Analysing networks / inputs

- Data – eye gaze
  - Network: 12-7-1, standard backprop.
- Analysing the weight matrix
  - Magnitude measures & brute force analysis
  - Functional measures & guided elimination
  - Sensitivity measures
- Comments
  - Weights ≠ functionality
  - ‘Functional’ measures
  - Behaviour >> ‘functional’ measures

Eye Gaze Data

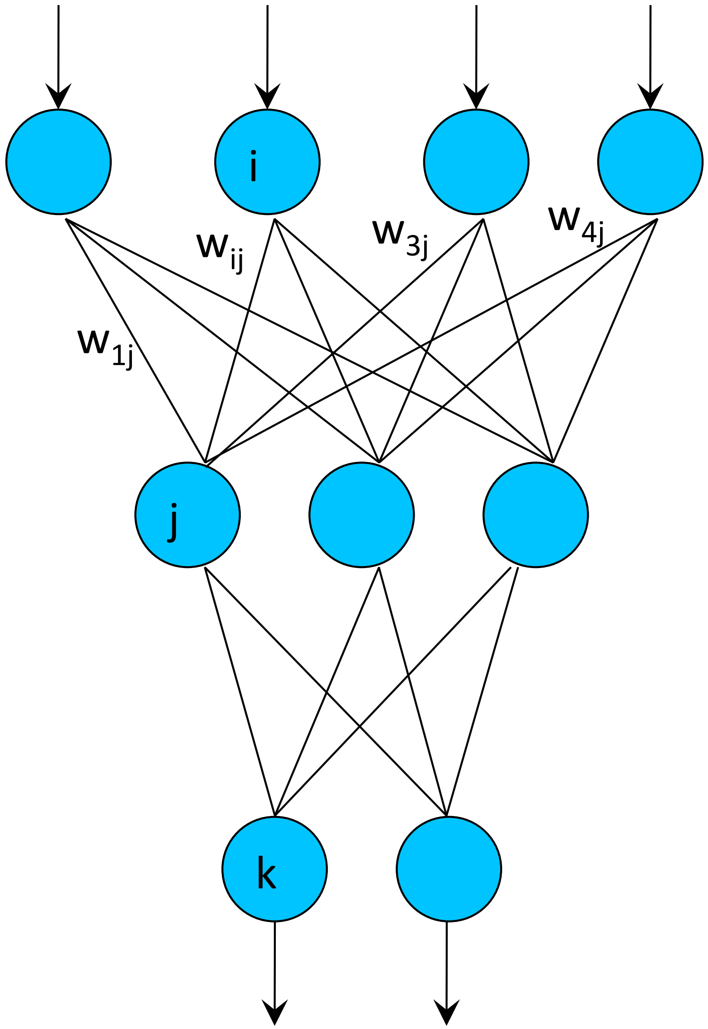
- Eye gaze detector at Westmead Hospital
- 10 schizophrenic and 10 normal individuals
- 4 responses, 10 secs long, recorded at 50 Hz
  - wire frame drawing
  - neutral affect face
  - happy face
  - sad face
- Task: separate schizophrenics from normals
- (Problem: medicated schizophrenics, and unmedicated normals)

Garson '91

- G_{ik} = contribution of input *i* to output *k*.
  - Sum the
    fraction of weight *i* to *j* to all weights to *j*
    modulated by weight *j* to *k*.
  - Divide by sum of all paths.
- Disadvantage
  - Positive and negative weights cancel, some contributions lost

$$G_{ik} = \frac{\displaystyle\sum_{j=1}^{nh} \frac{W_{ij}}{\sum_{p=1}^{ni} W_{pj}} . W_{jk}}{\displaystyle\sum_{q=1}^{ni} \left(\sum_{j=1}^{nh} \frac{W_{qj}}{\sum_{p=1}^{ni} W_{pj}} . W_{qj} \right)}$$

Example using Garson's formula



- What proportion of effect of inputs on hidden unit *j* is due to input *i* ?
  - Modify by effect of *j* on *k*
- Do for all paths from *i* to *k*

$$G_{ik} = \frac{\displaystyle\sum_{j=1}^{nh} \frac{W_{ij}}{\displaystyle\sum_{p=1}^{ni} W_{pj}} . W_{jk}}{\displaystyle\sum_{q=1}^{ni} \left(\sum_{j=1}^{nh} \frac{W_{qj}}{\displaystyle\sum_{p=1}^{ni} W_{pj}} . W_{qj} \right)}$$

