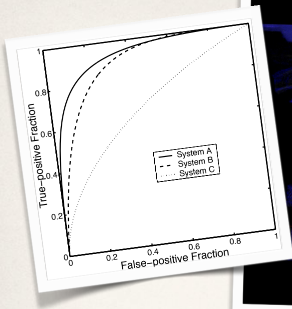
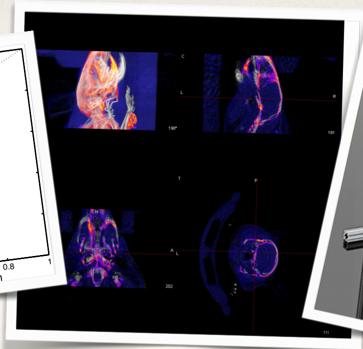
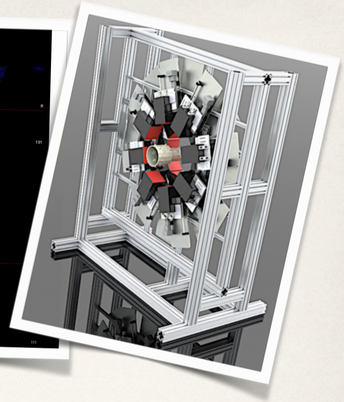


Observer Models and Realism Testing

Matthew Kupinski

July 22, 2014

Virtual Clinical Trial

- What is the relevant task?
  Detect tumors, detect and locate tumors, estimate parameters
- What objects will be imaged?
  The patient population and the statistics that govern it.
- What is the imaging system(s)?
  The physics and statistics of the system which determine how an object is mapped to image data.
- How will the information be extracted?
  The observer.
- What measure of performance will be used?
  The figure of merit

Virtual Clinical Trial

- What is the relevant task?
  Detect tumors, detect and locate tumors, estimate parameters
- What objects will be imaged?
  The patient population and the statistics that govern it.
- What is the imaging system(s)?
  The physics and statistics of the system which determine how an object is mapped to image data.
- **How will the information be extracted?**
  The observer.
- What measure of performance will be used?
  The figure of merit

Observers

- Decision-making strategy
- Options
  -Human
  -Model
  -Anthropomorphic model
  -Computer-aided diagnosis (CAD)

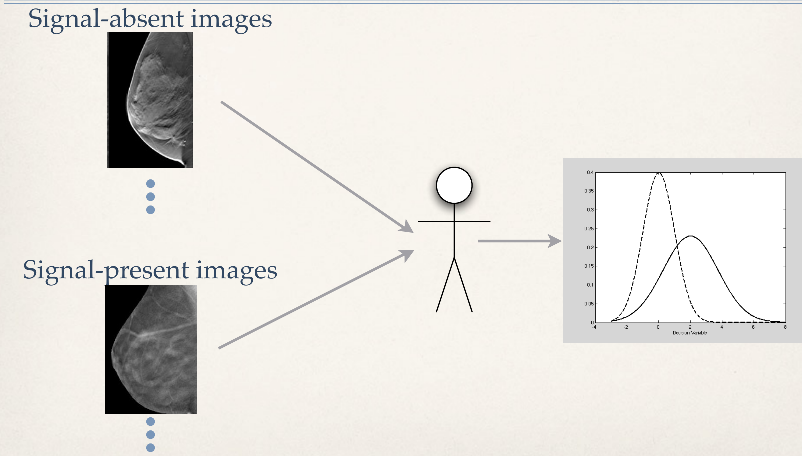
Observer review

- Classification
  $t = T(\boldsymbol{g})$

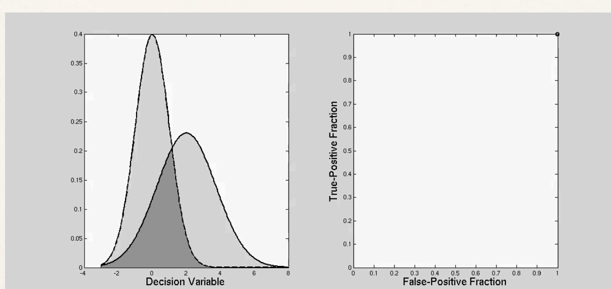
- Estimation
  $\widehat{\boldsymbol{\theta}} = \Theta(\boldsymbol{g})$

