Integration of diverse data types: an illustration with genomic data on breast cancer

Darlene R. Goldstein

Statistics Seminar, University of Vienna
7 December 2006
Integration of diverse data types: an illustration with genomic data on breast cancer

Darlene R. Goldstein

Outline

1 Introduction

2 Combination methods
   - Popular methods
   - Combining z-scores
   - Comparison

3 Example
   - Methodology
   - EDA
   - Results

4 This is joint work...
Microarray studies have typically been rather small (usually due to expense)

Several studies carried out on similar questions

Public data repositories should facilitate combining data

Result: will have the ability to use all available information to ‘find the genes’
Studies are very *different* wrt:

- inclusion/exclusion criteria
- measured outcomes
- probe sequences (genes) measured
- platforms, which cannot be transformed to comparable scales (Affymetrix, cDNA, Agilent, PCR, ...)
- gene expression quantification (normalization, summarization)

**Result:** ‘*finding the genes*’ still not trivial!
Integration of diverse data types: an illustration with genomic data on breast cancer

Darlene R. Goldstein

Outline
Introduction
Combination methods
Popular methods
Combining z-scores
Comparison
Example
Methodology
EDA
Results
References

This is joint work...

Analysis spectrum

The possibilities for combining information across studies can be viewed as occurring along a spectrum of levels of analysis, moving roughly in order of decreasing information content:

1. pooling raw data
2. pooling adjusted data (e.g. covariate adjustment, normalized signal intensities)
3. combining parameter estimates
4. combining effect sizes or test statistics
5. combining transformed p-values
6. combining statistic ranks
7. combining decisions (e.g. via intersecting Venn diagrams)
What/how to combine

- Avoid pooling data prior to analysis: make comparisons within study
  - Compare like with like
  - Avoid Simpson’s paradox
- Consider analysis goals: which deviations from the null you want to detect
  - Genes doing the same thing across studies (e.g. genes associated with increased survival)
  - Genes doing different things across studies (e.g. platform comparison)
- Use available information efficiently
  - Increase power
Popular methods of combination

- Combine *decisions*: ‘Venn diagram’
- Combine *parameter estimates*:
  - Fixed effects meta-analysis (FEMA)
  - Random effects meta-analysis (REMA)
- Combine *p-values*: Fisher $p$-value combination
- Combine *test statistics* (or $p$-values): Combining z-scores
Admissibility and power

- **Defn:** A test is *admissible* if and only if no other test dominates it; that is, if there is not another test which is never worse and is sometimes better.

- Extensive literature (1940s – 1960s) on methods of combining tests
  - **Methods:** Fisher [5], Liptak [10], *etc.*
  - **Admissibility:** Birnbaum [3], Marden [12]
  - **Power:** Bhattacharyya [1], Koziol and Perlman [9]
Integration of diverse data types: an illustration with genomic data on breast cancer

Darlene R. Goldstein

Outline
Introduction
Combination methods
Popular methods
Combining z-scores
Comparison
Example
Methodology
EDA
Results
References

This is joint work...

**Outline**

- Introduction
- Combination methods
- Popular methods
- Combining z-scores
- Comparison
- Example
- Methodology
- EDA
- Results
- References
- This is joint work...

**Venn diagram**

- Selects genes significant *in both (all) studies*
- This rule seems intuitive for biologists
- **Problem**: what does ‘reproducible’ mean?
- At the top are signal (true +) and noise (false +)
- *Not admissible for exponential families* [3]
Combining estimates: heterogeneity analysis

Before combining estimates from different studies, verify that they are *homogeneous*, i.e. do they all seem to be estimating the same underlying population parameter

- Graphical methods (e.g. forest plots) are useful when there are several *single outcome studies* to be combined
- For a *microarray study*, need one plot for each gene
- $\Rightarrow$ Use numerical assessment
Test of homogeneity

