Data Quality and Inverse Problems

Bayesian Image Reconstruction Workshop
U. Manchester, UK:
4 March 2013

Andy Adler

Professor & Canada Research Chair in Biomedical Engineering
Systems and Computer Engineering, Carleton University

Data Quality

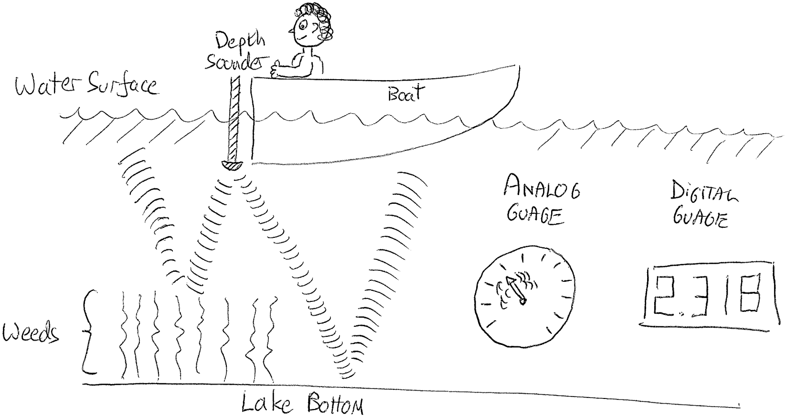
What is data quality?

- Wrong Data – Data Errors
- Uninformative Data
- Misinterpreted Data

Outline:

- Examples
- Bayesian Framework
- Electrode Errors
- Data Quality Measures
- Thoughts from Biometrics & Information Theory
- A way forward?

Example #1: Data Quality

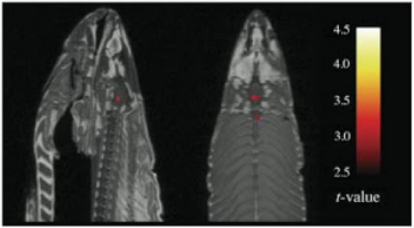


Depth Sounder – with analog and digital guages

Example #2: Data Quality

neuroskeptic.blogspot.ca/2009/09/fmri-gets-slap-in-face-with-dead-fish.html

Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction

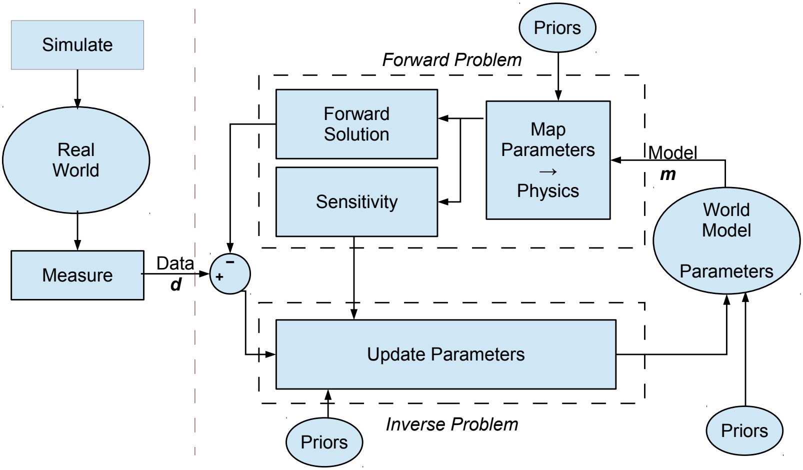


This is a poster presented by Bennett and colleagues at this year's Human Brain Mapping conference. It's about fMRI scanning on a dead fish, specifically a salmon. They put the salmon in an MRI scanner and *"the salmon was shown a series of photographs depicting human individuals in social situations. The salmon was asked to determine what emotion the individual in the photo must have been experiencing."*

According to the authors, subject, “not alive at the time of scanning”

neuroskeptic.blogspot.ca/2009/09/fmri-gets-slap-in-face-with-dead-fish.html

Inverse Problem Framework



Framework for inverse problems. *Note all the priors*

What's the problem?

