The EM algorithm

Analysis of a T cell frequency assay

Karl Broman

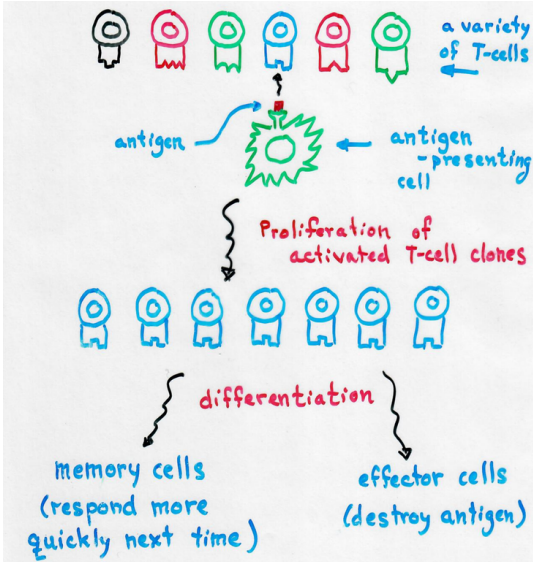
Biostatistics & Medical Informatics, UW–Madison

kbroman.org
github.com/kbroman
@kwbroman

Course web: kbroman.org/AdvData

Goal: Estimate the frequency of T-cells in a blood sample that respond to two test antigens.

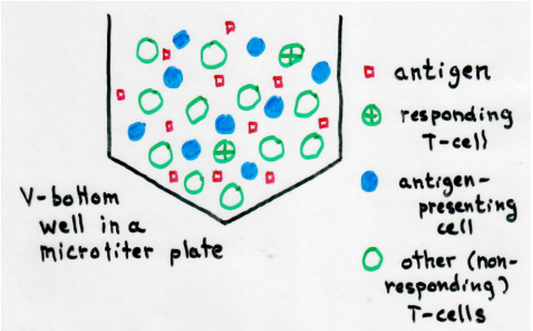
Real goal: Determine whether a vaccine causes an increase in the frequency of responding T-cells.



Broman K, Speed T, Tigges M (1996) J Immunol Meth 198:119-132 doi.org/b54v33

The assay

- Combine:
  - diluted blood cells + growth medium
  - antigen
  - ^{3}H-thymidine
- Replicating cells take up ^{3}H-thymidine.
- Extract the DNA and measure its radioactivity

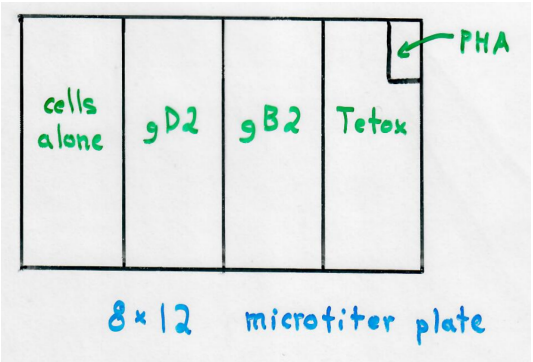


Usual approaches

- Use 3 wells with antigen and 3 wells without antigen, and take the ratio of the averages
- Limiting dilution assay
  - Several dilutions of cells
  - Many wells at each dilution

Our assay

Study a single plate or pair of plates at a single dilution.