- Cochran test for homogeneity tests for equality of estimates against alternative that at least one is different
- Test statistic \( Q = \sum_{i=1}^{k} w_i (\hat{\beta}_i - \bar{\beta})^2 \), where \( k \) is number of studies
- \( \hat{\beta}_i \) estimates the treatment effect
- \( w_i \) is the weight for study \( i \) (most commonly taken as the reciprocal of the variance of the outcome estimate)
- \( \bar{\beta} = \sum_i w_i \hat{\beta}_i / \sum_i w_i \) is the weighted average treatment effect
- Under the null, \( Q \sim \chi^2_{k-1} \)
Results of homogeneity test

- If null *is not rejected*, differences between studies are assumed to be due to *chance variation*

- $\Rightarrow$ Can combine estimates via a *fixed effects model* (FEMA)

- If null *is rejected*, then assume *extra variability* between studies (beyond chance error)

- Combination via a *random effects model* (REMA) is typically favored

- One homogeneity test *per gene*
FEMA vs. REMA

- FEMA and REMA: estimates are weighted averages, weights inversely proportional to variance.
- FEMA assumes no heterogeneity between studies.
- If results are heterogeneous, then there is assumed to be a distribution of effect sizes.
- REMA assumes individual studies may be estimating different effects.
- Weights take into account additional variability \( \tau^2 \) between studies: \( w_i^* = \frac{1}{(1/w_i) + \tau^2} \).
- When \( \tau = 0 \), REMA reduces to FEMA.
- For exponential families, under homogeneity REMA not admissible.
Fisher combined $p$-values

- Several possibilities for combining (transformed) $p$-values
- One commonly used method is due to Fisher [5]
- Studies have shown it performs pretty well under a variety of scenarios [1, 9]
- Fisher summary test statistic $S = -2 \sum_{i=1}^{k} \log(p_i)$
- Theoretical null distribution of $S$ should be $\chi^2_{2k}$
- Could also obtain a $p$-value for $S$ by resampling (Rhodes et al. [16])
Method of combining z-scores

- Can use when all test statistics have a *normal distribution*
- Can also be considered as part of class of methods based on *p*-value transformation
  - *BUT*: not generally efficient if have original test statistics and these are not normal
  - In particular, *should not use* to combine $\chi^2$ statistics
- *Admissible* for exponential families; can be *optimal*
- Weighted or unweighted (*i.e.* equal weights) versions
- Simplest (unweighted) case: Combined $Z = \sum Z_i / \sqrt{k}$ has a standard normal distribution under the null
Integration of diverse data types: an illustration with genomic data on breast cancer

Darlene R. Goldstein

Outline
Introduction
Combination methods
Popular methods
Combining z-scores
Comparison
Example
Methodology
EDA
Results
References

This is joint work...
Integration of diverse data types: an illustration with genomic data on breast cancer

Darlene R. Goldstein

Outline
Introduction
Combination methods
Popular methods
Combining z-scores
Comparison
Example
Methodology
EDA
Results
References

This is joint work...
Integration of diverse data types: an illustration with genomic data on breast cancer

Darlene R. Goldstein

Outline
Introduction
Combination methods
Popular methods
Combining z-scores
Comparison
Example
Methodology
EDA
Results
References

This is joint work...

Contours overplotted – zoom-in
Integration of diverse data types: an illustration with genomic data on breast cancer

Darlene R. Goldstein

Outline
Introduction
Combination methods
Popular methods
Combining z-scores
Comparison
Example
Methodology
EDA
Results
References

This is joint work...
Integration of diverse data types: an illustration with genomic data on breast cancer

Darlene R. Goldstein

Outline
Introduction
Combination methods
Popular methods
Combining z-scores
Comparison
Example
Methodology
EDA
Results
References

This is joint work...
Methodology for genome-scale survival data

- Need *raw (or suitably processed) data*, not just $p$-value from previous study
- Response variable: metastasis-free survival, no covariates
- Multiple probes of the same genes *made unique* by choosing the most variable
- Do *NOT* need to consider only the common probes: *missing data readily accommodated* in this framework
- For each gene fit a separate Cox model
- Can do $p$-value adjustment
Integration of diverse data types: an illustration with genomic data on breast cancer

Darlene R. Goldstein

Outline
Introduction
Combination methods
Popular methods
Combining z-scores
Comparison
Example
Methodology
EDA
Results
References
This is joint work...