With strong priors, algorithms give us pretty pictures, even when they are irrelevant.

Question:

- how can we know when to trust a pretty picture?
- how can we know when the data are junk?

Bayesian Formulation

- Forward Problem
  (Data d, Parameters m, noise n)

$$d = F(m) + n$$

- Noise Model

$$n \sim \mathcal{N}(0, \Sigma_n)$$

- Data Posterior Probablility

$$p(d|m) \propto exp\left(-\frac{1}{2}\|d - F(m)\|_W^2\right), \qquad W = {\Sigma_n}^{-1}$$

Bayesian Formulation

- Parameters Prior Probability

$$
\begin{aligned}
m &\sim \mathcal{N}(m, \Sigma_m) \\
p(m) &\propto exp\left(-\tfrac{1}{2}\|m - m_0\|_P^2\right), \qquad P = \Sigma_m^{-1}
\end{aligned}
$$

- Posterior Probability

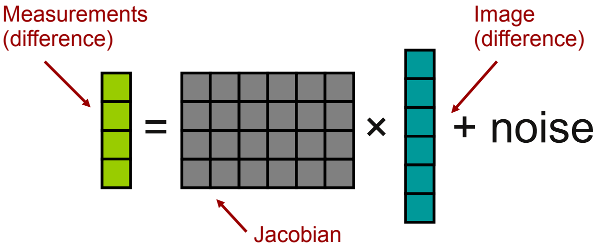
$$
\begin{aligned}
p(m|d) &\propto p(d|m)p(m) \\
&\propto exp\left(-\tfrac{1}{2}\|d - F(m)\|_W^2 - \tfrac{1}{2}\|m - m_0\|_P^2\right)
\end{aligned}
$$

- MAP solution minimizes norm

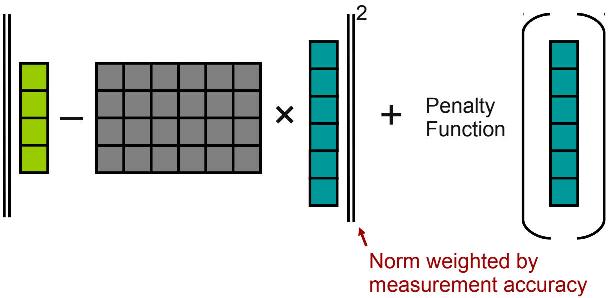
$$
\|d - F(m)\|_W^2 + \|m - m_0\|_P^2
$$