Milne '95

- M_{ik} = contribution of input *i* to output *k*.
  - Sum the
    fraction of weight *i* to *j* to all abs. weights to *j*
    modulated by weight *j* to *k*.
  - Divide by sum of abs.
    value of all paths.
- Advantage: sign
- Disadvantage
  - Divisor unclear meaning

$$M_{ik} = \frac{\displaystyle\sum_{j=1}^{nh} \frac{w_{ij}}{\sum_{p=1}^{ni} \left|w_{pj}\right|} . w_{jk}}{\displaystyle\sum_{q=1}^{ni} \left(\sum_{j=1}^{nh} \left| \frac{w_{qj}}{\sum_{p=1}^{ni} \left|w_{pj}\right|} . w_{qj} \right| \right)}$$

Wong et al '95 & Gedeon '96

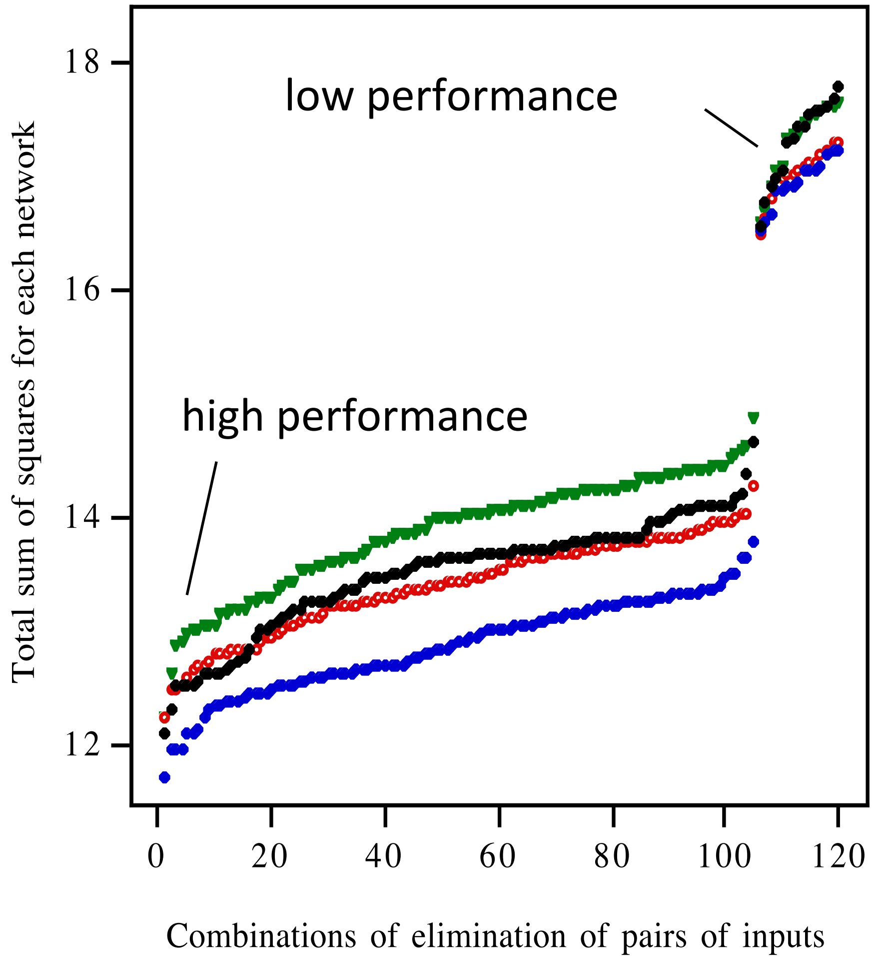
- P_{ij} = absolute value contribution of input *i* to hidden *j*.
  - Fraction of weight *i* to *j* to all weights to *j*.
- Q_{ik} = extend for input *i* to output *k*.
- Advantage
  - Magnitude contribution calculated
  - Sign clear from weight value

$$P_{ij} = \frac{\left|w_{ij}\right|}{\sum_{p=1}^{ni} \left|w_{pj}\right|}$$

$$Q_{ik} = \sum_{r=1}^{nh} \left(P_{ir} \times P_{rk}\right)$$

Brute Force Analysis

- eliminate
  - 1 input inconsistent
  - 2 inputs
    - 120 possib. x 4 runs
    - (different data set)
    - values sorted by total sum squares
    - discontinuity at significant loss of performance



