{\xi}^{1/2} + I_p \right] \beta;$$

now maximizing over $\beta$ yields the result that $\max_{\psi} imse$ is

$$\left( \sigma_{\varepsilon}^2 + \tau^2 \right) / n \text{ times}$$

$$I_{\nu} (\xi) = (1 - \nu) \text{trace} \left( AM_{\xi}^{-1} \right) + \nu \text{ch}_{\max} \left( K_{\xi} H_{\xi}^{-1} \right),$$

(2.2.11)

where $\nu = \tau^2 / \left( \sigma_{\varepsilon}^2 + \tau^2 \right)$ and $\text{ch}_{\max}$ denotes the maximum eigenvalue.

• This final result does not require $G$ to be non-singular. If $G$ is singular then one first perturbs it to make it non-singular, and then passes to the limit – details in Heo et al. (2001).
One is now to choose $\nu \in [0, 1]$, reflecting the relative importance to the experimenter of errors due to bias rather than to variance, and minimize $\mathbb{L}_\nu(\xi)$. This step is highly dependent on the form of the model being fitted. In some simple cases it can be done analytically, using variational methods to minimize $\mathbb{L}_\nu(\xi)$ over $\xi' = m$ subject to side conditions. For straight line regression see Huber (1981). For bivariate regression (without interactions) on a spherical design space, details are in Wiens (1990) — see below for a plot of $m(x)$, minimizing $\mathbb{L}_\nu(\xi)$ when $\nu = .36$. 
• In these and the other examples in Wiens (1990, 1992), a difficulty encountered is that it is not clear which of the eigenvalues of \( K_\xi H_\xi^{-1} \) — note that these depend on the design — will turn out to be the maximum one when evaluated at the final, minimizing design. The usual approach is to attempt to guess correctly which one is to be optimized, and to then verify at the end that the guess was correct. There are cases however — quadratic regression being one of the simplest — in which this approach fails. (Open problems here; Maryam is looking at some of them.)

• Another approach is to restrict to a smaller but more tractable class of designs. Heo et al. (2001) consider the class of designs with densities of the form

\[
m_\beta(x) = \max \left( 0, \sum_j \beta_j f_j \left( x_1^2, \ldots, x_q^2 \right) \right),
\]

and minimize (2.2.11) numerically over \( \beta \) subject to the requirement that the arguments of \( m \) be
exchangeable. Design points are then chosen in such a way that the resulting design has empirical moments matching those of the optimal $m_\beta(x)$, to as high an order as possible. For fitting a full second order model, including linear, quadratic, and interaction terms, in two variables $x_1$ and $x_2$ with $n = 48$ and $\nu = 1/6$, this yields the following.

‘Restricted’ minimax design for fitting a full second order model.
The situation is somewhat simpler, with little loss of generality, in a finite design space. In practice one is very often restricted to choosing from a finite, if large, set of levels of the independent variables. Fang and Wiens (2000) take an $N$-element design space $\chi$ and discretize the definition of $\Psi$ by replacing the integrals in (2.2.3), (2.2.4) and (2.2.7) by sums over $\chi$. Thus one is to maximize the imse

$$
\frac{1}{N} \sum_{i=1}^{N} E \left[ \{ \hat{Y} (x_i) - E [Y (x_i)] \}^2 \right],
$$

subject to

$$
\sum_{i=1}^{N} \psi^2 (x_i) \leq \frac{\tau^2}{n}, \quad (2.2.12)
$$

and

$$
\sum_{i=1}^{N} f (x_i) \psi (x_i) = 0. \quad (2.2.13)
$$

It is convenient to temporarily transform to an orthogonal basis. Let $Q_1$ be an $N \times p$ matrix whose columns form an orthogonal basis for the column space of the matrix with rows $\{ f' (x) \mid x \in \chi \}$. Augment this
by $Q_2 : N \times (N - p)$ whose columns form an orthogonal basis for the orthogonal complement of this space. Then $(Q_1 : Q_2)$ is an orthogonal matrix and from (2.2.13), $\psi = (\psi (x_1), \ldots, \psi (x_N))'$ is orthogonal to the column space of $Q_1$, hence is necessarily of the form

$$\psi = \left(\tau / \sqrt{n}\right) Q_2 c,$$

where (by (2.2.12)) $\|c\| \leq 1$. Define $\xi_i = \xi (x_i)$ and

$$\tilde{A} = \frac{1}{N} \sum_{i=1}^{N} f (x_i) f' (x_i),$$

$$\tilde{M}_\xi = \sum_{i=1}^{N} \xi_i f (x_i) f' (x_i),$$

$$D_\xi = \text{diag} (\xi_1, \cdots, \xi_N).$$

Then the analogue of (2.2.6) is

$$\text{imse} = \frac{\sigma^2}{n} \text{trace} \left( \tilde{A} \tilde{M}_\xi^{-1} \right)$$

$$+ \frac{\tau^2}{nN} c' \left[ Q_2' D_\xi Q_1 (Q_1' D_\xi Q_1)^{-2} Q_1' D_\xi Q_2 + I_{N-p} \right] c.$$

(2.2.14)
Carrying out the maximization over $c$ and returning to the original notation gives a direct analogue of (2.2.11): in terms of

$$
\tilde{K}_\xi = \sum_{i=1}^{N} \xi_i^2 f(x_i) f'(x_i)
$$

and

$$
\tilde{H}_\xi = \tilde{M}_\xi \tilde{A}_\xi^{-1} \tilde{M}_\xi
$$

the maximum loss is \( \left( \sigma_\varepsilon^2 + \tau^2 \right) / n \) times

$$
\tilde{L}_\nu (\xi) = (1 - \nu) \text{trace} \left( \tilde{A} \tilde{M}_\xi^{-1} \right) + \nu \text{ch}_{\text{max}} \left( \tilde{K}_\xi \tilde{H}_\xi^{-1} \right).
$$

The minimization is carried out numerically – by simulated annealing, or via a genetic algorithm – yielding exact, i.e. integer-valued, designs. See the next figure for an example, with $N = 40$, $n = 20$ and $\nu = .5$; this illustrates a recurring theme – a naive yet sensible, and ‘near optimal’ method of robustifying a design is to take the replicates in the classically optimal design for the model in question and spread these out into nearby clusters.
Minimax design \( (n = 20) \) for approximate cubic regression when bias and variance receive equal weight.

- Lots of other possibilities - other loss functions, different neighbourhood structures, etc. See the Handbook chapter for more details.
Appendix

Additional material is posted online.