Reconstruction in Pictures

- Forward Problem

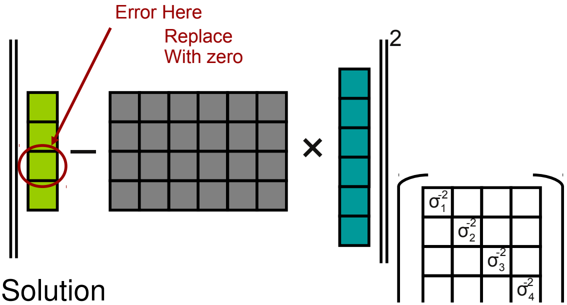


- MAP Solution Norm

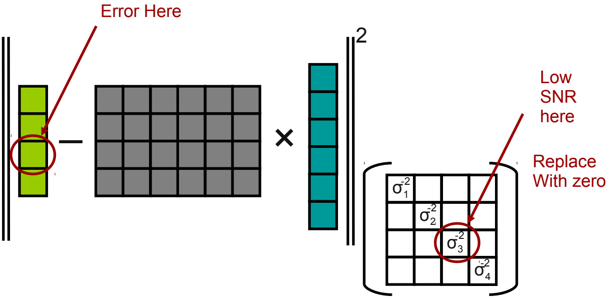


Reconstruction with Data Errors

- "Traditional" Solution

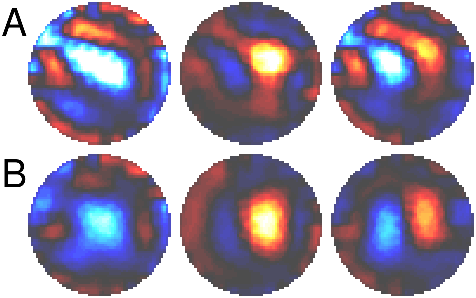


- Error Model Solution



Electrode Error compensation

- Offline compensation using “jack-knife” approach (2005)



EIT images in anaesthetised, ventilated dog
A: uncompensated, *B*: compensated. *Left*: ventilation *Centre*: saline (right lung) *Right*: ventilation and saline

- Automatic detection (via reciprocity comparison) (2009)
- New work to speed online calculation & use data quality

Data Quality Measure: Concept

- *Concept:* High Quality Data is Consistent
- *Idea:* Use IP to predict each data point from all others

$$\boxed{\begin{array}{c}\text{Original}\\ d\end{array}} \Rightarrow \boxed{\begin{array}{c}\text{Remove } i\\ d^{(i)}\end{array}} \Rightarrow \boxed{\begin{array}{c}\text{Solve}\\ \hat{m}^{(i)}\end{array}} \Rightarrow \boxed{\begin{array}{c}\text{Predict}\\ \hat{d}^{(i)}\end{array}}$$

- Calculate error

$$\epsilon_i = d_i - \hat{d}_i^{(i)}$$

Data Quality Measure: Linear Case

- ϵ from data d

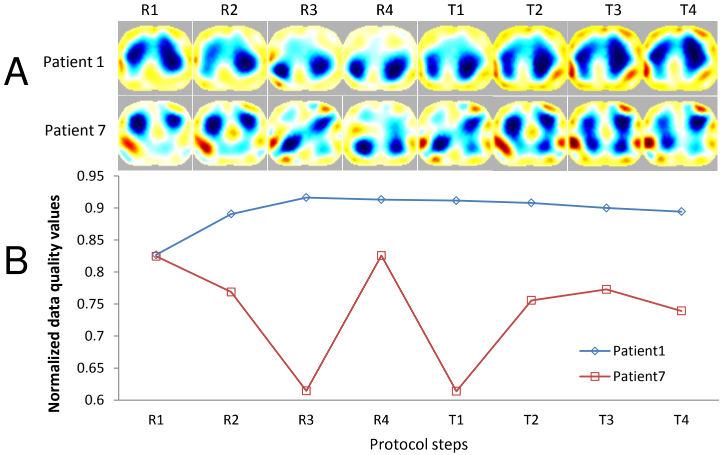
$$
\begin{aligned}
\epsilon_i &= d_i - \hat{d}_i^{(i)} && \text{where } \hat{d}^{(i)} \text{ is predicted without } i \\
\hat{d}^{(i)} &= J\hat{m}^{(i)} = JRd \\
\hat{d}^{(i)} &= JR^{(i)}d && \text{where } R^{(i)} \text{ is Rec. Matrix without } i \\
R &= \Sigma_m J^t \left(J\Sigma_m J^t + \Sigma_n\right)^{-1} \\
R^{(i)} &= \Sigma_m J^t \left(J\Sigma_m J^t + \alpha^2 I + \mu^2 \Xi^{(i)}\right)^{-1} && \text{where } \Xi^{(i)} \text{ is 1 at } (i,i) \\
\epsilon_i &= [d - \hat{d}^{(i)}]_i = [d - JR^{(i)}d]_i = [I - JR^{(i)}]_i d
\end{aligned}
$$

- Quality Matrix Q

$$
\begin{aligned}
\epsilon &= Qd \\
\text{where } Q_i &= I - J^t\Sigma_m J^t \left(J\Sigma_m J^t + \alpha^2 I + \mu^2 \Xi^{(i)}\right)^{-1}
\end{aligned}
$$

- Q calculation can be optimized

Example: Data quality measures



Clinical data and data quality metric for each stage of the protocol (R1–R4 — recruitment: PEEP⇑, T1–T4 — titration: PEEP⇓). *A*: EIT images (one-step Gauss-Newton solver with a 2D forward model), *B*: Calculated data quality.