Table: Publicly available breast cancer survival gene expression data

<table>
<thead>
<tr>
<th>Dataset symbol</th>
<th>No. of arrays</th>
<th>Institution</th>
<th>Reference</th>
<th>Platform</th>
<th>Data source</th>
<th>No. of geneID</th>
</tr>
</thead>
<tbody>
<tr>
<td>NKI</td>
<td>337</td>
<td>Nederlands Kanker Inst.</td>
<td>[20, 21]</td>
<td>Agilent</td>
<td>author’s website</td>
<td>13129</td>
</tr>
<tr>
<td>UCSF</td>
<td>161+8</td>
<td>UC San Francisco</td>
<td>[8]</td>
<td>cDNA</td>
<td>author’s website</td>
<td>6168</td>
</tr>
<tr>
<td>UNC</td>
<td>143+10</td>
<td>Univ. of Carolina</td>
<td>[6, 14, 23]</td>
<td>Agilent HuA1</td>
<td>author’s website</td>
<td>13785</td>
</tr>
<tr>
<td>STNO</td>
<td>115+7</td>
<td>Stanford-Norway</td>
<td>[17]</td>
<td>cDNA</td>
<td>author's website</td>
<td>5598</td>
</tr>
<tr>
<td>JRH1</td>
<td>99</td>
<td>John Radcliffe Hosp.</td>
<td>[18]</td>
<td>cDNA</td>
<td>journal's website</td>
<td>4109</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1866 = 1841 tumors</td>
<td>No. of the union of all geneID’s:</td>
<td>16776</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>+ 25 non-malignant breast tissues</td>
<td>No. of geneID’s common to genomic arrays:</td>
<td>1951</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Integration of diverse data types: an illustration with genomic data on breast cancer

Darlene R. Goldstein

Outline
Introduction
Combination methods
Popular methods
Combining z-scores
Comparison
Example
Methodology
EDA
Results
References

This is joint work...
One set vs. z-score combination of the rest
Integration of diverse data types: an illustration with genomic data on breast cancer

Darlene R. Goldstein

Outline
Introduction
Combination methods
Popular methods
Combining z-scores
Comparison
Example
Methodology
EDA
Results
References

This is joint work...
Integration of diverse data types: an illustration with genomic data on breast cancer

Darlene R. Goldstein

Outline
Introduction
Combination methods
Popular methods
Combining z-scores
Comparison
Example
Methodology
EDA
Results
References

This is joint work...

