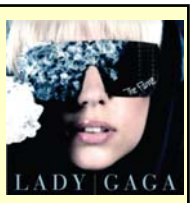


4. Dealing with epistatic interactions and non-linearities

gene x gene

gene x gene x gene

gene x gene x gene x gene

.....

(Alice in Wonderland)

Statistical Interaction (fixed effects models)

$$y_{ijk} = \mu + A_i + B_j + AB_{ij} + e_{ijk}$$

$$E(y_{ijk}|A_i, B_j, AB_{ij}) = \mu + A_i + B_j + AB_{ij}$$

$$\begin{aligned} E\left(y_{ijk} - y_{ij'k'}|A_i, B_j, AB_{ij}, A_{i'}, B_j, AB_{i'j}\right) &= \mu + A_i + B_j + AB_{ij} \\ &\quad - \left(\mu + A_{i'} + B_j + AB_{i'j}\right) \\ &= A_i - A_{i'} + AB_{ij} - AB_{i'j} \end{aligned}$$

Difference between levels of factor A depends on level of B

If factor *A* has *a* levels and factor *B* has *b* levels, the degrees of freedom are:
- *(a-1)*
- *(b-1)*
- *(a-1)(b-1)* [assuming no-empty cells]

Multi-SNP **Fixed** effects models?

(unraveling "physiological epistasis" a la Cheverud)

- Lots of "main effects"
- Splendid non-orthogonality
- Lots of 2-factor interactions
- Lots of 3-factor interactions
- Lots of non-estimability
- Lots of uninterpretable high-order interactions
- Run out of "degrees of freedom"

Analysis of SNPs with random effects models?

MEUWISSEN et al. (2001) ⇨ Will talk about this later

GIANOLA et al. (2003)

XU (2003)

} "Ridge regression-type"

--Use all SNP markers in statistical models
--Mechanistic basis to mixed effects linear model (genetic effects treated as random variables)
--Highly parametric models
--Strong assumptions made

What are ridge and Bayesian regression?
(given some variance components or tuning parameters)

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Large values of λ "shrink" regressions towards 0 (induces bias, but higher precision than OLS)

$$\hat{\boldsymbol{\beta}}_{RIDGE} = (\mathbf{X}'\mathbf{X} + \mathbf{I}\lambda)^{-1}\mathbf{X}'\mathbf{y}$$

Special case of Bayesian linear regression

$$\hat{\boldsymbol{\beta}}_{BAYES} = \left(\mathbf{X}'\mathbf{X} + \mathbf{B}^{-1}\frac{\sigma_e^2}{\sigma_\beta^2}\right)^{-1}\left(\mathbf{X}'\mathbf{y} + \frac{\sigma_e^2}{\sigma_\beta^2}\mathbf{B}^{-1}\boldsymbol{\beta}_0\right)$$

Bayes model assumes, a priori $\Rightarrow$ $\boldsymbol{\beta} \sim N(\boldsymbol{\beta}_0, \mathbf{B}\sigma_\beta^2)$

Typically assumed **0**

Typically identity matrix. However, can be given structure

ORDINARY LEAST-SQUARES

"Full model" $\Rightarrow$

$$y = X\beta + e$$
$$= X_1\beta_1 + X_2\beta_2 + e$$

"OLS" estimator $\Rightarrow$

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix}^{-1}\begin{bmatrix} X_1'y \\ X_2'y \end{bmatrix}$$
$$= [X'X]^{-1}X'y$$
$$E(\hat{\beta}|X) = [X'X]^{-1}X'E(y)$$
$$= [X'X]^{-1}X'X\beta = \beta$$

"OLS" is biased If full model holds and one fits "smaller" model (e.g., single marker Regressions)

$\Rightarrow$

$$y = X_1\beta_1 + e$$
$$E(\tilde{\beta}_1|X_1) = (X_1'X_1)^{-1}E(y)$$
$$= (X_1'X_1)^{-1}[X_1\beta_1 + X_2\beta_2]$$
$$= \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2$$

RIDGE REGRESSION

Can assess by cross-validation

$$\hat{\beta}_{Ridge} = [X'X + I\lambda]^{-1}X'y$$
$$= \left[I + (X'X)^{-1}\lambda\right]^{-1}(X'X)^{-1}X'y$$
$$= \left[I + (X'X)^{-1}\lambda\right]^{-1}\hat{\beta}_{OLS}$$

Shrinkage towards 0

$$E\left(\hat{\beta}_{Ridge}|X\right) = \left[I + (X'X)^{-1}\lambda\right]^{-1}E\left(\hat{\beta}_{OLS}\right)$$
$$= \left[I + (X'X)^{-1}\lambda\right]^{-1}\beta$$

Biased estimator but more precise

BAYESIAN REGRESSION (ASSUMING KNOWN VARIANCE COMPONENTS)