Data

LDA 713, plates 1 and 2
11,400 cells per well

cells alone			gD2			gB2			Tetox		PHA
179	249	460	2133	2528	2700	2171	1663	6200	761	9864	12842
346	1540	306	8299	1886	3245	1699	2042	3374	183	7748	10331
117	249	1568	1174	4293	979	1222	1536	2406	6497	2492	6188
184	414	308	2801	2438	1776	2193	3211	1936	2492	5134	927
797	233	461	1076	1527	2866	2205	2278	2215	3725	3706	4050
305	348	480	3475	902	3654	2046	1285	1187	9899	5891	3646
1090	159	89	1472	90	3639	657	2393	1814	3330	4174	2389
280	571	329	4448	3643	881	3462	2118	1013	8793	4313	672
178	111	630	4699	5546	5182	3982	3104	2496	4275	2831	9727
244	593	259	5622	560	1073	1479	2978	4362	5017	5074	10706
261	964	167	2991	3390	3986	2321	2157	3278	8216	3579	3538
221	544	299	1838	4368	322	1022	1554	2980	2732	6177	5212
533	228	615	1938	4046	333	3253	5091	2843	200	1110	5063
818	98	160	1032	3269	4918	1778	3810	2372	6355	1869	2695
234	472	243	4143	3351	1118	530	1174	1881	3447	4491	2945
169	481	478	3237	1565	2211	2460	2715	4793	3029	6225	4679

Traditional analysis

- Split wells into +/– using a cutoff (e.g., mean + 3 SD of “cells alone” wells)

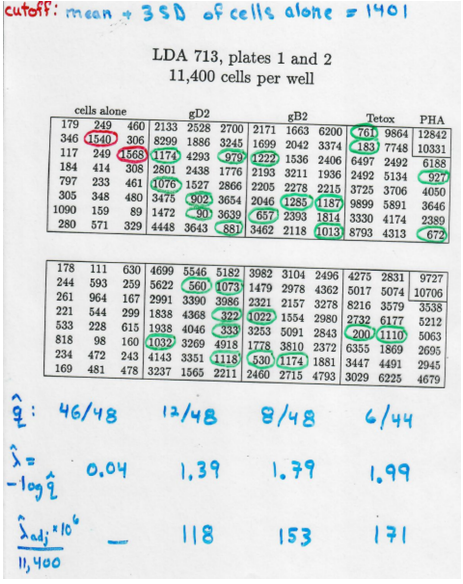
  positive = one or more responding cells
  negative = no responding cells

- Imagine that the number of responding cells in a well is Poisson(λ_i) for group *i*

  Pr(no responding cells) = $e^{-\lambda_i}$

  $$\hat{\lambda}_i = -\log\left(\frac{\text{\# negative wells}}{\text{\# wells}}\right)$$

Analysis



cutoff: mean + 3SD of cells alone = 1401

LDA 713, plates 1 and 2
11,400 cells per well

cells alone			gD2			gB2			Tetox		PHA
179	249	460	2133	2528	2700	2171	1663	6200	761	9864	12842
346	1540	306	8299	1886	3245	1699	2042	3374	183	7748	10331
117	249	1568	1174	4293	979	1222	1536	2406	6497	2492	6188
184	414	308	2801	2438	1776	2193	3211	1936	2492	5134	927
797	233	461	1076	1527	2866	2205	2278	2215	3725	3706	4050
305	348	480	3475	902	3654	2046	1285	1187	9899	5891	3646
1090	159	89	1472	90	3639	657	2393	1814	3330	4174	2389
280	571	329	4448	3643	881	3462	2118	1013	8793	4313	672

178	111	630	4699	5546	5182	3982	3104	2496	4275	2831	9727
244	593	259	5622	560	1073	1479	2978	4362	5017	5074	10706
261	964	167	2991	3390	3986	2321	2157	3278	8216	3579	3538
221	544	299	1838	4368	322	1022	1554	2980	2732	6177	5212
533	228	615	1938	4046	333	3253	5091	2843	200	1110	5063
818	98	160	1032	3269	4918	1778	3810	2372	6355	1869	2695
234	472	243	4143	3351	1118	530	1174	1881	3447	4491	2945
169	481	478	3237	1565	2211	2460	2715	4793	3029	6225	4679

$\hat{q}$: 46/48 12/48 8/48 6/44

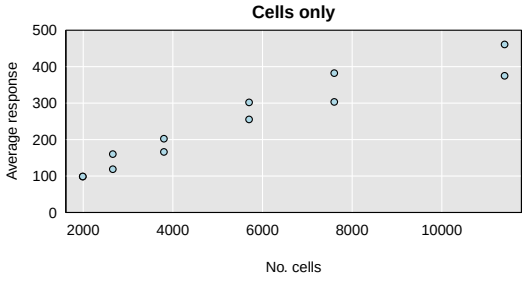
$\hat{\lambda} = -\log \hat{q}$: 0.04 1.39 1.79 1.99

