Massive Data Sets: Issues in Inference

Karen Kafadar, Indiana University

kkafadar@indiana.edu

`http://mypage.iu.edu/~kkafadar`

SAMSI Program on Massive Data Sets

Motivating Example #1: Beetles

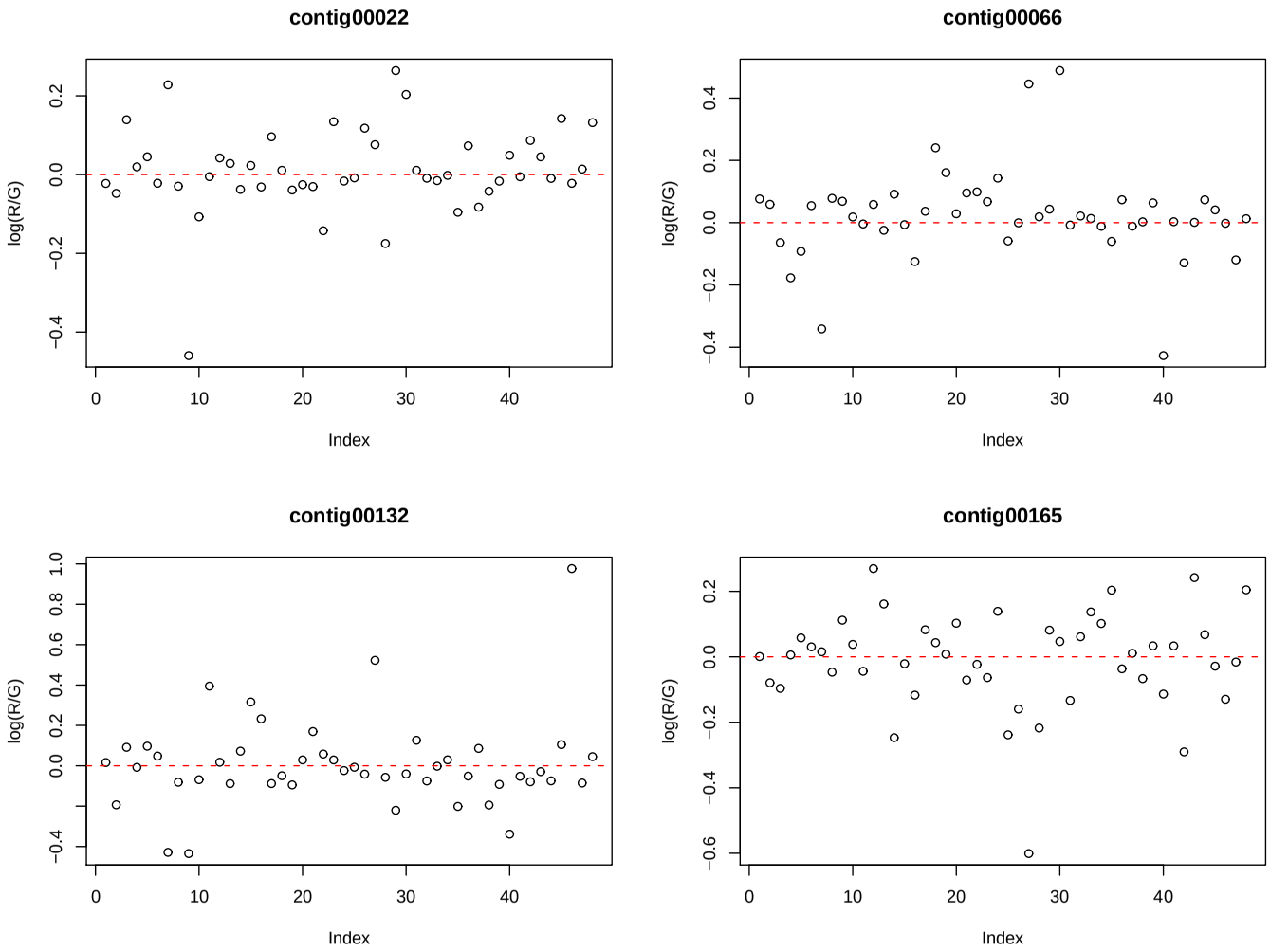
Experiment: $2^2 \times 4$ factorial design

1. **Nutrition**: High or Low
   ($\Rightarrow$ Large or Small beetles)
2. **Gender**: Female or Male
   (Females are Hornless; Large Males usually have Horns)*
3. **Tissues**: Abdomen, Head, Leg, Thorax

'Nuisance' factors: array, subarray, dye bias (red-green)