Prior

$$\beta \sim N(0, B\sigma^2_\beta)$$

$$\hat{\beta}_{Bayes} = \left[X'X + B^{-1}\frac{\sigma^2_e}{\sigma^2_\beta}\right]^{-1}\left(X'y + \frac{\sigma^2_e}{\sigma^2_\beta}B^{-1}\beta_0\right)$$

$$E\left(\hat{\beta}_{Bayes}|\beta\right) = \left[X'X + B^{-1}\frac{\sigma^2_e}{\sigma^2_\beta}\right]^{-1}\left(X'X\beta + \frac{\sigma^2_e}{\sigma^2_\beta}B^{-1}\beta_0\right)$$
$$= \left[I + (BX'X)^{-1}\frac{\sigma^2_e}{\sigma^2_\beta}\right]^{-1}(X'X)^{-1}\left(X'X\beta + \frac{\sigma^2_e}{\sigma^2_\beta}B^{-1}\beta_0\right)$$
$$= \left[I + (BX'X)^{-1}\frac{\sigma^2_e}{\sigma^2_\beta}\right]^{-1}\left(\beta + \frac{\sigma^2_e}{\sigma^2_\beta}(BX'X)^{-1}\beta_0\right)$$

Conditionally biased

ILLUSTRATION OF SOMEPOINTS

Standard analysis (fixed **X**) but random β

Genotype

$$y = f + e = X\beta + e$$
$$\beta \sim N(0, I\sigma^2_\beta)$$
$$E(y|X) = X\beta$$
$$Var(y|X) = Var(f) + Var(e)$$
$$= XX'\sigma^2_\beta + I\sigma^2_e$$

Prediction of marker effects: BLUP (iid marker effects)

$$\left[X'X + \frac{\sigma^2_e}{\sigma^2_\beta} I\right]\hat{\beta} = X'y$$
$$\left[I + \frac{\sigma^2_e}{\sigma^2_\beta}(X'X)^{-1}\right]\hat{\beta} = (X'X)^{-1}X'y$$
$$\hat{\beta} = \left[I + \frac{\sigma^2_e}{\sigma^2_\beta}(X'X)^{-1}\right]^{-1}\tilde{\beta}_{OLS} \Rightarrow \text{SHRINKAGE}$$

Prediction of signal ($X\beta$) to phenotype

$$Var(X\beta|y) = XVar(\beta|y)X'$$
$$= X\left[I + \frac{\sigma^2_e}{\sigma^2_\beta}(X'X)^{-1}\right]^{-1}X'\sigma^2_e$$

Prediction of future record

$$y^* = X^*\beta + e^*$$

$$E(X^*\beta + e^*|y, X, X^*) = X^*E(\beta|y, X)$$
$$= X^*\left[I + \frac{\sigma_e^2}{\sigma_\beta^2}(X'X)^{-1} \right]^{-1} \tilde{\beta}_{\text{OLS}}$$

$$Var(X^*\beta + e^*|y, X, X^*) = X^*Var(\beta|y, X)X^* + I^*\sigma_e^2$$

GAUSSIAN PROCESS ANALYSIS (IID MARKER EFFECTS)

$$y = f + e = X\beta + e$$
$$\beta \sim N(0, I\sigma_\beta^2)$$
$$X \sim F$$

Assume **X** and **β** are independent

$$E(y|X, \beta) = X\beta$$
$$E(y|\beta) = E_X E(y|X, \beta) = E(X)\beta$$
$$E(y) = E_\beta[E(X)\beta] = E(X)E(\beta) = 0$$

$$Var(y) = Var(f) + Var(e) = Var(f) + I\sigma_e^2$$

$$\begin{aligned}
Var(f) &= Var(X\beta) \\
&= E_X(Var(X\beta|X) + Var_X[E(X\beta|X)] \\
&= E_X[XVar(\beta)X']) + Var_X[XE(\beta)] \\
&= E_X[XX'\sigma_\beta^2] + Var_X(0) \\
&= \sigma_\beta^2 E_X[XX'],
\end{aligned}$$

BP= "best predictor" (MULVN assumed)

$$\begin{aligned}
\hat{f} &= BP(f) \\
\left[\frac{1}{\sigma_e^2}I + Var^{-1}(f)\right]\hat{f} &= \frac{1}{\sigma_e^2}y \\
\left[I + \frac{\sigma_e^2}{\sigma_\beta^2}E_X^{-1}[XX']\right]\hat{f} &= y \\
E_X^{-1}[XX']\left[E_X[XX'] + \frac{\sigma_e^2}{\sigma_\beta^2}I\right]\hat{f} &= y \\
\left[E_X[XX'] + \frac{\sigma_e^2}{\sigma_\beta^2}I\right]\hat{f} &= E_X[XX']y
\end{aligned}$$