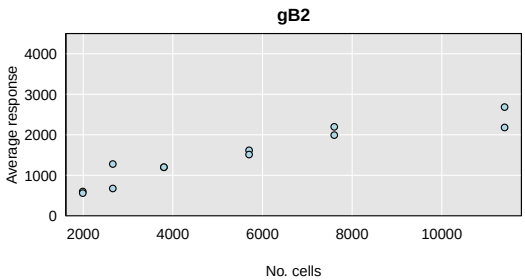
$\frac{\hat{\lambda}_{adj} \times 10^6}{11,400}$: — 118 153 171

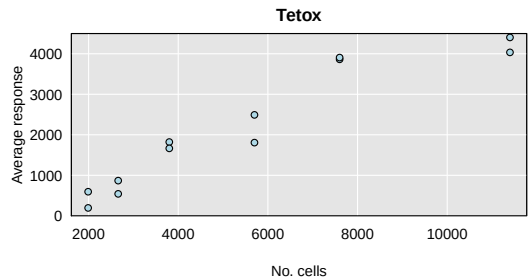
Problems

- Hard to choose cutoff
- Potential loss of information

Response vs no. cells







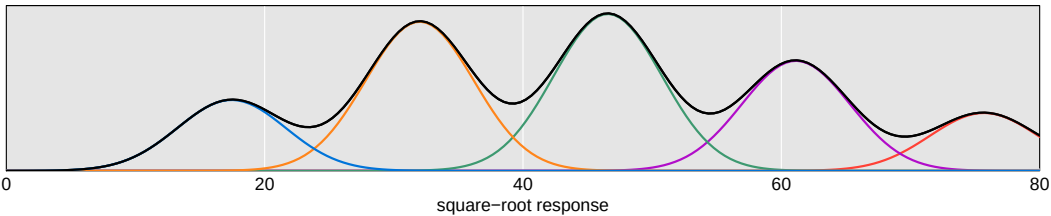
Model

k_{ij} = Number of responding cells (unobserved)
y_{ij} = square-root of response

Assume $k_{ij} \sim \text{Poisson}(\lambda_i)$

$$y_{ij} \mid k_{ij} \sim \text{Normal}(a + bk_{ij}, \sigma)$$

(k_{ij}, y_{ij}) mutually independent



log Likelihood

$$
\begin{aligned}
l(\boldsymbol{\lambda}, a, b, \sigma) &= \sum_{i,j} \log \Pr(y_{ij}|\lambda_i, a, b, \sigma) \\
&= \sum_{i,j} \log \left[\sum_k \Pr(k|\lambda_i) \Pr(y_{ij}|k, a, b, \sigma) \right] \\
&= \sum_{i,j} \log \left[\sum_k \left(\frac{e^{-\lambda_i} \lambda_i^k}{k!} \right) \phi \left(\frac{y_{ij} - a - bk}{\sigma} \right) \right]
\end{aligned}
$$

EM algorithm

- Iterative algorithm useful when there is missing data that if observed would make things easy
- Dempster et al. (1977) JRSS-B 39:1-22 doi.org/gfxzrv
- Start with some initial estimates
- E-step: expected value of missing data given current estimates
- M-step: MLEs replacing missing data with their expected values
- Advantages
  - often easy to code
  - usually super stable
  - log likelihood is non-decreasing