Magnitude measures - brute force

- Use brute force method as basis to compare.
- Proportion match to most important:
  - Q >> M ≈ G
- (Prop. match to least important: ◊ Q ≈ G > M)

model	Most significant...					Least signif.				
B	5	1	10	14	11	...	7	8	13	9	6
Q	11	10	2	12	14	...	13	5	4	8	7
G	2	4	6	7	11	...	9	13	14	1	15
M	11	15	7	13	12	...	8	10	2	4	5

- I.e. Q >> G > M

Functional measures – vector angles

- I – vector components from each pattern
  for each input create 1,334 dimensional vector
- C – aggregate of I, average angle to others
- W – vector components from input weights
  – modified weight distinctiveness – Gedeon '96b
- U – aggregate of W, again averaged angles

Analysis – functional measures

- Rank techniques: U > W >> C > I
  - Analysis of network better than merely analysing the data & aggregation is useful
  - Validated by elimination suggested by each of the techniques
- Next:
  - Sensitivity measures
    - Effects of perturbing inputs instead of elimination.

Sensitivity

- Perturb in sequence
  - single artificial pattern
  - single average pattern
  - all patterns, single input
  - all, pairs of inputs
  - all, triples of inputs
  - all, fours
  - all, fives
- Accumulate Δs

1)	2)	3)	4)	5)	6)	7)
0.5	av.	Δ1	Δ2	Δ3	Δ4	Δ5
9	9	9	9	9	9	9
3	3	3	3	3	3	3
8	8	1	7	7	7	7
10	2	7	10	10	10	10
1	10	10	8	1	1	1
6	6	6	1	8	8	8
2	1	8	6	5	5	5
4	7	11	11	2	2	2
11	4	4	4	12	12	12
7	11	12	12	4	11	11
5	12	2	5	11	4	4
12	5	5	2	6	6	6

Analysis – sensitivity

- Use sum of squared difference of ranks to compare

I	C	W	U	Mag	Sens.
11	3	9	6	4	9
10	10	8	4	5	3
3	8	6	3	11	7
8	11	3	2	6	10
4	7	11	7	10	1
1	1	4	5	1	8
6	2	2	1	2	5
5	9	5	10	8	2
2	6	10	9	3	12
9	12	7	8	7	11
7	5	12	12	12	4
12	4	1	11	9	6
290	272	318	322	406	$\sum(X-S)^2$
2	1	3	4	5	Rank
I	C	W	U	Mag	Model

Summary so far

- Functional
  - Rank techniques: U > W >> C > I
- Sensitivity
  - Rank techniques: C > I >> W > U
  - Results no better than from training pattern set, worse than measure on network weights.
- Sensitivity to an input does not necessarily correlate with importance of an input.
  - Affect output without affecting classification?