Under multivariate normality

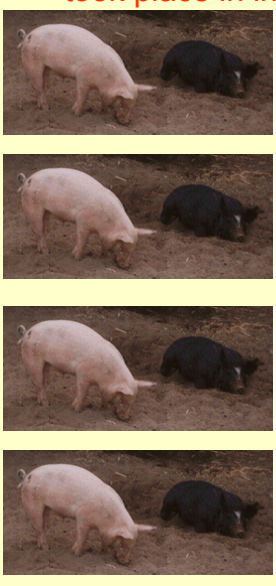
$$\begin{aligned}
Var(f|y) &= Var(f) - Cov(f,y)Var^{-1}(y)Cov'(f,y) \\
&= Var(f) - Var(f)[Var(f) + I\sigma_e^2]^{-1}Var(f) \\
&= \sigma_\beta^2 E_X[XX'] - \sigma_\beta^2 E_X[XX'][\sigma_\beta^2 E_X[XX'] + I\sigma_e^2]^{-1}\sigma_\beta^2 E_X[XX'] \\
&= \sigma_\beta^2 E_X[XX'] - \sigma_\beta^2 E_X[XX']\frac{E_X^{-1}[XX']}{\sigma_\beta^2}\left[I + \frac{\sigma_e^2}{\sigma_\beta^2}E_X^{-1}[XX']\right]^{-1}\sigma_\beta^2 E_X[XX'] \\
&= \left\{I - \left[I + \frac{\sigma_e^2}{\sigma_\beta^2}E_X^{-1}[XX']\right]^{-1}\right\}\sigma_\beta^2 E_X[XX'].
\end{aligned}$$

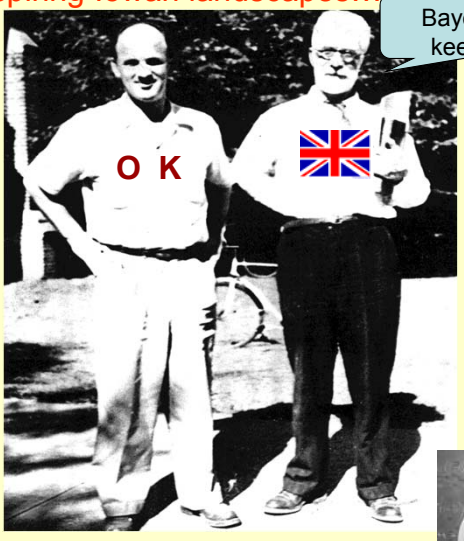
Future record:

$$
\begin{aligned}
f^* &= X^*\beta + e^* \\
E(f^*|f) &= E(f^*) + Cov(X^*\beta, \beta X')Var^{-1}(f)f \\
E(f^*|y) &= E_{f|y}E(f^*|f,y) = E_{f|y}E(f^*|f) \\
&= E_{f|y}[Cov(X^*\beta, \beta X')Var^{-1}(f)f] \\
&= Cov(X^*\beta, \beta X')Var^{-1}(f)\hat{f}
\end{aligned}
$$

$$
\begin{aligned}
Cov(X^*\beta, \beta X') &= E_{X,X^*}[Cov(X^*\beta, \beta X')|X, X^*] \\
&\quad + Cov_{X,X^*}[E(X^*\beta), E(X\beta)'|X, X^*] \\
&= \sigma^2_\beta E_{X,X^*}[X^*X'] + Cov_{X,X^*}(0,0) \\
&= \sigma^2_\beta E_{X,X^*}[X^*X']
\end{aligned}
$$

Dealing with interactions ("statistical epistasis"): much of this took place in inspiring Iowan landscapes…



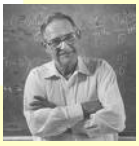


Bayesians, keep out!

SOME CORN

PIGS AGAIN

MORE PIGS HERE



C. C. C

$\Sigma_i\Sigma_j\Sigma_k\Sigma_l pig^2_{ijkl} - (\Sigma_i\Sigma_j\Sigma_k\Sigma_l pig_{ijkl})^2$ / as many pigs as you got

RANDOM EFFECTS MODELS
FOR ASSESSING EPISTASIS REST ON:
Cockerham (1954) and Kempthorne (1954)

--Orthogonal partition of genetic variance into additive, dominance, additive x additive, etc. **ONLY** if

- ❑No selection
- ❑No inbreeding
- ❑No assortative mating
- ❑No mutation
- ❑No migration
- ❑Linkage equilibrium

A standard decomposition of phenotypic value in quantitative genetics (Falconer & Mackay, 1996) is

$$y = \mu + a + d + i + e,$$

where a, d and i are additive, dominance and epistatic effects, respectively, and e is a residual, reflecting environmental (residual) variability. This linear de-

The i effect can be decomposed into additive × additive, additive × dominance, dominance × dominance, etc., deviates. In what has been termed 'statistical epistasis' (Cheverud & Routman, 1995), these deviates are assumed to be random draws from some distributions

The degrees of freedom of the distribution are NOT GIVEN by the number of levels.

There is now 1 df for each type of genetic effect.

$$N(0,\sigma^2_a)$$
$$N(0,\sigma^2_d)$$
$$N(0,\sigma^2_{aa})$$
$$N(0,\sigma^2_{ad})$$
$$N(0,\sigma^2_{dd})$$
$$\ldots.$$
$$N(0,\sigma^2_{ddd\ldots d})$$