Normal/Poisson model

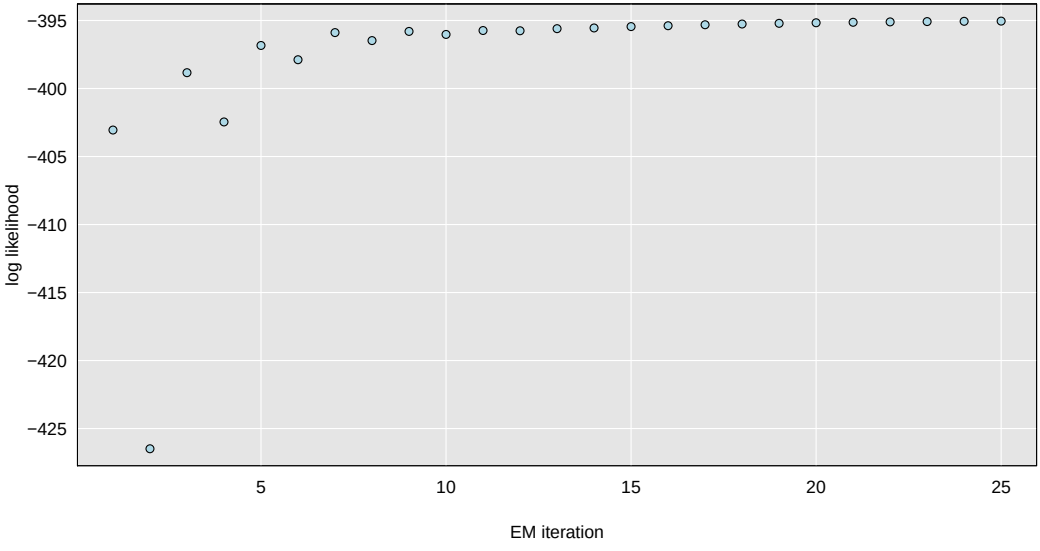
E-step:

$$\begin{aligned}\Pr(k=s|y,\lambda,a,b,\sigma) &= \frac{\Pr(k=s|\lambda)\Pr(y|k=s,a,b,\sigma)}{\sum_s \Pr(k=s|\lambda)\Pr(y|k=s,a,b,\sigma)} \\ &= \frac{\left(\frac{e^{-\lambda}\lambda^s}{s!}\right)\phi\left(\frac{y-a-bs}{\sigma}\right)}{\sum_s \left(\frac{e^{-\lambda}\lambda^s}{s!}\right)\phi\left(\frac{y-a-bs}{\sigma}\right)}\end{aligned}$$

$$\mathrm{E}(k|y,\lambda,a,b,\sigma) = \frac{\sum_s s\left(\frac{e^{-\lambda}\lambda^s}{s!}\right)\phi\left(\frac{y-a-bs}{\sigma}\right)}{\sum_s \left(\frac{e^{-\lambda}\lambda^s}{s!}\right)\phi\left(\frac{y-a-bs}{\sigma}\right)}$$

M-step: Regress y on $\mathrm{E}(k|y)$

Oops, that didn't work



EM algorithm, more formally

- Calculate expected complete-data log likelihood, given observed data and observed parameters, and then maximize that.

$$l^{(s)}(\theta) = \mathrm{E}\{\log f(y, k|\theta)|y, \hat{\theta}^{(s)}\}$$

- In practice, it's usually a linear combination of the sufficient statistics, so you focus on those.
- Here, we need not just $\sum k$ and $\sum ky$, but also $\sum k^2$.