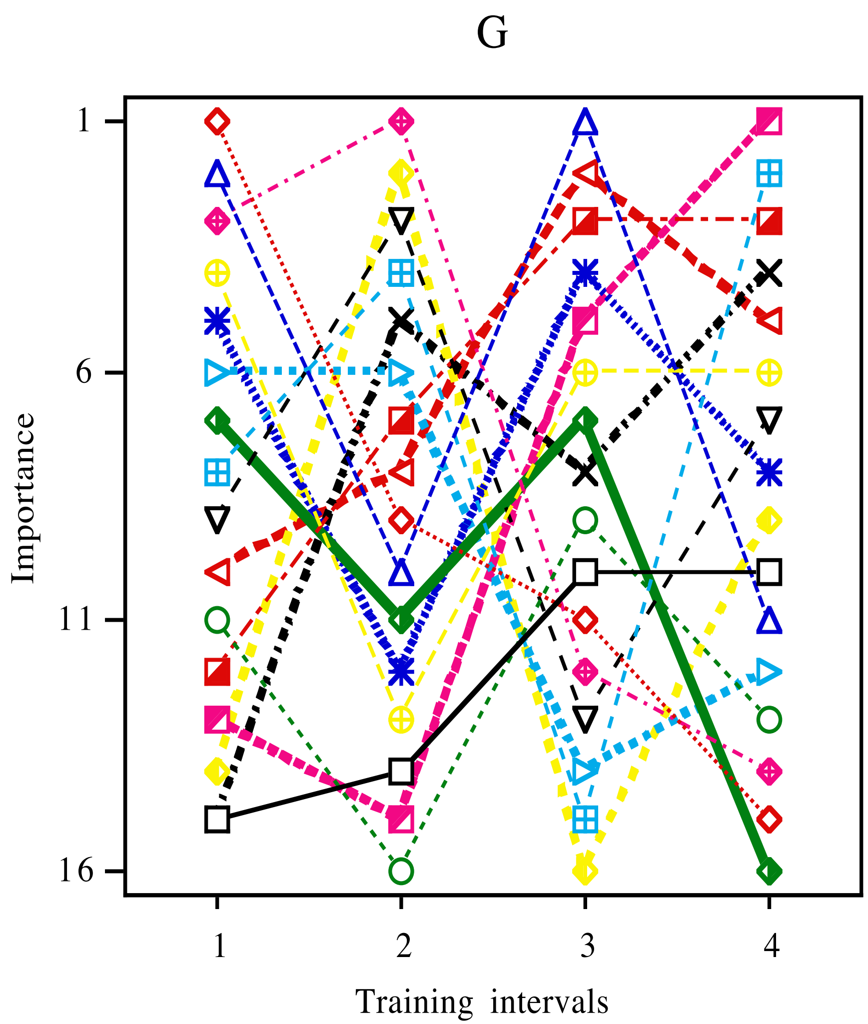
Stability of Mag. Techniques

- Observation: significance of inputs change during normal network training.
- Premise: significance of inputs will change slowly on overtraining beyond best result on test set.

Stability - G

- Very unstable.
- Few inputs are similar over the adjacent intervals.

$$G_{ik} = \frac{\sum_{j=1}^{nh} \frac{W_{ij}}{\sum_{p=1}^{ni} W_{pj}} . W_{jk}}{\sum_{q=1}^{ni} \left(\sum_{j=1}^{nh} \frac{W_{qj}}{\sum_{p=1}^{ni} W_{pj}} . W_{qj} \right)}$$

