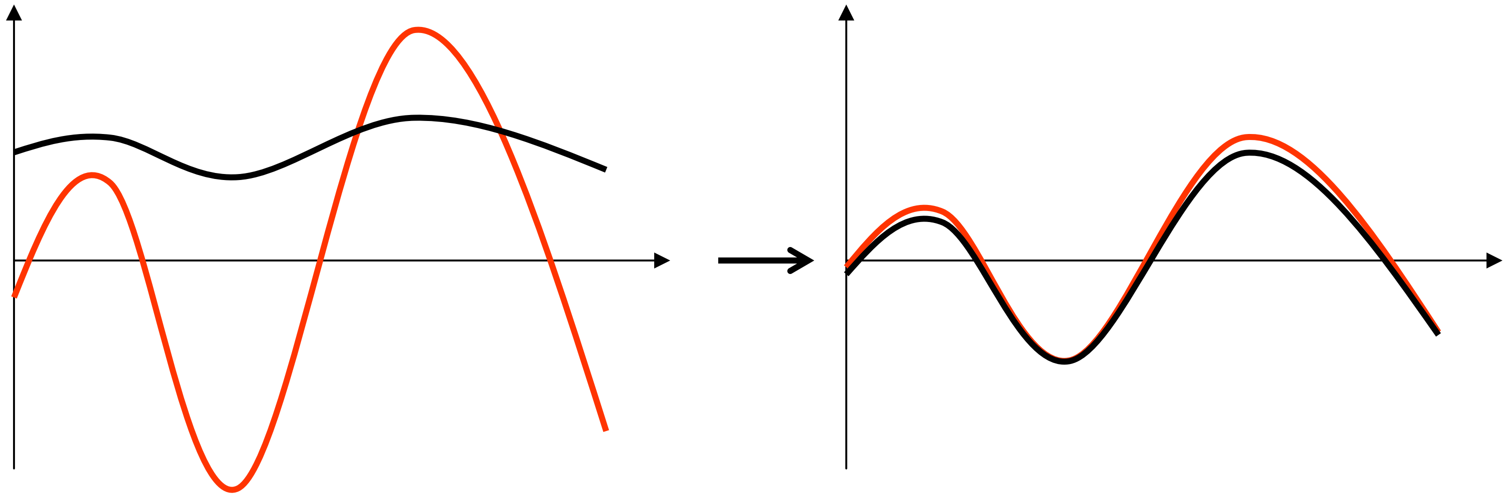
Preprocessing: Other Transformations and Filtering

1. Further improve comparability of measurements across experimental conditions and/or across genes with **centering and standardization**:

- by column of the data matrix (i.e. chip, experimental condition or replicate)
- by row in the data matrix (i.e. gene)

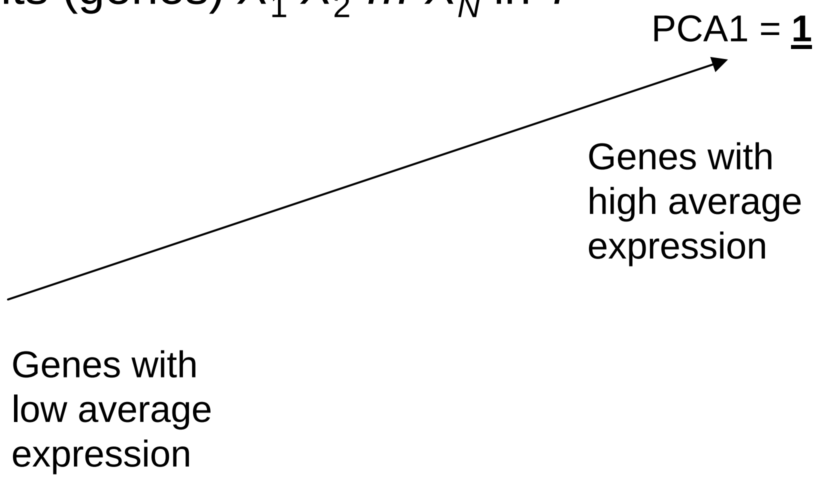
If both, need iterative procedures. Eliminate location and variation magnitude effects: restrict analysis to “pure shapes” (very common). Used in a number of applications.

By column: eliminate location or scale effects that “survived” normalization.
By row: especially when using differential expression and co-expression of genes as means to investigate regulation and co-regulation, what matters is not the average level of expression or the variation magnitude in absolute terms… what matters is the shape of a gene’s expression profile



Think of the gene profiles as a cloud of N points (genes) $X_1\ X_2\ \ldots\ X_N$ in T dimensions (conditions).

