AUTOMATIC PHENOTYPING: ESTIMATING FRUIT NUMBERS BY CLUSTERING REPEATED, INCOMPLETE OBSERVATIONS

Chris Glasbey, Yu Song, Graham Horgan (BioSS, UK)
Gerie van der Heijden, Gerrit Polder (Biometris, The Netherlands)
Pepper plants in greenhouse (EU-FP7 project SPICY)

⇒

SPYSEE cameras on trolley

colour image every 5cm
How many pepper fruits?
Fruits found automatically by image processing:
Locations of fruits

![Graph showing the locations of fruits with coordinates. The x-axis ranges from 100 to 400, and the y-axis ranges from 300 to 700. The graph shows several points marked with asterisks.]

* image 0
Locations of fruits

image 0 and 1
Locations of fruits

Repeated, incomplete observations: Cluster to estimate number of fruits
Model for data from single experimental plot:

\[
\begin{pmatrix}
    x_{ij} \\
    y_{ij}
\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}
    \alpha_{k(ij)} + i\gamma_{k(ij)} \\
    \beta_{k(ij)}
\end{pmatrix}, \begin{pmatrix}
    \sigma_x^2 & 0 \\
    0 & \sigma_y^2
\end{pmatrix}\right)
\]

for \( k = 1, \ldots, K \)

\((x_{ij}, y_{ij})\): column, row coordinates of \( j \)th fruit in \( i \)th image

\( k(ij) \): correct fruit label of observation \((i, j)\)

\((\alpha_k, \beta_k)\): true column, row coordinates of fruit \( k \) in image 0

\( \gamma_k \): true shift in column coordinate between consecutive images

\((\sigma_x^2, \sigma_y^2)\): error variances (common to all 400 plots)
Model for data from single experimental plot:

\[
\begin{pmatrix}
    x_{ij} \\
    y_{ij}
\end{pmatrix}
\sim N\left( \begin{bmatrix}
    \alpha_{k(ij)} + i\gamma_{k(ij)} \\
    \beta_{k(ij)}
\end{bmatrix}, \begin{bmatrix}
    \sigma^2_x & 0 \\
    0 & \sigma^2_y
\end{bmatrix} \right)
\]

for \( k = 1, \ldots, K \)

How to estimate \( K \)?

1. estimate \( (\sigma^2_x, \sigma^2_y) \) using data from all plots
2. estimate \( K \) and other parameters separately for each plot
Consider all pairs of observations \( \{(i_1, j_1), (i_2, j_2)\} \) st \( i_1 < i_2 \)

\[ \Delta = y_{i_2,j_2} - y_{i_1,j_1} \]

\[ \hat{\gamma} = \frac{x_{i_2,j_2} - x_{i_1,j_1}}{i_2 - i_1} \]

Model: \( y_{i,j} \sim N(\beta_{k(i,j)}, \sigma_y^2) \)

So \( (\Delta \mid |\Delta| \leq \Delta_M) \sim \begin{cases} N(0, 2\sigma_y^2) & \text{if } k(i_1, j_1) = k(i_2, j_2) \\ U(-\Delta_M, \Delta_M) & \text{otherwise} \end{cases} \)
All triplets of observations $\{(i_1, j_1), (i_2, j_2), (i_3, j_3)\}$ st $i_1 < i_2 < i_3$

$$(\hat{\alpha}, \hat{\gamma}) = \arg \min_{(\alpha, \gamma)} \sum_{l=1}^{3} (x_{i_l,j_l} - \alpha - i_l \gamma)^2$$

$$S_x^2 = \sum_{l=1}^{3} (x_{i_l,j_l} - \hat{\alpha} - i_l \hat{\gamma})^2$$

Model: $x_{ij} \sim N(\alpha_{k(ij)} + i\gamma_{k(ij)}, \sigma_x^2)$

So $(S_x \mid S_x \leq S_M) \sim \begin{cases} N^+(0, \sigma_x^2) & \text{if } k(i_1, j_1) = k(i_2, j_2) = k(i_3, j_3) \\ U(0, S_M) & \text{otherwise} \end{cases}$
Algorithm

1. Initialise cluster size $n \rightarrow \#\text{images}$

2. Find cluster $\{(i_1, j_1), (i_2, j_2), \ldots, (i_n, j_n)\}$ st $i_1 < i_2 < \ldots < i_n$ and $50 \leq \hat{\gamma} \leq 130$, which minimises

$$S^2 = \frac{1}{\hat{\sigma}_x^2} \sum_{l=1}^{n} (x_{i_l,j_l} - \hat{\alpha} - i_l\hat{\gamma})^2 + \frac{1}{\hat{\sigma}_y^2} \sum_{l=1}^{n} (y_{i_l,j_l} - \bar{y})^2$$

3. If $S^2 \leq \chi_{2n-3}^2(95\%)$ then accept, remove the $n$ data points and go to 2

4. $n \downarrow (n - 1)$, and go to 2 provided $n \geq 2$

5. $\hat{K} = \text{number of clusters found} + \text{number of remaining singletons}$
Example

\[ \hat{K} = 2 + 5 + 3 = 10 \]
Choice of $S^2 \leq \chi^2_{2n-3}(95\%)$ threshold?

Algorithm applied to 100 simulations of data for different thresholds:

<table>
<thead>
<tr>
<th>threshold %</th>
<th>$\hat{K}$ bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>75</td>
<td>0.83</td>
</tr>
<tr>
<td>90</td>
<td>0.24</td>
</tr>
<tr>
<td>95</td>
<td>0.00</td>
</tr>
<tr>
<td>99</td>
<td>-0.19</td>
</tr>
<tr>
<td>99.5</td>
<td>-0.21</td>
</tr>
<tr>
<td>99.9</td>
<td>-0.27</td>
</tr>
</tbody>
</table>
10 plots with fruits identified by eye

Correlation = 95%
400 plots, with fruits located automatically by image processing

Correlation = 74%
SUMMARY

• Automatic phenotyping is important, but challenging

• Our fruit counting algorithm is fast, and of useable accuracy

• It could be extended to include covariates, such as fruit colour

• Would more sophisticated estimators of $K$ do better?

For further details and related work, see papers on http://www.bioss.ac.uk/people/chris.html