Example: Robust Algorithms

$$\|d - F(m)\|_{\ell_d} + \|m - m_0\|_{\ell_m}$$

- ℓ_1 norm for the image prior allows "blocky" reconstructions
- ℓ_1 norm for the data mismatch gives improved robustness to outliers

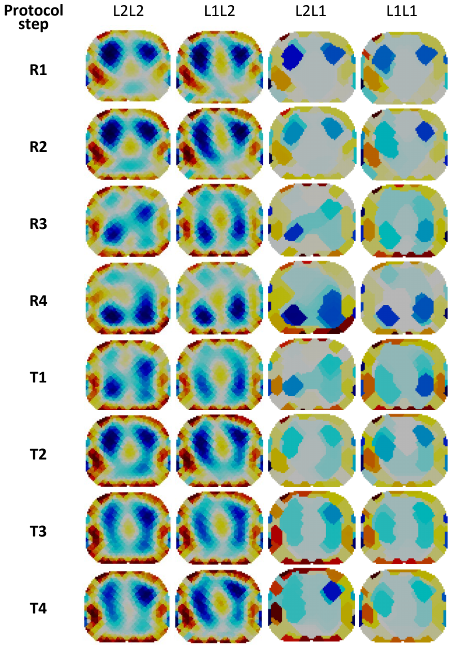


Figure: Reconstructions with mixed (data/image) norms for clinical data for each stage of the protocol (R1–R4 — recruitment: PEEP$\Uparrow$, T1–T4 — titration: PEEP$\Downarrow$).

A way forward?

- Inverse Problems are hard; priors are useful; users like pretty pictures
  $\Rightarrow$ the situation will get worse
- Complex systems fail in complex ways
- Very complex systems (human brains) fail in extremely sophisticated ways. These errors are carefully researched (psychology, neuroscience)
- Idea: we need a new research area — *inverse problem* problems IP^2
  $\Rightarrow$ goal: understand/classify situations where IPs fail.

Low Quality Data are Less Informative:

Measuring Quality via Information Content

Andy Adler, Richard Youmaran
Carleton University, Ottawa, Canada

From: Presentation at NIST Biometric Biometric Quality Workshop II, Nov 2007

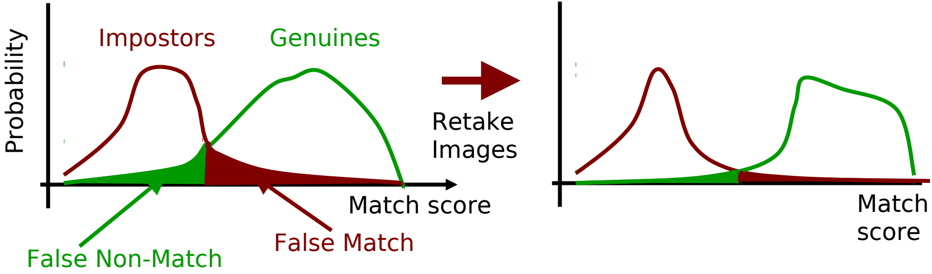
Biometric Sample Quality

Biometric Sample Quality measures:

- character
  - inherent features
- Fidelity
  - accuracy of features
- Utility
  - predicted system performance

INCITS, Biometric Sample Quality Standard Draft, M1/06-0003

Utility Quality



Since the algorithm errors were less, the retaken images had higher quality

But, could we have done better with the first images?