If we do not center and standardize the data matrix by row (gene), many analyses will be dominated by a “small vs large transcription” signal – very strong $\underline{\mathbf{1}}$ first PCA



PCA1 = $\underline{\mathbf{1}}$

Genes with high average expression

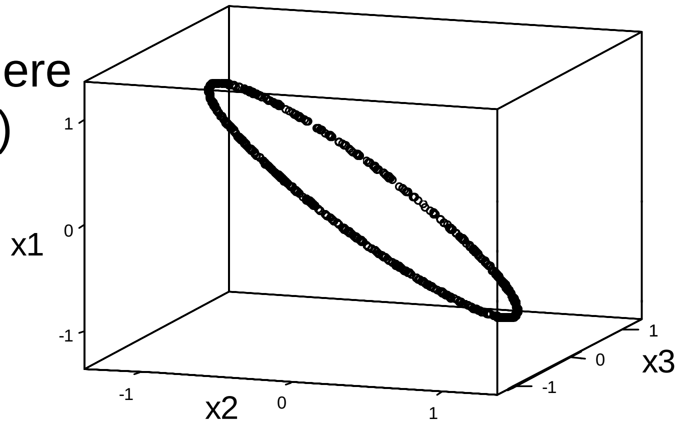
Genes with low average expression

BUT: centering and standardizing the data matrix by row (gene) “creates” geometrical structure:

- Centering creates a linear constraint; points live on a hyperplane (dim T-1)
- Standardizing forces points on a hypersphere

(with both, points will live on the intersection)

$$X_i{}'\ \underline{\mathbf{1}} = 0$$
$$||\ X_i\ ||^2 = 1$$

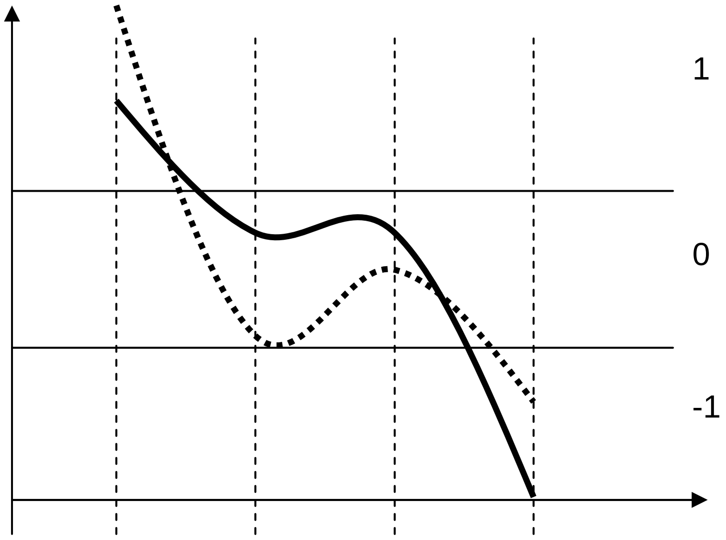


2. Further decrease the effect of unwanted sources of variation by eliminating “detail”, and systematic errors with it:

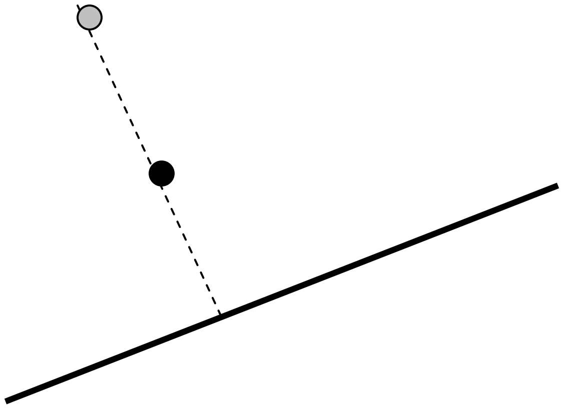
- **Quantization**; discretizing continuous data into (ordered) classes

- **Low-dimensional reconstructions**; approximating expression profiles through a small number of characteristic patterns.

BUT: an arbitrary quantization or low-dimensional reconstruction may induce misleading similarities in gene profiles.



Two profiles are discretized to 1 0 0 –1 .
Are they similar?



Two profiles share a 1-dimensional reconstruction. Are they similar?