Matrix representation

$$\begin{aligned}\mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + (\mathbf{a} + \mathbf{d} + \mathbf{i}_{aa} + \mathbf{i}_{ad} + \mathbf{i}_{dd}) + \mathbf{e} \\ &= \mathbf{X}\boldsymbol{\beta} + \mathbf{g} + \mathbf{e}, \end{aligned} \quad (1)$$

where $\boldsymbol{\beta}$ is some nuisance location vector (equal to μ if it contains a single element); $\mathbf{X}$ is a known incidence matrix; $\mathbf{a}$ and $\mathbf{d}$ are vectors of additive and dominance effects, respectively; $\mathbf{i}_{aa}$, $\mathbf{i}_{ad}$ and $\mathbf{i}_{dd}$ are epistatic effects, and $\mathbf{g} = \mathbf{a} + \mathbf{d} + \mathbf{i}_{aa} + \mathbf{i}_{ad} + \mathbf{i}_{dd}$ is the 'total' genetic value. Assuming that $\mathbf{g}$ and $\mathbf{e}$ are uncorrelated, the variance–covariance decomposition is

Variance-covariance

$$\mathbf{V}_y = \mathbf{V}_g + \mathbf{V}_e, \quad (2)$$

where $\mathbf{V}_y$, $\mathbf{V}_g$ and $\mathbf{V}_e$ are the phenotypic, genetic and residual variance–covariance matrices, respectively. Further,

Decomposition

$$\mathbf{V}_g = \mathbf{A}\sigma^2_a + \mathbf{D}\sigma^2_d + (\mathbf{A}\#\mathbf{A})\sigma^2_{aa} + (\mathbf{A}\#\mathbf{D})\sigma^2_{ad} + (\mathbf{D}\#\mathbf{D})\sigma^2_{dd}. \quad (3)$$

Here, $\mathbf{A}$ is the numerator relationship matrix; $\mathbf{D}$ is a matrix due to dominance relationships which can be computed from entries in $\mathbf{A}$, and the remaining matrices involve Hadamard (element by element) products of matrices $\mathbf{A}$ or $\mathbf{D}$. Thus, under CK, all

DO THESE ASSUMPTIONS HOLD?

RANDOM EFFECTS MODELS
FOR ASSESSING EPISTASIS REST ON:
Cockerham (1954) and Kempthorne (1954)

--Orthogonal partition of genetic variance into additive, dominance additive x additive, etc. **ONLY** if

- ❑No selection
- ❑No inbreeding
- ❑No assortative mating
- ❑No mutation
- ❑No migration
- ❑Linkage equilibrium

ALL ASSUMPTIONS VIOLATED!

Just consider Linkage disequilibrium

Digression: linkage disequilibrium

Let the genotypes at the first locus be A_1A_1, A_1A_2 and A_2A_2 and B_1B_1, B_1B_2 and B_2B_2 at the second locus. Let the frequency of the A_1 allele be p_{A_1}, of the A_2 allele be p_{A_2}, $(p_{A_1} + p_{A_2} = 1)$ and at the B locus the equivalent frequencies are p_{B_1} and p_{B_2}, $(p_{B_1} + p_{B_2} = 1)$. The four possible gametes are A_1B_1, A_1B_2, A_2B_1, A_2B_2, with respective frequencies $p_{A_1B_1}$, $p_{A_1B_2}$, $p_{A_2B_1}$, $p_{A_2B_2}$ and $p_{A_1B_1} + p_{A_1B_2} + p_{A_2B_1} + p_{A_2B_2} = 1$. Notice that

$$\begin{aligned} p_{A_1} &= p_{A_1B_1} + p_{A_1B_2}, \\ p_{A_2} &= p_{A_2B_1} + p_{A_2B_2}, \\ p_{B_1} &= p_{A_1B_1} + p_{A_2B_1}, \\ p_{B_2} &= p_{A_1B_2} + p_{A_2B_2}. \end{aligned}$$

If the allelic state at locus A is independent of that at locus B, one expects $p_{A_1B_1} = p_{A_1}p_{B_1}$, $p_{A_1B_2} = p_{A_1}p_{B_2}$, and so on. The system is said to be in *linkage equilibrium*: the alleles at loci A and B are independent and their joint frequency is given by the product of their

marginal frequencies. If this is not the case, the dependence between alleles at loci A and B is measured by their covariance, known as *linkage disequilibrium* and symbolised by D. Define the random variable X which takes the value 1 if in gametes, A_i is present at locus A and zero otherwise, and the random variable Y which takes the value 1 if B_j is present at locus B zero otherwise. The expected value of X is p_{A_i} and that of Y is p_{B_j}. The expected value of (XY) is $p_{A_iB_j}$ and the covariance between X and Y is by definition

$$\begin{aligned} D &= Cov(X,Y) \\ &= E(X,Y) - E(X)E(Y) \\ &= p_{A_iB_j} - p_{A_i}p_{B_j}. \end{aligned} \tag{1}$$