Utility Quality

- The ability of the system to use the data to achieve low error rates

- Dependent on processing algorithm
- Doesn’t measure “inherent” quality in the data

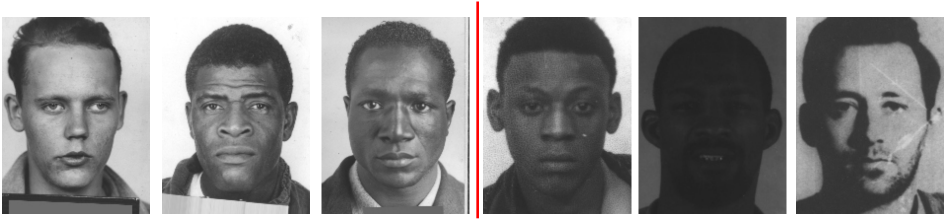
Character / Fidelity

Descriptions of “inherent” quality of a biometric sample

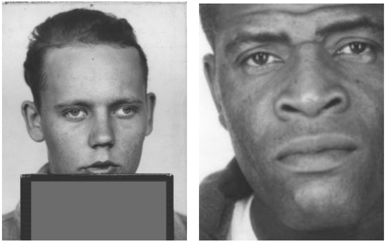
- Character Quality Problems
  - Blur
  - Shadows
  - Poor lighting
- Fidelity Quality Problems
  - A good image of the wrong part

Example: *Character Quality*

← Best Faces **Human Selections** Worst Faces →



Example: *Fidelity Quality*



How can we measure character quality?

- Probing question:

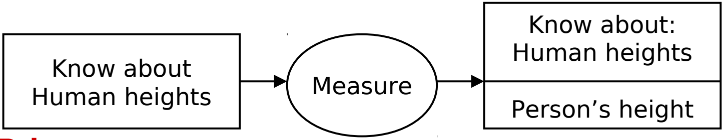
  Why do we worry about low quality data?

- Answer:

  They have less information .

Definition: *MI* (Mutual Information)
Measurement Information:

- the decrease in uncertainty about the identity of an individual based on a measurement of biometric features.



Prior:
Uncertainty is 1:6 billion

$$q(y) \Longrightarrow p(y)$$

Posterior:
Uncertainty is less

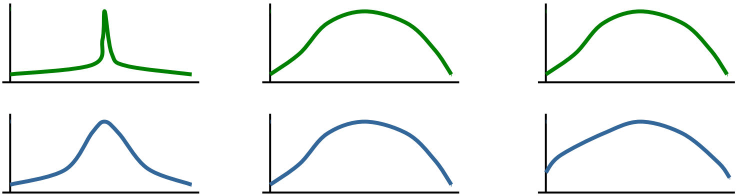
- Measure **KLD (**Kullback-Leibler divergence)

  the “extra bits” of information needed to represent *p(y)* wrt *q(y).* Average over population to get MI (mutual information)

Example #1: measure *Height*

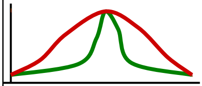
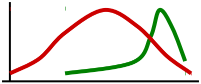
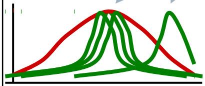
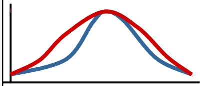
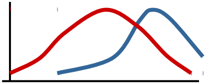
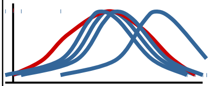
- Measure #1 (at doctor's office, ie. accurate)
- Measure #2 (via telescope, ie. inaccuate)

Individual Variability (+device errors)	Population Variability	Overall Distribution