Challenge: Main effects & 2-factor interactions (2-fi)
either already known or not interesting

- Which contigs show nutrition differences in G×T conditions?
- Classify contigs: N only; N in AF only; A only; H only; A & H only; ... ($2^{15} = 32{,}768$ possible subsets!)


contig00022
log(R/G)
Index
contig00066
log(R/G)
Index
contig00132
log(R/G)
Index
contig00165
log(R/G)
Index

Motivating Example #2: HEP Experimental Data

- Colliding beams of electrons (SLAC) or protons (CERN) accelerated at very high energies (MeV/GeV/TeV)
- Collisions yield short-lived particles that decay into more short-lived particles in any one of 100,000 ways ("events")
- Most events well-characterized (particles, speeds, lifetimes)
- Others less well understood (e.g. those with B-mesons)

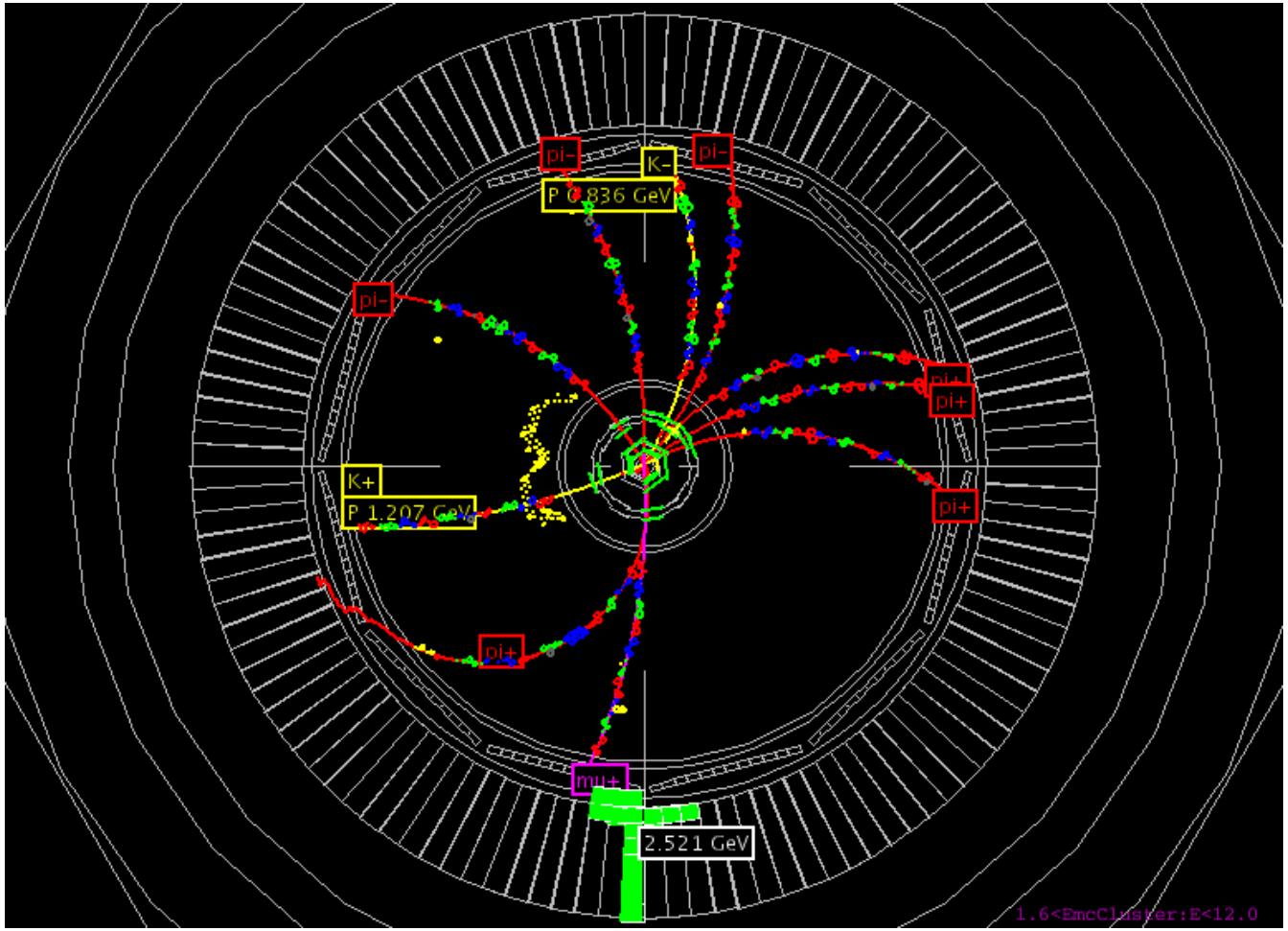
Goals:

- Find specific "target" events of interest amidst millions (per minute) of "uninteresting" events
- Even better: Did something "new" happen?