- Combination
  $t = T(\boldsymbol{g})$
  $\widehat{\boldsymbol{\theta}} = \Theta(\boldsymbol{g})$

Classification



ROC analysis



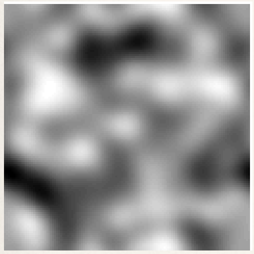
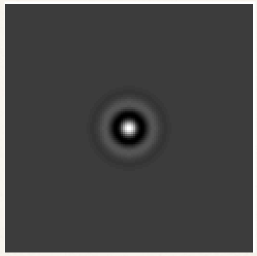
Hotelling observer

$$t = \boldsymbol{w}^{\dagger}\boldsymbol{g}$$

$$w = K_{\boldsymbol{g}}^{-1}\Delta\overline{\boldsymbol{g}}$$

- Is the ideal observer when the image data are Gaussian distributed
- Maximizes the SNR among all linear observers
- Requires knowledge of the first- and second-order statistics
- A linear template applied to the image

Ideal linear observers

 $\times$  $= t$

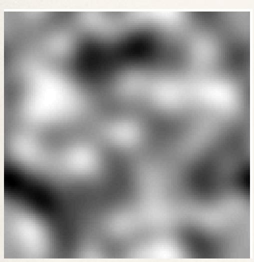
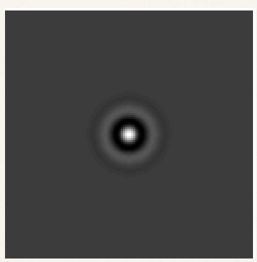
Channelized Hotelling observer

$$\boldsymbol{v} = T\boldsymbol{g}$$

$$t = \boldsymbol{\xi}^{\dagger}\boldsymbol{v} \qquad \boldsymbol{\xi} = K_{\xi}^{-1}\Delta\overline{\boldsymbol{\xi}}$$

- Dimension of $\boldsymbol{v}$ is much smaller than $\boldsymbol{g}$
- Ideal linear observer on channel outputs
- Channels can be chosen to mimic human visual system
- Internal noise model is needed to match absolute performance

Ideal linear observers

 $\times$  $= t$

Limitations of Hotelling observers

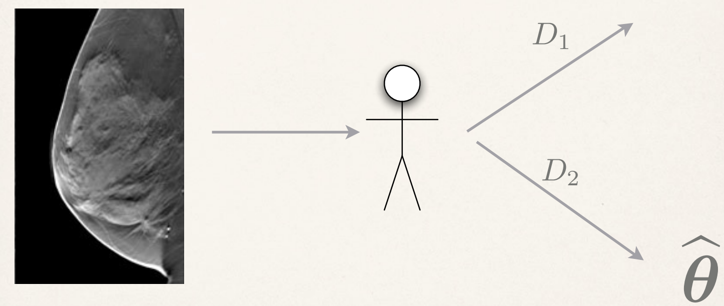
- Signal location is fixed
- Performance is limited in the presence of signal-shape variability
- Signal must often be very weak to achieve AUC < 1.0

Limitations of Hotelling observers

- Signal location is fi
- Performance is lim hape variability
- Signal must often be very weak to achieve AUC < 1.0

