

### 3 Discuss the origin of noise and its implications on data analysis

#### 3.1 Pages in handouts

5,6.

#### 3.2 Answer

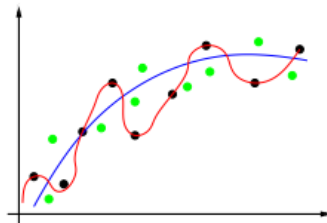
##### 3.2.1 Origin of noise

Noise results from measurement errors (repeating an experiment leads for technical reasons always to different observations), missclassifications (experts get for example the phenotype wrong), and simplifications during modelling. The latter “modelling noise” refers to ignoring certain influence factors in the model which do however alter observations. Although this should be avoided if possible, sometimes these factors are very difficult to control and thus ignored.

Example of the latter: gene expression changes in response to the metabolic state of the model organism (not all organisms are really equally fed at the same time). Such factors must never be confounded with the question at hand since that would lead to misleading results. If they can not be controlled perfectly they will inevitably add random variation to the observations and thus constitute part of the “noise”.

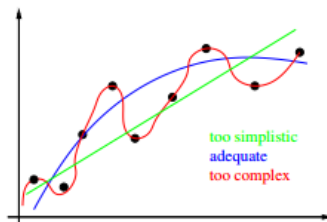
##### 3.2.2 Implications on data analysis

Data analysis attempts finding an appropriate abstraction which explains some given observations (training data) reasonably well. An optimally chosen model should explain new observations (the green dots in the figure below) and avoid fitting the noise component in the measurements.



In the above sketch, the more complex (red) model fits the training data (black dots) better than the blue one. This improvement is though a result from poor averaging and trying to explain the noise.

A good fit is obtained by local averaging of observations and appropriately adjusting the model complexity (blue model).



In the above sketch the too complex and the too simplistic model are inadequate.