In locally weighted regression, we build a function pointwise.

1. Take a point X_0, find the *k* nearest neighbors, which constitute a neighborhood $N(X_0)$. *k* is a percentage of the full dataset, chosen by the user
2. Calculate the largest distance between $X_0$and all points in $N(X_0)$.
$$\Delta(X_0) = max_{i=N(X_0)}|X_0 - X_i|$$
3. Assign weights to each point in neighborhood $N(X_0)$ using a tri-weight function W. Let $u = \frac{|X_0 - X_i|}{\Delta(X_0)}$ - normalized distance
$$W(u) = \begin{cases} (1-u^3)^3, for\ 0 \leq u < 1 \\ 0, \quad otherwise \end{cases}$$
4. Calculate a weighted least squares regression of Y on the $N(X_0)$. Take the fittest value $\widehat{Y_0} = S(X_0)$ as whatever function of X_0.
5. Go to 1. And repeat for all other points (X values) in the dataset.

For a 2-channel microarray

1. Fit a lowess regression to the MA plot
2. Obtain the fitted values $\widehat{M}_k$
3. Adjust the probe values by $M_k^{norm} = M_k - \widehat{M}_k$. Remember, that $M_k = \log(X_k) - \log(Y_k)$

For a single channel microarray

Global scaling method:

- Choose a baseline chip, calculate $\tilde{X}_{base}$ - 2% trimmed mean (Tilde represents trimmed mean).
- For each other array, multiply $\tilde{X}_i$ (2% trimmed mean for the *i* array) by a corresponding scaling factor $SF_i = \tilde{X}_{base}/\tilde{X}_i$.

Cyclic loess:

- MA plots are useful for normalizing _between_ gene chips.
- If the gene chips are considered "replicates", we would expect the intensities to be approximately the same - for each pairwise MA plots should be centered around 0
- Fit a loess curve to each MA plot and save each the $\widehat{M}_k$ (estimated) value
- $M_k^{norm} = M_k - \widehat{M}_k$ is used to adjust the M_k (known) data
- The adjustment is partitioned equally among all possible MA plots that include the chips
$$X_{ki}^{loess-norm} = 2^{A_k + M_k'/2}$$

where M_k' is average over the different loess fits. This brings the adjusted data on the raw scale.

Invariant set method:

- Rather than using all the probes for normalization, one may want to restrict attention to the set of probes that are *invariant* (stable, presumed not to chage gene expression) across the set of chips.
- Same as base normalization in relation to probes that are non-differentially expressed
- First proposed by Li and Wong
- Detection of invariant set:
    - For two chips – rank PM probe intensities separately
    - The ranked expression for the *ith* chip R_{ik}, $k = 1, \ldots, G$.
    - For two chips, we have R_{1k} and R_{2k} => $\Delta k = R_{1k} - R_{2k}$
    - For multiple chips, $\Delta k = \max(R_{ik}) - \min(R_{ik})$
    - A probe set is "invariant" if its $\frac{\Delta k}{G} \leq 0.003$. This threshold is manually picked. The effect for low and high expressed genes would be different – rank difference in low expression space is different from the high expression space
- Fit a piecewise linear running median line

Quantile normalization:

- How you assess is a sample comes from the normal distributions? With QQ plots – plot the *ordered values* against the corresponding quantiles from that distribution
- Quantile points are defined as

$$\frac{i - {}^{1}\!/_{2}}{n}, if\ n \geq 11$$

or

$$\frac{i - {}^{3}\!/_{8}}{n + {}^{1}\!/_{4}}, ifn \leq 10$$

- *Quantile normalization idea:* If the same distribution is expected for all replicate gene chips, the quantiles for each chip should agree
- Let X represent a matrix with N columns and G rows
- Sort each column in X and define the sorted matrix X_{sort}.
- Project each row of X_{sort} to get your overall quantiles
- Get X_{norm} by re-arranging each column to have the same order as the original X.