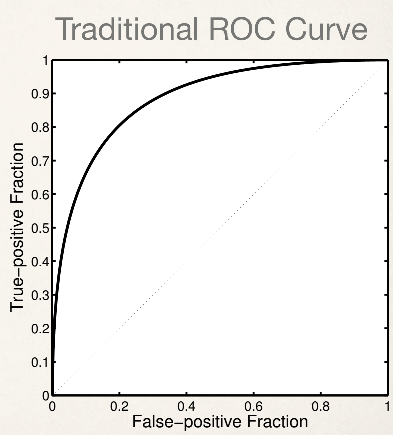
Task is limited

Combined tasks



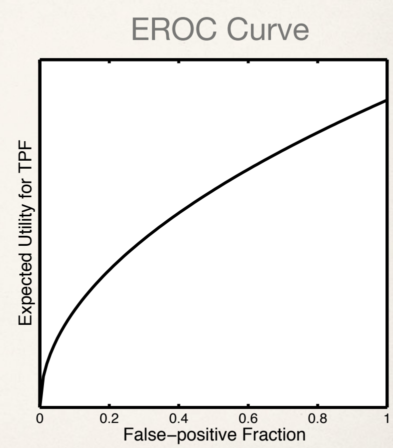
EROC

- TPF vs. FPF
- AUC from 2AFC test
- Ideal observer is likelihood ratio



EROC

- Expected utility for TPF vs. FPF
- AEROC from 2AFC test
- Ideal observer is known
- Specializes to LROC for signal localization tasks



Scanning linear observer

$$t(\boldsymbol{g},\boldsymbol{\theta}) = \Delta\overline{\boldsymbol{g}}(\boldsymbol{\theta})^{\dagger}K_g^{-1}(\boldsymbol{g}-\overline{\boldsymbol{g}}) - \frac{1}{2}\Delta\overline{\boldsymbol{g}}(\boldsymbol{\theta})^{\dagger}K_g^{-1}\Delta\overline{\boldsymbol{g}}(\boldsymbol{\theta})$$

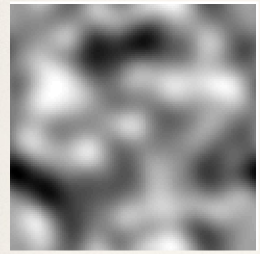
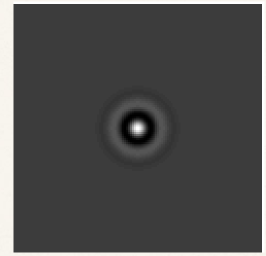
$$t(\boldsymbol{g}) = \max_{\boldsymbol{\theta}}\ t(g,\boldsymbol{\theta})$$

$$\widehat{\boldsymbol{\theta}}(\boldsymbol{g}) = \arg\max_{\boldsymbol{\theta}}\ t(\boldsymbol{g},\boldsymbol{\theta})$$

- Maximizes area under the EROC curve when the image statistics are Gaussian distributed and a delta utility function is assumed
- Non-linear observer due to the maximization steps

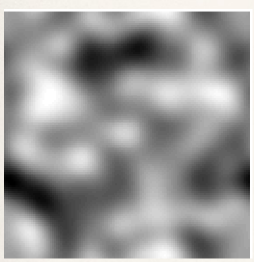
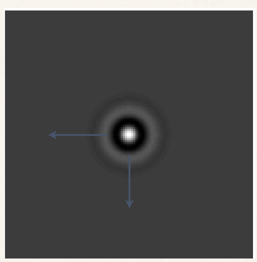
Scanning linear observer

Estimating signal position

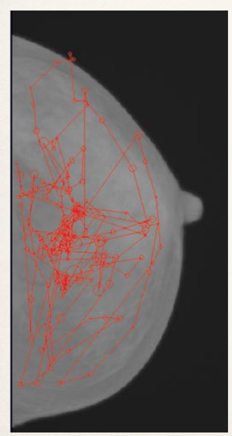
 $\times$  $= t(\boldsymbol{g},\boldsymbol{\theta})$

Scanning linear observer

Estimating signal position

 $\times$  $= t(\boldsymbol{g},\boldsymbol{\theta})$

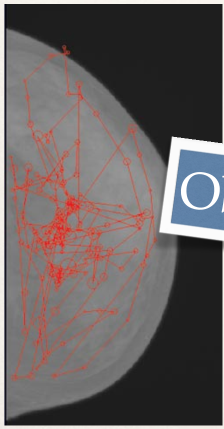
Anthropomorphic models?