EM algorithm, again

E step: we also need

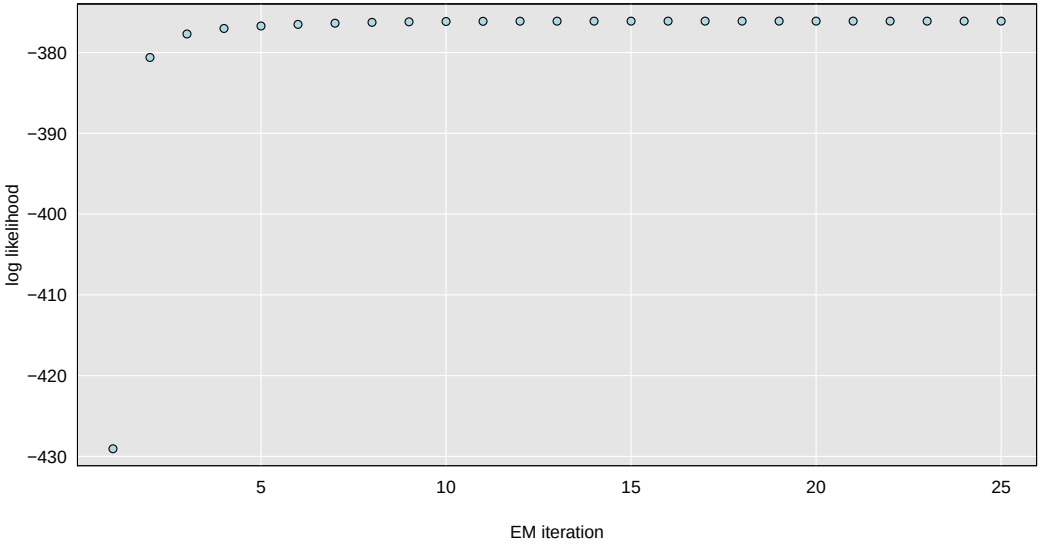
$$\mathrm{E}(k^2|y,\lambda,a,b,\sigma) = \frac{\sum_s s^2 \left(\frac{e^{-\lambda}\lambda^s}{s!}\right)\phi\left(\frac{y-a-bs}{\sigma}\right)}{\sum_s \left(\frac{e^{-\lambda}\lambda^s}{s!}\right)\phi\left(\frac{y-a-bs}{\sigma}\right)}$$

M step: we want $\hat{\beta} = (X'X)^{-1}(X'y)$

where $(X'X)$ is like $\begin{pmatrix} n & \sum k \\ \sum k & \sum k^2 \end{pmatrix}$