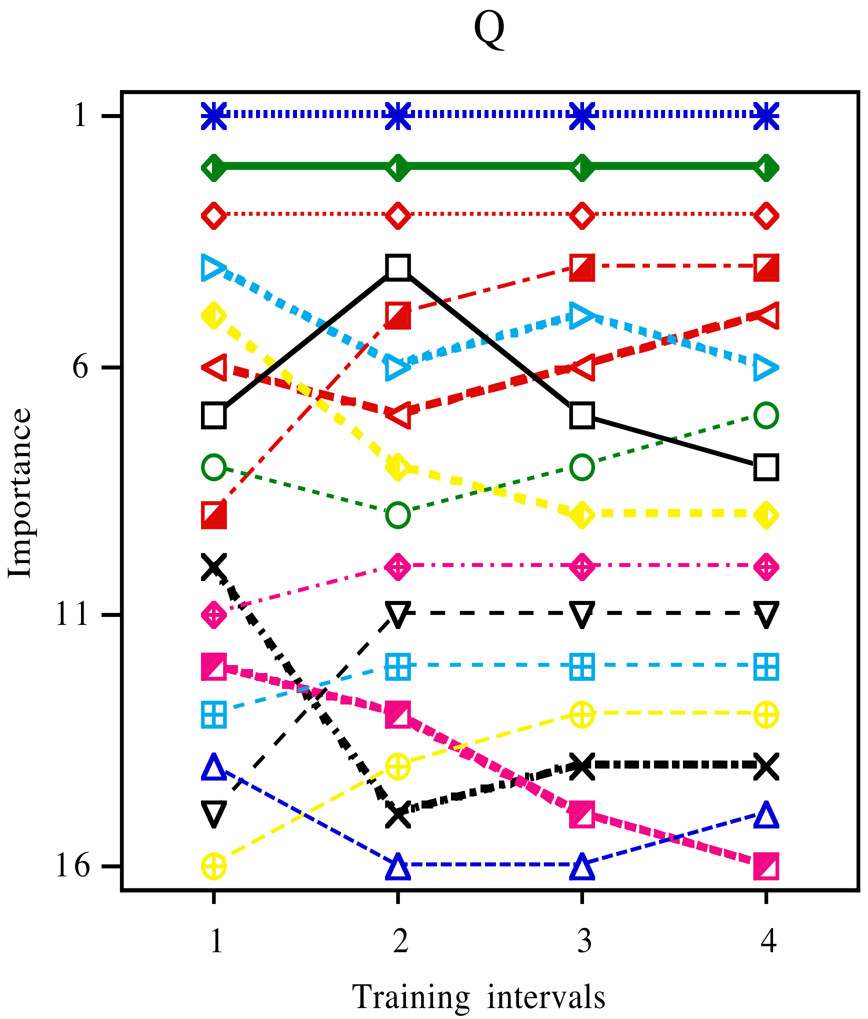


Stability - Q

- Quite stable.
- Most inputs have <u>same</u> values over adjacent intervals.

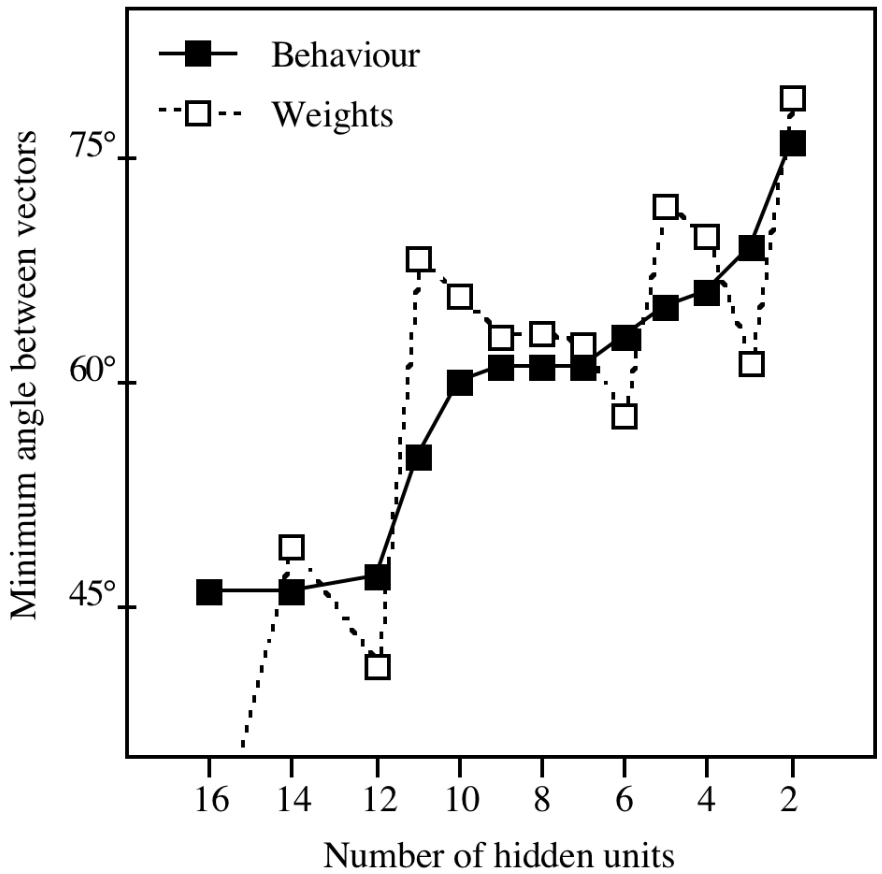
$$P_{ij} = \frac{|w_{ij}|}{\sum_{p=1}^{ni} |w_{pj}|}$$

$$Q_{ik} = \sum_{r=1}^{nh} \left(P_{ir} \times P_{rk}\right)$$



Weights versus behaviour

- Comparison of changes in input weight matrix versus pattern activations (using the image compression example)
- Weight matrix is only coarsely approximating the behaviour of the hidden units while the no. of hidden units are reduced
- Behaviour >> Weights



All weights versus behaviour

- Similar, input weights approximate all weights
- Conclude the true functionality of the neural network 'black box' is found from behaviour not examining its innards

Correlation Matrix for $X_1 \dots X_4$

	behav	out wts	in wts	all wts
behav	1			
out wts	.83	1		
in wts	.84	.77	1	
all wts	.9	.96	.76	1