3. Get rid of “inert bulk” that can affect detection of interesting signals by standard methods; **filtering** out genes that

- have very low expression in all conditions of interest (e.g. absent calls in affy, or otherwise evaluated small normalized affy signals)
- have very small change wrt a “baseline” in all conditions of interest (e.g. small normalized log-ratios in spotted arrays),
- show very little variation in expression or log-ratios across conditions of interest

IMPORTANT: NEED TO DO THIS BEFORE ROW (GENE) STANDARDIZATION

Preprocessing step: reduce the number of genes considered in further analyses, eliminate completely uninteresting genes – not an aim in itself (identifying genes presenting significant differential expression). Criteria and methods can be less stringent and/or rigorous, one will tend to retain a larger portion of the genes…

Retaining false positives: in preprocessing, it is better to err towards retaining false positives… just avoid that their number is so large as to obscure patterns of, and relationships among, true positives. However, when identifying differentially expressed genes (as main aim), false positives can be as bad as false negatives:

- wrong conclusions
- expenditure for further experimental validation.

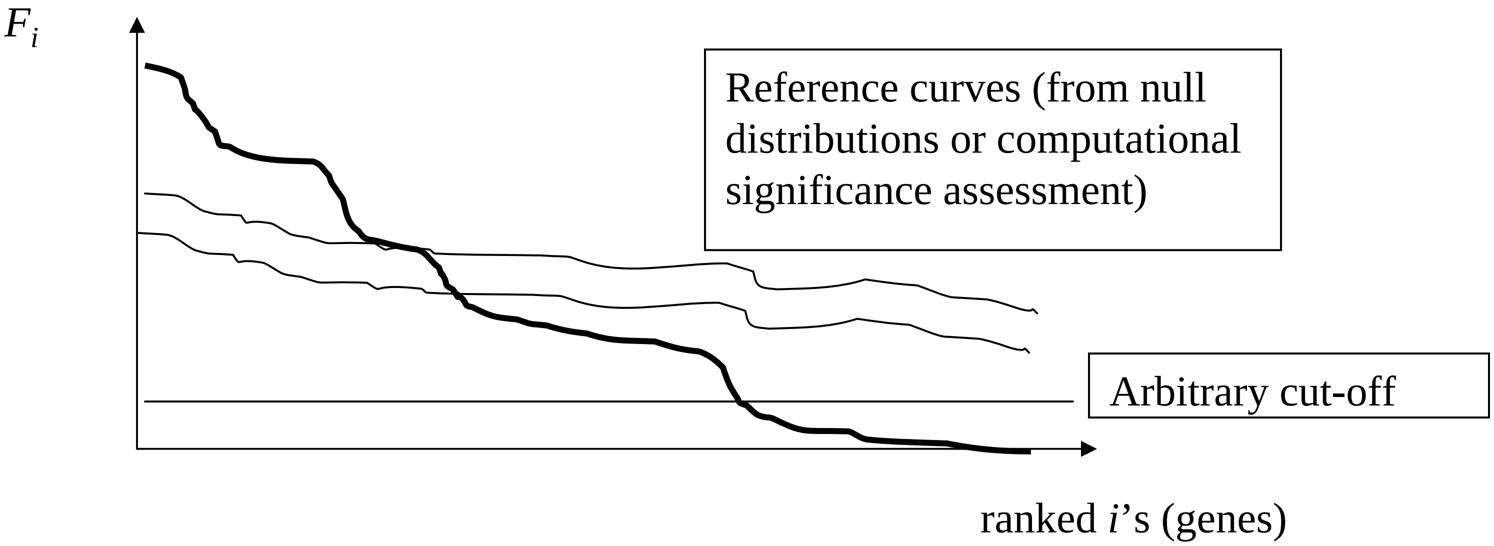
(Note: false positive and false negative rates are in trade-off)

General idea: **use a statistic to create a ranking** (partial ordering of the genes)

For preprocessing, create some arbitrary cut-off along the ranking in terms of

- a value for the statistic
- a number of genes
- a quantile (a percentage of the genes)

But for identifying differentially expressed gene will need to employ a testing mechanism: how many top-ranking genes have a “significant” value of the statistic? Much more later. Note: VERY multiple testing problem



Questions to remember:

- What is the appropriate “scale” to look at our measurements, given the questions we want to address, and the data analysis methods we want to employ?

- Do we introduce “structure” in the data by imputing missing values, or applying transformations (e.g. row centering and standardization)?

- What is the definition of unnecessary “detail”?

- Many data preprocessing steps (including some normalization and missing value imputation techniques; centering & standardizing by row; low-dimensional reconstructions; some definitions of “bulk” for filtering) move across the columns of the data matrix. Preprocessing of the numbers for an individual chip (condition) is “context dependent”; relative to:

  - the set of chips it will be analyzed with.
  - the data analysis methods and models we want to employ.

Do not perform these steps prior to data-basing of microarray information!