For example, if we set arbitrarily $i = 1$, $j = 1$, then

$$p_{A_1B_1} = p_{A_1}p_{B_1} + D, \tag{2}$$

and

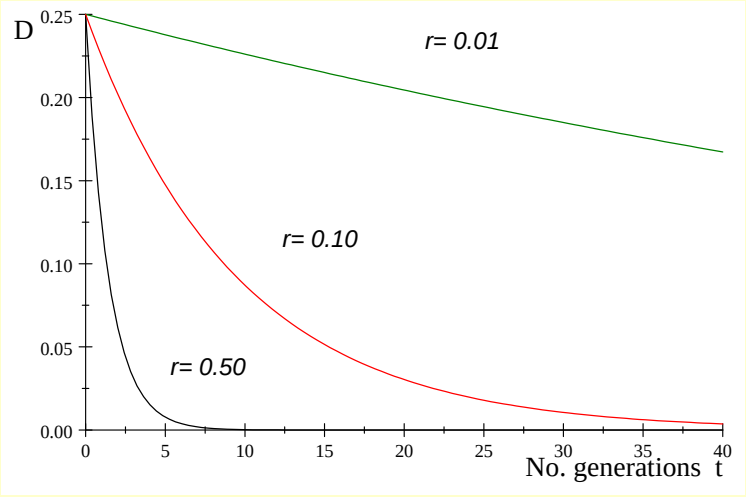
$$\begin{aligned} D &= p_{A_1B_1} - p_{A_1}p_{B_1} \\ &= p_{A_1B_1}(p_{A_1B_1} + p_{A_1B_2} + p_{A_2B_1} + p_{A_2B_2}) - (p_{A_1B_1} + p_{A_1B_2})(p_{A_1B_1} + p_{A_2B_1}) \\ &= p_{A_1B_1}p_{A_2B_2} - p_{A_1B_2}p_{A_2B_1}, \end{aligned} \tag{3}$$

the difference between the product of the frequencies of the *coupling* and *repulsion* gametic phases. Choosing $i = 1$, $j = 1$ resulted in (2) and in

$$\begin{aligned} p_{A_1B_2} &= p_{A_1}p_{B_2} - D, \\ p_{A_2B_1} &= p_{A_2}p_{B_1} - D, \\ p_{A_2B_2} &= p_{A_2}p_{B_2} + D. \end{aligned}$$

Evolution of linkage disequilibrium as a function of recombination rate

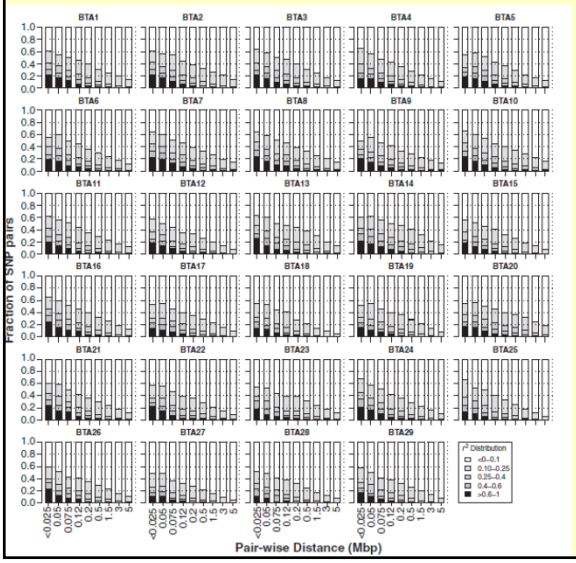
$$D_t = (1-r)^t D_0$$



ANIMAL GENETICS Immunogenetics, Molecular Genetics and Functional Genomics
doi:10.1111/j.1365-2052.2009.02011.x

The pattern of linkage disequilibrium in German Holstein cattle

S. Qanbari*, E. C. G. Pimentel*, J. Tetens†, G. Thaller†, P. Lichtner‡, A. R. Sharifi* and H. Simianer*



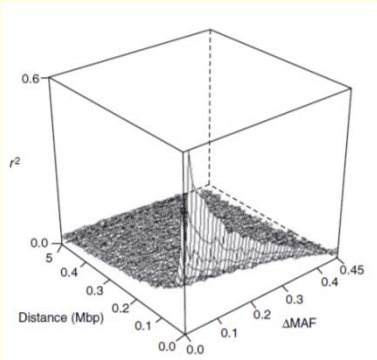


Figure 5 Three-dimensional surface plot depicting the decay of linkage disequilibrium vs. inter-marker distance and minor allelic frequency interval.

A VIEW OF LINEAR MODELS (as employed in q. genetics)

Mathematically, can be viewed as a "local" approximation of a complex process

$$f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f^{(3)}(a)}{3!}(x-a)^3 + \ldots + \frac{f^{(n)}(a)}{n!}(x-a)^n + \ldots.$$