- Human scan pattern is difficult to model

Courtesy of E. Krupinski

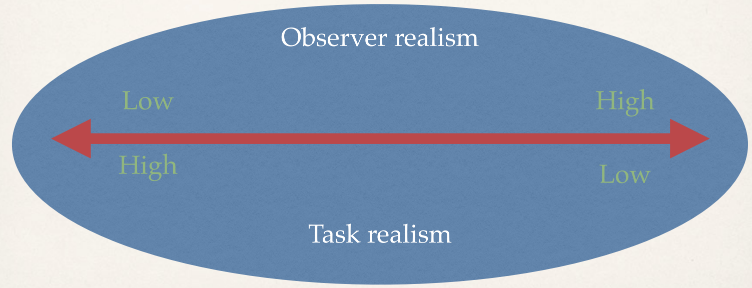
Anthropomorphic models?



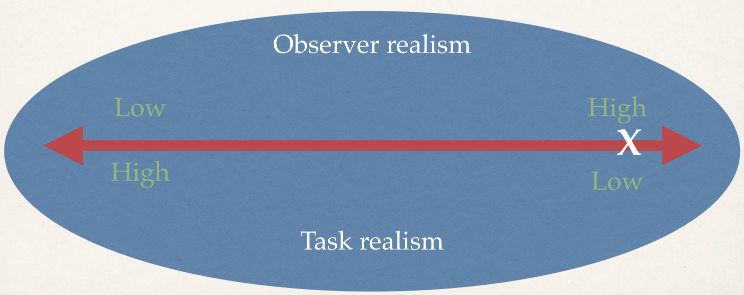
Observer is limited to model

Courtesy of E. Krupinski

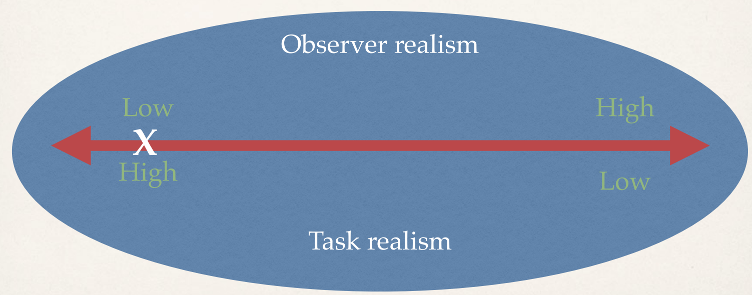
Problem



Problem



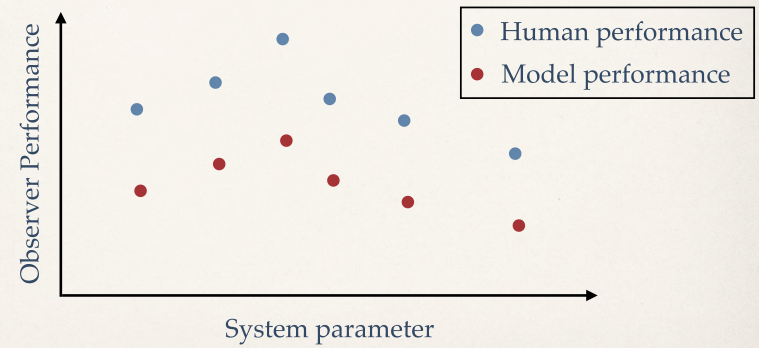
Problem



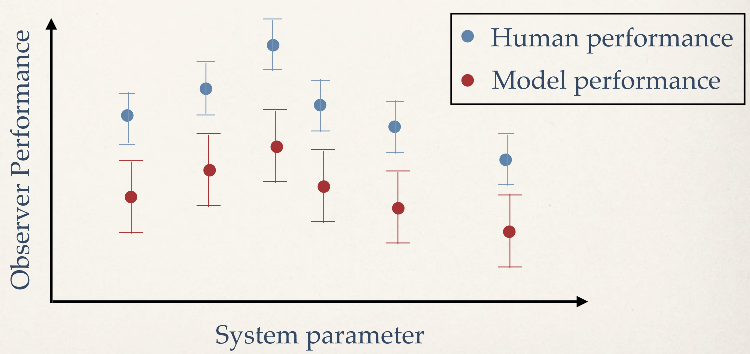
Solution

- Does absolute performance matter?
- What matters is that the same decision is made using the model observer
- Rank ordering
- **May allow for simpler tasks and observers**

