Mathematical Supplement to
Centralization: A New Method for the Normalization of Gene Expression Data

Alexander Zien
Fraunhofer Institute for Algorithms and Scientific Computing (FhI-SCAI)
Schloß Birlinghoven, 53754 Sankt Augustin, Germany
Alexander.Zien@scai.fhg.de

December 6, 2002

Core to the **Centralization** method [1] for the normalization of gene expression data is the estimation of most consistent sample scaling factors based on possibly inconsistent pairwise estimates. Here, the maximum likelihood approach that is employed in order to find the most probable consistent scaling vector is described in detail.

1 Estimation of Most Consistent Sample Scaling Factors

Let $i, j \in A$ be two microarray measurements of any two samples with unknown scaling factors s_i, s_j. Thus, the true relative scaling of the two samples is $r_{i,j} := \frac{s_i}{s_j}$. Let $m_{g,i}$ denote the background-corrected measurement value of gene g in sample i. Now $G(i,j) \subset G(i) \cap G(j)$ is defined to be the set of genes that are considered to be expressed and reliably measured (in the linear part of the dynamic range) in both samples, as judged by the measurements. Formally, this may be defined by using a lower bound m_{min} and an upper bound m_{max} on the expression levels, for example:

$$G(i,j) \quad := \quad \{g \in G(i) \cap G(j) \mid m_{min} \leq m^*_{gi} \leq m_{max} \wedge m_{min} \leq m^*_{gj} \leq m_{max}\} \tag{1}$$

Let $m_{i,j} := |G(i,j)|$ be the number of these genes. The other genes are excluded, since ratios of values that are dominated by background noise (as well as saturated intensities) are incorrectly biased towards one. If a measurement for a gene g is available for only one of the samples, the gene is also excluded from $G(i,j)$.

In order to estimate $r_{i,j}$, the set of quotients of the background-corrected measurements,

$$Q_{i,j} \quad := \quad \left\{ q_g \,\middle|\, q_g := \frac{m_{g,i}}{m_{g,j}}, g \in G(i,j) \right\}, \tag{2}$$

will be used. Let $(q_1, \ldots, q_{m_{i,j}})$ be the ascendingly sorted list of these quotients. The idea is to regard each of the ratios $q_g \in Q_{i,j}$ as an estimate of the pairwise relative bias $r_{i,j}$; In general, any measure of central tendency of the values in $Q_{i,j}$, including median, mean, trimmed mean and weighted mean, may yield a sensible estimate $\hat{r}_{i,j}$ for $r_{i,j}$. However, care must be taken whenever values are averaged: the arithmetic mean should be computed in the space of log ratios in order to keep the symmetry of up- and downregulation.

In the following it is assumed that reasonable estimates $\hat{r}_{i,j}$ of the true quotients $r_{i,j}$ are given for all pairs $i, j \in A$ of microarrays. Now the task is to determine estimates $\hat{s}_a$ of the array-dependent multiplicative errors s_a for all arrays $a \in A$. With such values, the measurement values $m_{g,a}$ can be made mutually comparable between different arrays a by rescaling them accordingly via

$$m_{g,a} \quad \rightarrow \quad s_a m_{g,a}. \tag{3}$$

While this does not recover the true number of mRNA molecules for a gene in the sampled cells, it leads to consistent multiples of this number. This enables comparisons of the expression of any fixed gene between different arrays.

If the quotients $Q_{i,j}$ were free of measurement errors and the assumptions of well-behavedness were totally and exactly fullfilled, the estimated relative scalings $\hat{r}_{i,j}$ would equal the true values $r_{i,j}$. In particular, they would be triangle-consistent, i.e. $\hat{r}_{i,k} = \hat{r}_{i,j}\hat{r}_{j,k}$. Then, the values $s_1, \ldots, s_n$ could be determined by $s_l := q_{l,i}$ for all $1 \leq l \leq n$ with any fixed $i \in \{1, \ldots, n\}$. However, due to measurement errors and because the assuption of well-behaved regulation is only roughly correct, usually the values $\hat{r}_{i,j}$ are inaccurate and do not satisfy the triangle condition. In consequence, the result of the above computation depends on the ordering of the samples, which is unsatisfactory.