Linear approximation

Quadratic approximation

n^{th} order approximation

Example

Response variate

Model residual

$$y = g(x) + e$$

Some function of a covariate x

Suppose $g(x) = \sin(x) + \cos(x)$

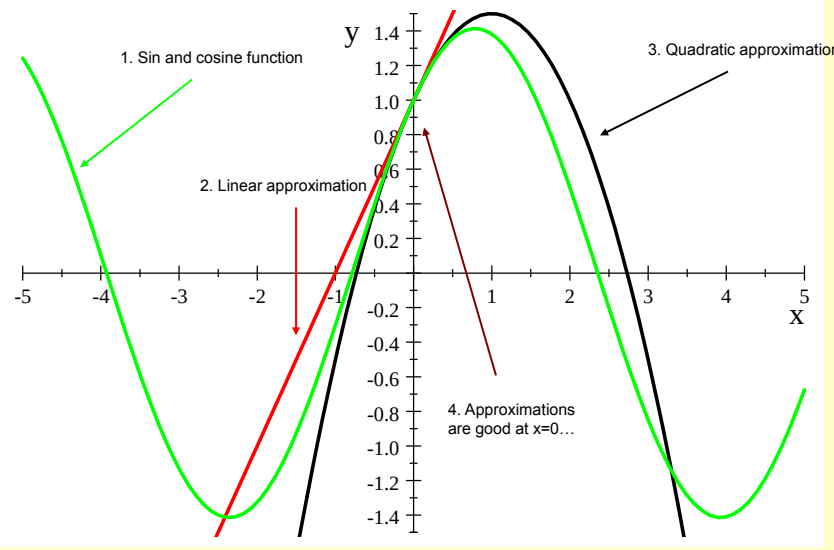
$$\frac{d}{dx}[\sin(x)] = [\cos x]$$
$$\frac{d}{dx}[\cos(x)] = [-\sin x]$$
$$\frac{d}{dx}[\sin(x) + \cos(x)] = [\cos x - \sin x]$$
$$\frac{d}{dx}[\cos x - \sin x] = [-\cos x - \sin x]$$
$$\frac{d^2}{(dx)^2}[\sin(x) + \cos(x)] = [-\cos x - \sin x]$$

Second-order Taylor series expansion about 0

$$[\sin(x) + \cos(x)] \approx [\sin(0) + \cos(0)] + [\cos 0 - \sin 0](x - 0) + \frac{1}{2}[-\cos 0 - \sin 0](x - 0)^2$$

$$= 1 + x - \frac{x^2}{2}$$

How good are the linear and quadratic approximations? Recall that a Taylor series provides a local approximation only…



Finding structure from noisy data
we have environmental noise…:

evaluate function sin(x)+cos(x) at x=0, 0.5 and 1

True values are:

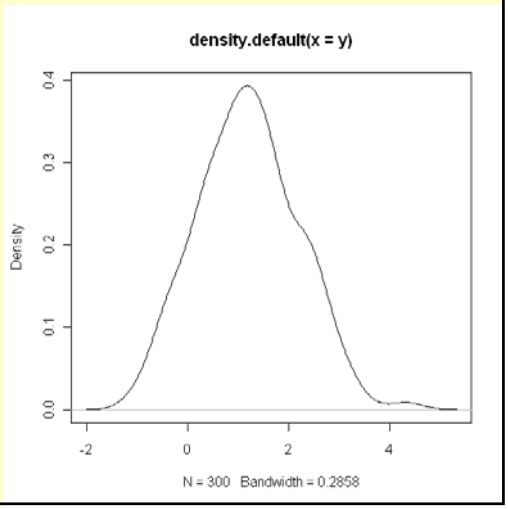
```
> sin(0)+cos(0)
  [1] 1

> sin(0.5)+cos(0.5)
  [1] 1.357008

> sin(1)+cos(1)
  [1] 1.381773
```

VERY CLOSE TO EACH OTHER
NOISE CAN MASK SIGNALS!

Create an R data set (N=300) from adding 100 N(0,1) residuals to each of the 3 values

```
> y0<-sin(0)+cos(0)       +rnorm(100,0,1)
> y05<-sin(0.5)+cos(0.5)+rnorm(100,0,1)
> y1<-sin(1)+cos(1)       +rnorm(100,0,1)
> y<-c(y0,y05,y1)
```

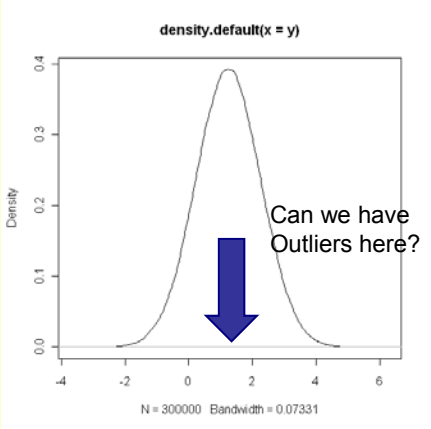
MEASURING MACHINE 1



Create a larger R data set (N=300000) by adding 100000 N(0,1) residuals to each of the 3 values

```
> y0<-sin(0)+cos(0)      +rnorm(100000,0,1)
> y05<-sin(0.5)+cos(0.5) +rnorm(100000,0,1)
> y1<-sin(1)+cos(1)      +rnorm(100000,0,1)
> y<-c(y0,y05,y1)
```

CANNOT SEE UNDERLYING STRUCTURE.
LARGE NOISE (ERROR VARIANCE)

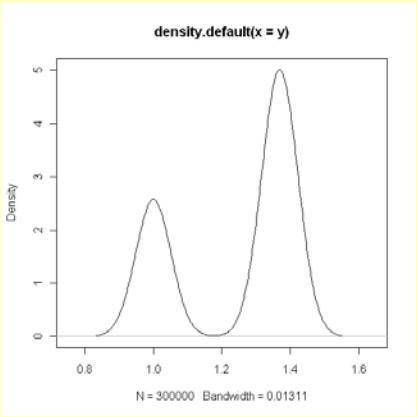


Now we get a more precise measuring instrument with variance 0.05

```
> y0<-sin(0)+cos(0)      +rnorm(100000,0,.05)
> y1<-sin(1)+cos(1)      +rnorm(100000,0,.05)
> y05<-sin(0.5)+cos(0.5) +rnorm(100000,0,.05)
```