Compare "likelihoods" of possible events?

- Tens of thousands of possible "events"
- Likelihoods not simple! Ex:

$$f(q^2 = (p_\ell + p_\mu)^2, \cos\theta_\ell, \cos\theta_\nu, \chi) =$$
$$\frac{3G_F^2|V_c b|^2}{8(4\pi)^4} \frac{\rho_{D*}}{(q^2 - m_l^2)^2} q^2 B \times [((1 + \cos^2\theta_\ell) + \frac{m_\ell^2}{q^2}\sin^2\theta_\ell)\sin^2\theta_V(|H_+|^2 +$$
$$|H_-|^2) + 4(\sin^2\theta_\ell + \frac{m_\ell^2}{q^2}\cos^2\theta_\ell)\cos^2\theta_V|H_0|^2 - 2\cos\theta_\ell\sin^2\theta_V(|H_+|^2 -$$
$$|H_-|^2) - 2(1 - \frac{m_\ell^2}{q^2})\sin^2\theta_\ell\cos 2\chi\sin^2\theta_V Re(H_+ H_-^*) + (1 -$$
$$\frac{m_\ell^2}{q^2})\sin 2\theta_\ell\cos\chi\sin 2\theta_V Re(H_+ + H_-)H_0^* -$$
$$\sin\theta_\ell\cos\chi\sin 2\theta_V Re(H_+ - H_-)H_0^* + 4\frac{m_\ell^2}{q^2}\cos^2\theta_V|H_t|^2$$


pi-
K-
pi-
P 0.836 GeV
pi-
pi+
pi+
K+
P 1.207 GeV
pi+
pi+
mu+
2.521 GeV
1.6<EmcCluster:E<12.0

Issue #1: More data $\Rightarrow$ more hypotheses!

Challenge for *Confirmatory* inference:

- FDR $= \mathrm{E}(V/R|R > 0) \cdot \mathrm{P}\{R > 0\}$: control $\mathrm{E}(V/R)$
  better power, but: $\uparrow m =$ #hypotheses $\Rightarrow \uparrow R =$ #rejections
  $\Rightarrow \uparrow V =$ #false rejections
- k-FWER: lower power, but fixed V (control $\mathrm{P}\{V > k\}$)
- Hybrid: Generalized error rate (Meskaldji, Thiran, Morgenthaler)
  FDR until $R = k$, then k-FWER; i.e., control $\mathrm{E}(V/s(R))$
  where $s(R) = R$ if $R \leq k$; $s(R) = k$ thereafter

Issue #2: Sequential updates in parameter estimates

- More data than memory or disk space to store it
- Rate of increase in data collection >> rate of increase in processing speed
- Subsamples dictated by time of collection
- $\Rightarrow$ sequential updates in parameter estimates
- Do additional data reduce uncertainty? or indicate trends?
- Do we really have millions of data points from a steady-state population?
- How quickly can we distinguish between using new data to update parameter estimates versus to indicate process changes?

Exploratory data analysis leads to confirmatory inference:

- Which segments of the data are interesting?
  cf. JWT, 'Which part of the sample contains the information?'
  PNAS 1965:127–134
- Computer diagnostics ("cognostics"): Have the computer indicate "interesting" (outliers, correlations, trends, ...)
- Characterize "interestingness" of graphical display?
- Typical approach: start small, scale up
  (develop procedures on small data set; apply to larger amounts)
- Reverse? Start big: drill down if interesting, skip otherwise
- Inference on the whole data set, not on specific units
- Brillinger & Tukey (1984 *CWJWT II*):
  REAL DATA OFTEN FAIL to be Gaussian IN MANY WAYS.

References

Exploring Data Tables, Trends, and Shapes DC Hoaglin, F Mosteller, JW Tukey, eds., Wiley 1985.

Moczek A: "Evolution and Development: *Onthophagus* Beetles and the Evolutionary Developmental Genetics of Innovation, Allometry, Plasticity," *Ecology and Evolution of Dung Beetles* (ed. LW Simmons, TJ Ridsdill-Smith), Blackwell 2009.

Tukey, JW (many: See bibliography in Brillinger 2002, *Ann.Stat.*)