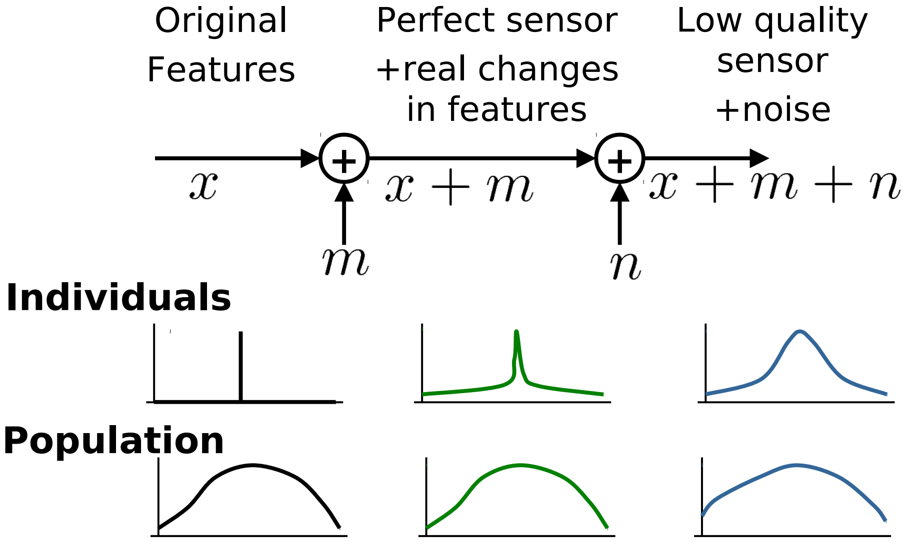


MI for height data

Most People
Are average

	Average (5½’ tall)	Tall (6½’ tall)	Mean MI
Measure #1: accurate	 MI = 0.23b	 MI = 2.7b	 MI = 0.3b
Measure #2: inaccurate	 MI = 0.05b	 MI = 1.1b	 MI = 0.07b

Quality Loss Model



Formula page ...

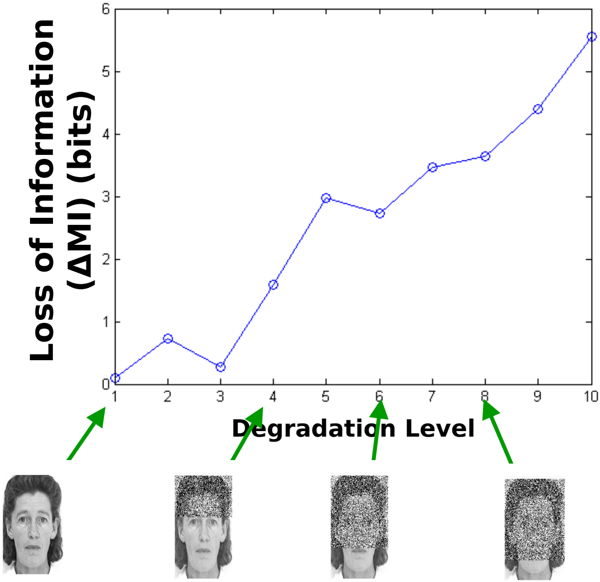
- KLD: $$D(p\|q) = \int p(\mathbf{y}) log_2 \frac{p(\mathbf{y})}{q(\mathbf{y})} d\mathbf{y}$$
- MI: $$= \underset{q}{E}\left[D(p\|q)\right]$$
- Gaussian Models:
$$= \frac{1}{2} log_2 |\boldsymbol{\Sigma}_q \boldsymbol{\Sigma}_p^{-1}| + tr\left(\boldsymbol{\Sigma}_p \boldsymbol{\Sigma}_q^{-1}\right)$$

When signal>noise

When noise>signal, *ignore*

- With noise model:
$$= \frac{1}{2} log_2 |\boldsymbol{\Sigma}_x (\boldsymbol{\Sigma}_m + \boldsymbol{\Sigma}_n)^{-1} + \mathbf{I}|$$

Results: loss of MI with addition of image noise



Comment: *Quality*

- *Quality* is a value laden term
- Can we tell users this?

Comment: *Quality*

- *Quality* is a value laden term
- Can we tell users this?