Rank ordering



Rank ordering



Ranking Systems

- A *judge* is a program producing the estimates of the performance for a system (may include multiple readers)
- Judge j produces a rank R_{js} for system s
- We want to know how consistent these rankings are among the judges

Ranking Systems

$$\bar{R}_s = \frac{1}{J}\sum_{j=1}^{J} R_{js} \qquad \bar{R} = \frac{1}{S}\sum_{s=1}^{S} \bar{R}_s = \frac{S+1}{2}$$

$$SS_t = J\sum_{s=1}^{S}\left(\bar{R}_s - \bar{R}\right)^2$$

$$SS_e = \frac{1}{J\,(S-1)}\sum_{j=1}^{J}\sum_{s=1}^{S}\left(R_{js} - \bar{R}\right)^2 = \frac{S\,(S+1)}{12}$$

The Friedman statistic is $Q = \dfrac{SS_t}{SS_e}$

Ranking Systems

- The Friedman statistic is usually used in hypothesis testing where the null hypothesis is that the rankings of the judges are uncorrelated
- Under the null hypothesis the distribution of Q can be approximated by a chi-square distribution with S - 1 degrees of freedom
- A large value for Q indicates agreement among the judges

Ranking Systems

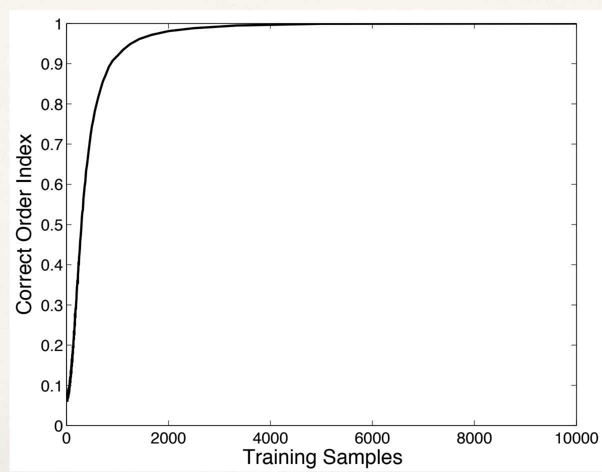
- We want a large value for Q, but how large?
- The maximum value for Q is

$$Q_{max} = J\,(S-1)$$

Define a sample variance for each system

$$\sigma_s^2 = \frac{1}{J}\sum_{j=1}^{J}\left(R_{js} - \bar{R}_s\right)^2$$

Ranking Systems



Task-Based Turing Test

- Problem: Is the simulated data realistic enough?
- Question: Realistic enough for what?
- Answer: Realistic enough to give results for estimators in simulation that match what we get with real data
- We must quantify how well the results match

Task-Based Turing Test

- Consider parameter μ with estimator μ_{est}
- μ_{est} has a PDF $pr_1(\mu_{\text{est}})$ with the simulated data
- μ_{est} has a PDF $pr_2(\mu_{\text{est}})$ with the real data
- Want to measure how well these PDFs match

Task-Based Turing Test

To measure the match between the two PDFs we will use the Wicoxon statistic

$$W = \frac{1}{N_r} \sum_{r=1}^{N_r} W_r$$

$$W_r = \frac{1}{N_1 N_2} \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} step\left[\mu_{est}\left(g_{2j}\right) - \mu_{est}\left(g_{1i}\right)\right]$$

