Fitting a bivariate normal distribution to a 2D scatterplot

Florian Hahne
October 17, 2016

1 Overview

Using FACS (fluorescence-activated cell sorter) one can measure certain properties of each individual cell in a population of cells. Examples for these properties:

- Forward light scatter (FSC): this measures a cell’s size
- Sideward light scatter (SSC): this measures a cell’s granularity
- Several fluorescence channels (typically 3 to 4) that measure the abundance of fluorophores, which may be bound to specific antibodies for surface or intracellular markers, or be encoded by a GFP-tagged transcript.

First, we load example data from a FACS analysis that was performed by Mamatha Sauermann at the German Cancer Research Center in Heidelberg.

```r
> library(prada)
> sampdat <- readFCS(system.file("extdata", "fas-Bcl2-plate323-04-04.A01",
+ package="prada"))
> fdat <- exprs(sampdat)
```

The scatterplot of FSC vs SSC is often used for quality control. It is shown in Fig. 1.

```r
> plot(fdat[,"FSC-H"], fdat[,"SSC-H"], pch=20, col="#303030",
+ xlab="FSC", ylab="SSC", main="Scatter plot FSC vs SSC")
```

The cell population is often contaminated by cell debris or conjugates. These can be identified by their size: they are either much smaller or much larger than the main population, or they have an unusual degree of granularity. Segmentation is often performed manually by looking at the FSC-SCC scatterplot.

Here we describe an automated algorithm for this task.

2 Fitting

The package *prada* provides the functions `fitNorm2` and `plotNorm2`. We assume that the shape of the main population in the FSC vs SSC plot can be approximated by a normal distribution. The function `fitNorm2` fits a bivariate normal distribution into the data (by robust estimation
To select the cells from within the ellipse, the list item
> nfit3 <- fitNorm2(fdat[,"FSC-H"], fdat[,"SSC-H"], scalefac=3)
> plotNorm2(nfit, selection=TRUE, ellipse=TRUE)

We can plot this with the function plotNorm2 (see Fig 2). It shows the ellipse, and the set of
discarded points is marked by a red dot. Also the center of the normal distribution is marked
by the red cross.
> plotNorm2(nfit, selection=TRUE, ellipse=TRUE)

To select the cells from within the ellipse, the list item nfit$sel is a logical vector with the
same length as the number of data points.
> cleanfdat <- fdat[nfit$sel,]

Fig. 3 shows again a scatter plot of the two fluorescense channels FL1 and FL4 this time using
the 'clean' data set cleanfdat.
> par(mfrow=c(1,2))
> xlim <- range(fdat[,"FL1-H"])
> ylim <- range(fdat[,"FL4-H"])
> plot(fdat[,"FL1-H"], fdat[,"FL4-H"], pch=20, col="#303030", xlab="FL1",
+ ylab="FL4", main="all data", xlim=xlim, ylim=ylim)
> plot(cleanfdat[,"FL1-H"], cleanfdat[,"FL4-H"], pch=20, col="#303030", xlab="FL1",
+ ylab="FL4", main="clean data only", xlim=xlim, ylim=ylim)
Figure 2: Selection of the main population, using two different values of the parameter `scale-fac`.

Figure 3: Scatter plots of FL1 vs FL4.
3 Scatterplots

If you think that scatterplots with thousands of points are hard to read and annoying to view in a PDF viewer, have a look at the function `smoothScatter` (see Fig. 4):

```r
> smoothScatter(fdat[,c("FSC-H", "SSC-H")], nrpoints=50)
```

![Figure 4: Smooth scatter plots.](image_url)