Distribution of combined z

Histogram of Combined z

Normal Q-Q Plot
### Preliminary results – Top 25 genes

<table>
<thead>
<tr>
<th>symbol</th>
<th>Z</th>
<th>NKI</th>
<th>DUKE</th>
<th>UCSF</th>
<th>STNO</th>
<th>JRH1</th>
<th>MGH</th>
<th>UPP</th>
<th>STOCK</th>
<th>EMC</th>
<th>UNC</th>
<th>JRH2</th>
</tr>
</thead>
<tbody>
<tr>
<td>*AURKA</td>
<td>9.67</td>
<td>6.33</td>
<td>1.09</td>
<td>2.33</td>
<td>3.05</td>
<td>1.83</td>
<td>1.56</td>
<td>3.38</td>
<td>3.28</td>
<td>4.52</td>
<td>3.55</td>
<td>1.16</td>
</tr>
<tr>
<td>*CCNB2</td>
<td>9.17</td>
<td>5.56</td>
<td>3.95</td>
<td>2.33</td>
<td>3.05</td>
<td>1.83</td>
<td>1.56</td>
<td>3.38</td>
<td>3.28</td>
<td>4.52</td>
<td>3.55</td>
<td>1.16</td>
</tr>
<tr>
<td>*MELK</td>
<td>8.82</td>
<td>4.51</td>
<td>4.10</td>
<td>2.77</td>
<td>3.64</td>
<td>1.83</td>
<td>3.38</td>
<td>3.28</td>
<td>4.52</td>
<td>3.55</td>
<td>1.16</td>
<td>0.66</td>
</tr>
<tr>
<td>*MYBL2</td>
<td>8.79</td>
<td>4.94</td>
<td>3.20</td>
<td>0.56</td>
<td>3.38</td>
<td>2.73</td>
<td>1.23</td>
<td>4.37</td>
<td>3.02</td>
<td>2.61</td>
<td>3.01</td>
<td>0.11</td>
</tr>
<tr>
<td>*BUB1</td>
<td>8.70</td>
<td>4.43</td>
<td>1.15</td>
<td>1.24</td>
<td>3.65</td>
<td>2.63</td>
<td>0.79</td>
<td>2.88</td>
<td>4.24</td>
<td>3.37</td>
<td>2.78</td>
<td>1.69</td>
</tr>
<tr>
<td>*AURKB</td>
<td>8.47</td>
<td>5.01</td>
<td>4.12</td>
<td>−0.12</td>
<td>3.56</td>
<td>2.09</td>
<td>3.44</td>
<td>3.71</td>
<td>1.15</td>
<td>3.00</td>
<td>0.84</td>
<td>1.56</td>
</tr>
<tr>
<td>*RACGAP1</td>
<td>8.47</td>
<td>5.48</td>
<td>0.48</td>
<td>4.24</td>
<td>3.76</td>
<td>4.91</td>
<td>1.99</td>
<td>1.56</td>
<td>3.68</td>
<td>4.49</td>
<td>2.19</td>
<td>0.94</td>
</tr>
<tr>
<td>CENPA</td>
<td>8.40</td>
<td>5.75</td>
<td>2.43</td>
<td>2.35</td>
<td>3.41</td>
<td>3.70</td>
<td>2.84</td>
<td>2.19</td>
<td>3.68</td>
<td>4.49</td>
<td>2.71</td>
<td>1.15</td>
</tr>
<tr>
<td>DDX39</td>
<td>8.35</td>
<td>5.49</td>
<td>3.29</td>
<td>1.09</td>
<td>3.53</td>
<td>4.49</td>
<td>2.71</td>
<td>1.15</td>
<td>3.68</td>
<td>4.49</td>
<td>2.71</td>
<td>1.15</td>
</tr>
<tr>
<td>*UBE2C</td>
<td>8.32</td>
<td>5.63</td>
<td>3.56</td>
<td>1.15</td>
<td>1.62</td>
<td>0.66</td>
<td>3.68</td>
<td>3.48</td>
<td>3.43</td>
<td>1.70</td>
<td>0.94</td>
<td>0.86</td>
</tr>
<tr>
<td>*FEN1</td>
<td>8.15</td>
<td>5.31</td>
<td>1.43</td>
<td>0.81</td>
<td>1.92</td>
<td>1.99</td>
<td>4.49</td>
<td>3.28</td>
<td>2.47</td>
<td>3.05</td>
<td>1.00</td>
<td>0.45</td>
</tr>
<tr>
<td>DLG7</td>
<td>8.13</td>
<td>4.31</td>
<td>2.64</td>
<td>0.88</td>
<td>3.14</td>
<td>1.27</td>
<td>3.18</td>
<td>3.96</td>
<td>3.75</td>
<td>1.81</td>
<td>0.77</td>
<td>0.74</td>
</tr>
<tr>
<td>p762E1312</td>
<td>8.12</td>
<td>6.10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.68</td>
<td>4.00</td>
<td>2.52</td>
</tr>
<tr>
<td>*TRIP13</td>
<td>8.02</td>
<td>4.97</td>
<td>3.11</td>
<td>0.53</td>
<td>2.90</td>
<td>0.71</td>
<td>4.33</td>