MEASURING MACHINE 2

STRUCTURE IS REVEALED BUT
WE CANNOT DIFFERENTIATE
BETWEEN TWO OF THE UNDERLYING
VALUES



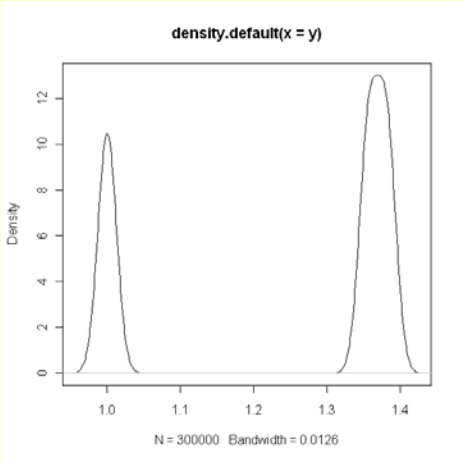
…SO WE BUY ANOTHER INSTRUMENT WITH VARIANCE 0.001!

```
> y0<-sin(0)+cos(0)        +rnorm(100000,0,.001)
> y1<-sin(1)+cos(1)        +rnorm(100000,0,.001)
> y05<-sin(0.5)+cos(0.5)   +rnorm(100000,0,.001)
> y<-c(y0,y05,y1)
```

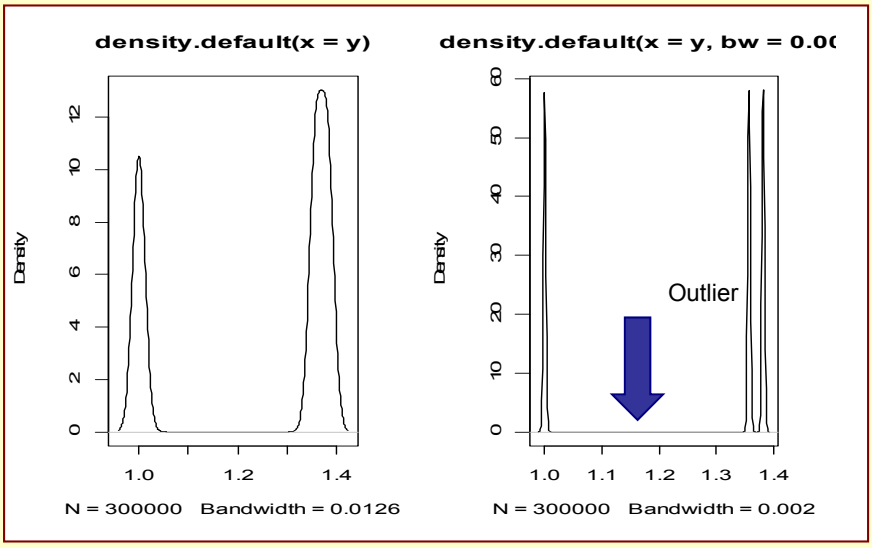
MEASUREMENT MACHINE 3

STILL CANNOT DIFFERENTIATE BETWEEN THE

```
> sin(0.5)+cos(0.5)
   [1] 1.357008

> sin(1)+cos(1)
   [1] 1.381773
```



HOWEVER, NON-PARAMETRIC DENSITY ESTIMATES DEPEND ON SOME BANDWIDTH PARAMETER. BY REDUCING IT, WE CAN SEE THE ENTIRE STRUCTURE OF THE PROBLEM…

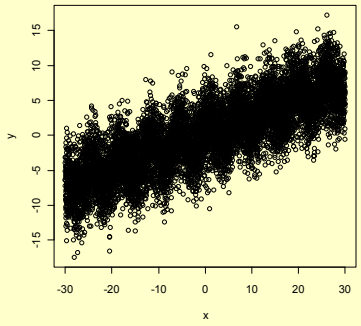


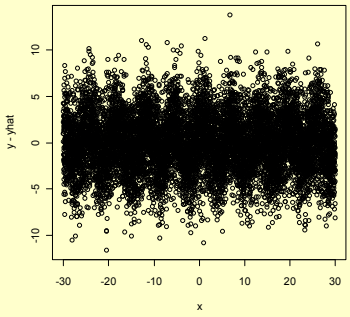
FINDING "STRUCTURE" WITH A LINEAR MODEL

We are given (x,y) data (n=10,000). It looks like this and we run a linear regression

yhat= 0.07936+0.24814*x

```
> cor(x,y)
[1] 0.8064256

> cor(y,yhat)
[1] 0.8064256
```