and $(X'y)$ is like $\begin{pmatrix} \sum y \\ \sum ky \end{pmatrix}$

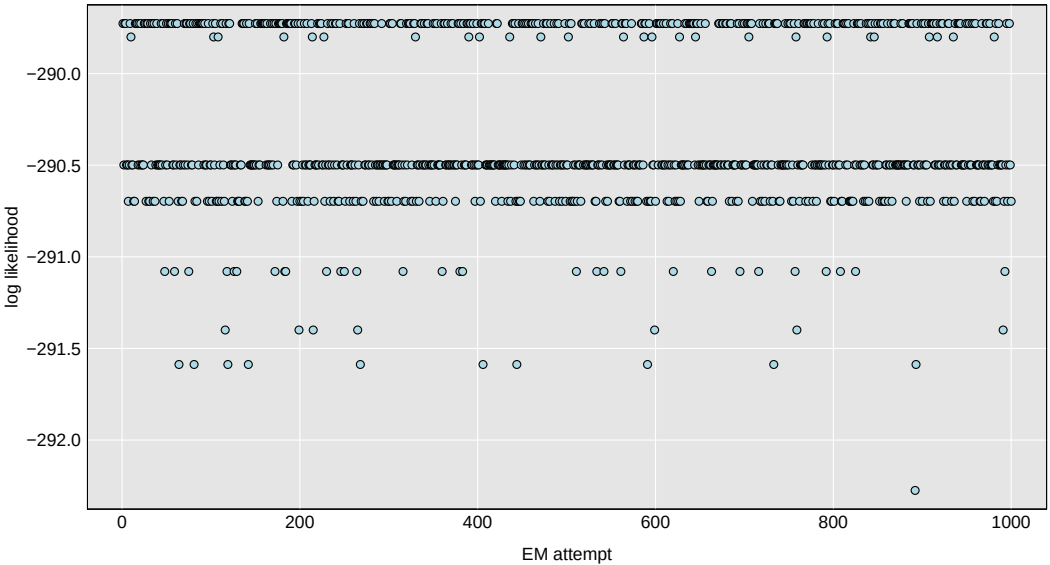
Ah, that's better



Difficulties

- Starting values
- Multiple modes

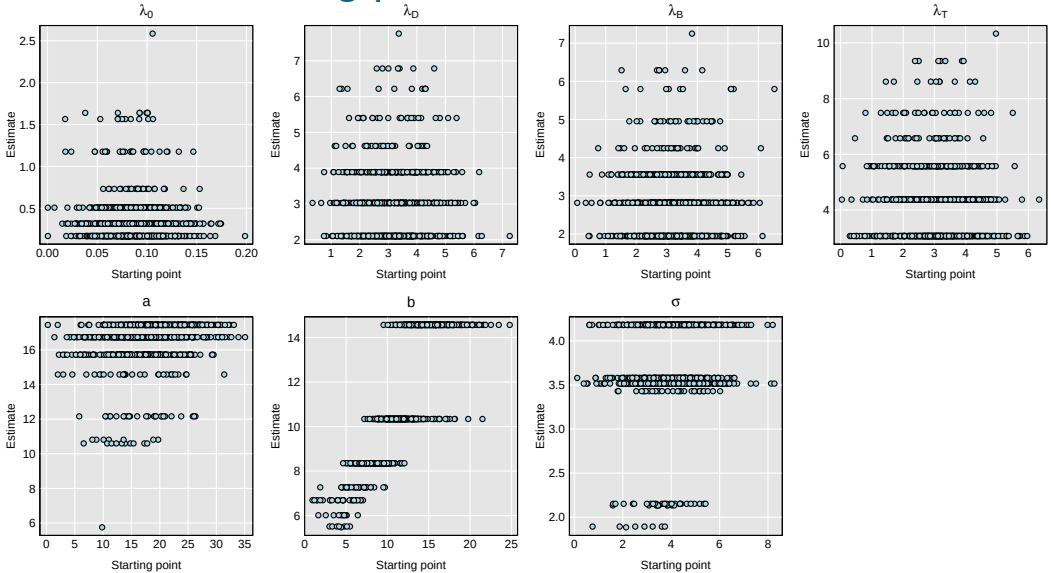
Multiple modes



Multiple modes

	λ_0	λ_D	λ_B	λ_T	a	b	σ	log lik	no. hits
1	0.32	3.03	2.82	4.37	16.73	10.34	3.52	–289.73	331
2	1.18	5.40	4.95	7.49	12.16	6.69	2.15	–289.80	26
3	0.17	2.10	1.95	3.07	17.44	14.56	4.18	–290.50	415
4	0.51	3.89	3.56	5.58	15.72	8.35	3.58	–290.70	180
5	0.73	4.62	4.25	6.58	14.58	7.27	3.43	–291.08	30
6	1.64	6.79	6.29	9.35	10.81	5.51	1.89	–291.40	7
7	1.57	6.22	5.80	8.61	10.60	6.02	2.13	–291.59	10
8	2.59	7.76	7.25	10.34	5.75	5.47	1.88	–292.27	1

Estimate vs. starting point



Principles

- Start with an understanding of the problem and data
- Think about a model for the data-generating process

Lessons

- The EM algorithm is really useful
- Use the log likelihood as a diagnostic when implementing an EM algorithm

Impact

- I'm pretty sure that the vaccine they were working on didn't work well.
- R package `npem`, but I never put it on CRAN, and no one has ever asked me about it.
- Our paper has like 9 citations: no one has ever really used the method.

Further things

- Standard errors should always be required.
  - But usually painful to obtain
  - We used the SEM algorithm of Meng and Rubin (1991) doi.org/10.1080/01621459.1991.10475130
- Could more formally investigate the appropriate transformation
  - See Box and Cox (1964) doi.org/10.1111/j.2517-6161.1964.tb00553.x
  - Box-Cox transformation is $g(y) = (y^c - 1)/c$ for $c \neq 0$ and $= \log y$ for $c = 0$
  - Key issue is change-of-variables in the density; as a result you add $\sum_{ij}(c-1)\log y_{ij}$ to the log likelihood