<td>3.79</td>
<td>1.34</td>
<td>2.68</td>
<td>1.01</td>
<td>0.45</td>
</tr>
<tr>
<td>*GPI</td>
<td>7.97</td>
<td>4.12</td>
<td>3.16</td>
<td>0.75</td>
<td>3.77</td>
<td>1.76</td>
<td>1.75</td>
<td>3.61</td>
<td>3.34</td>
<td>0.16</td>
<td>3.58</td>
<td>0.45</td>
</tr>
<tr>
<td>CCNE2</td>
<td>7.96</td>
<td>5.31</td>
<td>2.90</td>
<td></td>
<td>2.46</td>
<td>3.01</td>
<td>4.27</td>
<td>1.55</td>
<td>1.58</td>
<td>0.94</td>
<td>0.58</td>
<td>0.58</td>
</tr>
<tr>
<td>PRC1</td>
<td>7.96</td>
<td>5.80</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.35</td>
<td>3.72</td>
<td>3.50</td>
<td>2.16</td>
</tr>
<tr>
<td>CCNB1</td>
<td>7.84</td>
<td>4.76</td>
<td>3.23</td>
<td>−1.33</td>
<td>2.41</td>
<td>0.51</td>
<td>4.30</td>
<td>3.71</td>
<td>3.12</td>
<td>1.81</td>
<td>2.28</td>
<td>1.58</td>
</tr>
<tr>
<td>SEC61G</td>
<td>7.83</td>
<td>4.61</td>
<td>1.47</td>
<td>1.37</td>
<td>3.74</td>
<td>2.13</td>
<td>2.72</td>
<td>3.48</td>
<td>2.84</td>
<td>0.57</td>
<td>0.87</td>
<td>0.87</td>
</tr>
<tr>
<td>CENPF</td>
<td>7.83</td>
<td>3.44</td>
<td>1.53</td>
<td>1.41</td>
<td>2.93</td>
<td>1.93</td>
<td>2.90</td>
<td>4.37</td>
<td>2.65</td>
<td>2.13</td>
<td>1.46</td>
<td>1.46</td>
</tr>
<tr>
<td>GINS2</td>
<td>7.79</td>
<td>5.21</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.16</td>
<td>4.00</td>
<td>3.36</td>
<td>0.64</td>
</tr>
<tr>
<td>ZWINT</td>
<td>7.75</td>
<td>4.59</td>
<td>1.80</td>
<td>0.52</td>
<td>1.32</td>
<td>4.63</td>
<td>3.28</td>
<td>2.95</td>
<td>2.50</td>
<td>1.65</td>
<td>0.15</td>
<td>1.70</td>
</tr>
<tr>
<td>SPAG5</td>
<td>7.74</td>
<td>5.02</td>
<td>2.48</td>
<td>0.71</td>
<td>0.91</td>
<td>4.20</td>
<td>3.73</td>
<td>2.78</td>
<td>3.24</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>KIF23</td>
<td>7.69</td>
<td>3.53</td>
<td>2.02</td>
<td>−0.26</td>
<td>4.06</td>
<td>2.49</td>
<td>0.04</td>
<td>3.32</td>
<td>4.02</td>
<td>2.27</td>
<td>2.85</td>
<td>1.17</td>
</tr>
<tr>
<td>UBE2S</td>
<td>7.64</td>
<td>4.45</td>
<td>2.62</td>
<td>1.06</td>
<td>1.66</td>
<td>0.59</td>
<td>4.42</td>
<td>4.22</td>
<td>2.36</td>
<td>0.99</td>
<td>1.77</td>
<td>0.99</td>
</tr>
</tbody>
</table>
Combined Z compared to Fisher $p$
References


Integration of diverse data types: an illustration with genomic data on breast cancer

Darlene R. Goldstein

Outline
Introduction
Combination methods
Popular methods
Combining z-scores
Comparison
Example
Methodology
EDA
Results
References

This is joint work...


Integration of diverse data types: an illustration with genomic data on breast cancer

Darlene R. Goldstein

Outline
Introduction
Combination methods
Popular methods
Combining z-scores
Comparison
Example
Methodology
EDA
Results
References

Acknowledgments

Institut Suisse de recherche expérimentale sur le cancer and Swiss Institute of Bioinformatics

- Asa Wirapati
- Mauro Delorenzi

Institut de mathématiques, École Polytechnique Fédérale de Lausanne

- Tom Mountford