MRMC Analysis applies

$$\sigma_W^2 = \frac{C_1}{N_1} + \frac{C_2}{N_2} + \frac{C_3}{N_r} + \frac{C_4}{N_1 N_2} + \frac{C_5}{N_1 N_r} + \frac{C_6}{N_2 N_r} + \frac{C_7}{N_1 N_2 N_r}$$

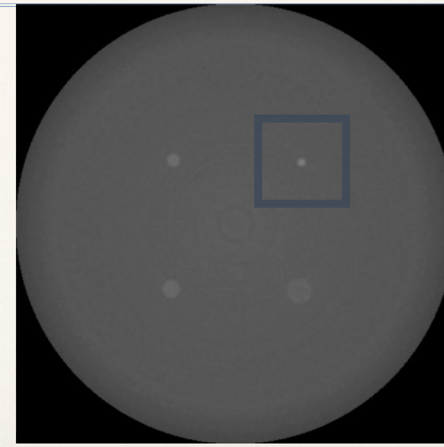
Task-Based Turing Test

- W is usually used to reject the null hypothesis that the two distributions are the same
- W is also an estimate of the AUC for the two PDFs
- If the two PDFs are the same, then the AUC is 0.5

Task-Based Turing Test

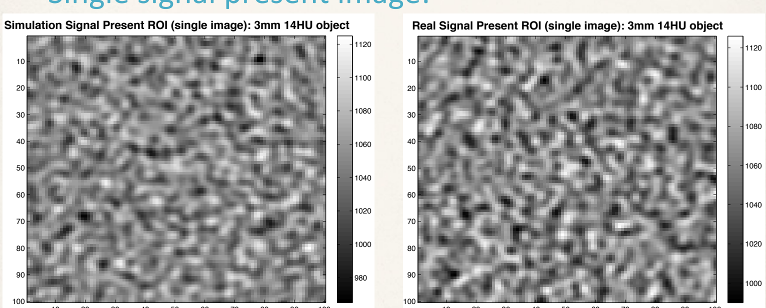
- If 0.5 is within one standard deviation of W we will consider the two PDFs to be a good match
- As the number of cases and readers increases the standard deviation decreases and this becomes a more demanding test
- Using MRMC analysis we can compute the numbers of readers and cases needed to reach the level of accuracy desired

Task-Based Turing Test



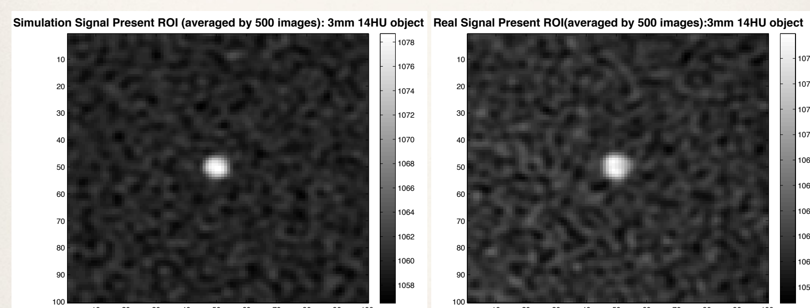
Task-Based Turing Test

Single signal present image:

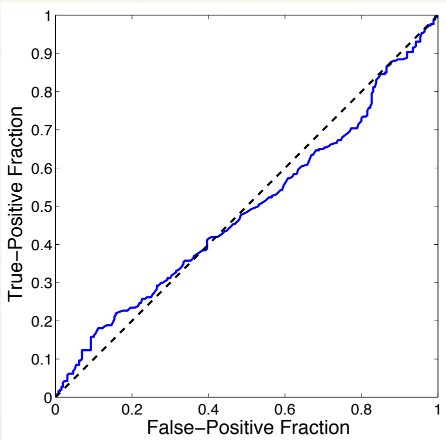


Task-Based Turing Test

Signal present image (500 averaged):



Task-Based Turing Test



Summary

- Observer models can predict human performance for relatively simple tasks
- Consistent ranking may be more important than absolute quantitation
- Task-based methods can be used to test model realism