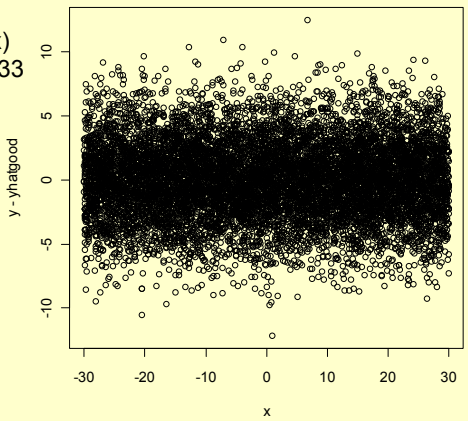


RESIDUALS DISPLAY
SINUSOIDAL BEHAVIOR

TRUE MODEL

```
> e<-rnorm(10000,0,sqrt(9))
> x<-runif(10000,-30,30)
> a<-0.10
> b<-0.25
>y<-a+b*x+sin(x)+cos(x)+e

> model<-lm(y~x+sin(x)+cos(x))

>Coefficients:
>(Intercept)          x      sin(x)      cos(x)
> 0.1030       0.2489       0.9518      0.9433
```

RESIDUALS LOOK RANDOM



WE GENERATE A NEW SAMPLE AT THE SAME VALUES OF X

```
> enew<-rnorm(10000,0,sqrt(9))
>ynew<-a+b*x+sin(x)+cos(x)+enew
```

CALCULATE PREDICTIVE MEAN SQUARED ERROR

```
> msepredbadmodel<-sum((ynew-yhat)**2/10000)
> msepredbadmodel
[1] 9.725709

> msepredgoodmodel<-sum((ynew-yhatgood)**2/10000)
> msepredgoodmodel
[1] 8.729272
```

CALCULATE PREDICTIVE CORRELATIONS

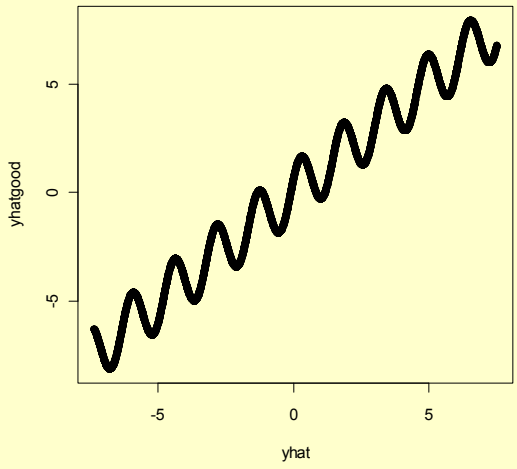
```
> cor(yhat,ynew)
[1] 0.8070097
> cor(yhatgood,ynew)
[1] 0.828854
```

MSE(Good)/MSE(Bad)=0.8975
MSE(Bad)/MSE(Good)=1.1141

Cor(BAD)/Cor(GOOD)=0.9736

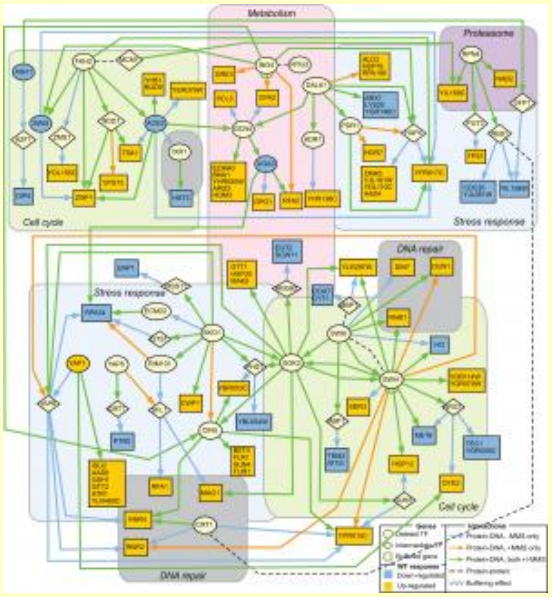
```
> lm(yhat~yhatgood)
Coefficients:
(Intercept)     yhatgood
   0.005653     0.953468
```

DO NOT TRUST CORRELATIONS!



RECALLING COMPLEXITY…

How one
Would model
something like this?



Heal Thyself: Systems Biology Model Reveals How Cells Avoid Becoming Cancerous. ScienceDaily (May 21, 2006)

What to do in genomic-assisted analysis of complex genetic signals?

- Include all markers, model all possible interactions? Unrealistic…
- Select sets of influential markers via model selection
  - ➔ Huge search space
  - ➔ Frequentist methods "err" probabilistically
  - ➔ Bayesian model selection (RJMC) difficult to tune
- Use LASSO (least absolute shrinkage and selection operator): Tibshirani (1996). What about interactions?
- Explore model-free techniques that have been used successfully in many domains
  - ➔ **semi-parametric regression**
  - ➔ **machine learning**: focus on prediction, learning mappings from inputs to outputs

DEFINITION OF MACHINE LEARNING (Wikipedia)

Machine learning: subfield of artificial intelligence concerned with design and development of algorithms that allow computers (machines) to improve their performance over time (to learn) based on data,

A major focus of machine learning research is to automatically produce (induce) models, such as rules and patterns, from data. Hence, machine learning is closely related to fields such as data mining, statistics, inductive reasoning, pattern recognition, and theoretical computer science.