A solution that is independent of the ordering of the microarrays or other arbitrary circumstances can be found through a maximum likelihood approach. According to this method, a set $\hat{\theta}$ of parameter values is computed that maximizes the probability of the observed data under a given parametric model. The maximum likelihood method may be formalized like this:

$$\hat{\theta} \quad \in \quad \arg_\theta \max_\theta P(\text{data}|\theta, \text{model}). \tag{4}$$

In the case of centralization, the parameters are the scaling factors s_i for the arrays and measures $\sigma^2_{i,j}$ of variances of the log ratios in $Q_{i,j}$. Thus, $\theta = (\mathbf{s}, \Sigma)$, where $\mathbf{s} := (s_1, \ldots, s_n)$ and $\Sigma := (\sigma_{i,j})_{i,j \in \{1,\ldots,n\}}$. Although the data are actually the measured expression levels, it will be more convenient to use the sets $Q_{i,j}$ of ratios, which are directly derived from the primary data. Thus, the data will be represented by $\mathbf{Q} := (Q_{i,j})_{i,j \in \{1,\ldots,n\}}$. Application of the maximum likelihood approach to the present situation therefore yields the following estimation $\hat{\mathbf{s}}$ of the true scaling factors:

$$\hat{\mathbf{s}} \quad \in \quad \arg_{\mathbf{s}} \max_{\mathbf{s},\Sigma} P(\mathbf{Q}|\mathbf{s}, \Sigma, \text{model}), \tag{5}$$

where the model still remains to be specified. The model characterizes how the probability of observing data depends on the parameter values. It thus consists of a parametric mathematical function.

In bioinformatics, the mathematical models used in probabilistic methods frequently form a compromise between an accurate representaion of reality and what is mathematically and/or algorithmically feasible. For example, most prominent sequence alignment methods assume that the character distributions at the different sequence positions are indepent, although in reality dependencies are known to exist. However, this allows to sum positional scores, which represent logarithms of (products and quotients of) probabilities. Similar compromises will be taken for centralization, as is detailed now.

First, the likelihood functions is assumed to be decomposable into n^2 (conditionally) independent terms, one for each pairwise comparison of arrays:

$$P\left(\mathbf{Q}|\,\mathbf{s}, \Sigma\right) \quad := \quad \prod_{i,j=1}^{n} P\left(Q_{i,j}|\,s_i, s_j, \sigma_{i,j}\right) \tag{6}$$

This assumption of independence can be considered to be violated, because the $\frac{1}{2}(n^2 - n)$ essentially different sets of quotients $Q_{i,j}$ $(\hat{\mu}_{i,j}, \hat{\sigma}_{i,j})$ are computed from only n microarrays, and the $\frac{1}{2}(n^2 - n)$ different variances relate to only n independent multiplicative errors s_i. On the other hand, the dependencies among the values are symmetrical, and not systematically biased towards certain arrays or pairs of arrays. Thus, while the optimal probability value will not be meaningful in itself, one can still expect to obtain a close to optimal scaling s through maximizing the product probability. In any case, it simplifies the coming calculations.

Second, the distribution of the log-ratios of each pairwise comparison is assumed to follow a normal distribution, where the mean corresponds to the log of the pairwise scalings, $\log \frac{s_i}{s_j}$, and the variance $\sigma^2_{i,j}$ results from the amount of measurement error and biological change between the two measurements. Thus,

$$q \quad \sim \quad \mathcal{N}\left(\log \frac{s_i}{s_j}, \sigma^2_{i,j}\right) \tag{7}$$

for $q \in Q_{i,j}$, or more explicitly:

$$P\left(q \in Q_{i,j}|\,s_i, s_j, \sigma_{i,j}\right) \quad := \quad \frac{1}{\sigma_{i,j}\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\log q - \log \frac{s_i}{s_j}}{\sigma_{i,j}}\right)^2\right). \tag{8}$$

This is a possible formalization of the essential properties of well-behaved gene regulation: changes of expression are symmetric (with respect to up- and downregulation), and most of the genes are close to unregulated. Of course, other distributions than the normal distribution can also deliver one or both of these properties. But as was shown in the preceding section, there are theoretical arguments for the normal distribution, and it also has the advantage that it makes the calculation of maximum likelihood parameter values possible and easy, as will be shown during this section.

While Equation 8 assigns a probability to each ratio $q \in Q_{i,j}$, Equation 6 requires a single probability for each entire set $Q_{i,j}$. One possibility is to model the genes $g \in G(i,j)$ as being independently regulated:

$$P(Q_{i,j}|\, s_i, s_j, \sigma_{i,j}) \quad := \quad \prod_{q \in Q_{i,j}} P(q \in Q_{i,j}|\, s_i, s_j, \sigma_{i,j}) \tag{9}$$

However, it is known that dependencies between the expression levels of genes exist in reality, for example caused by shared transcription factor binding sites in their promoters. It can even be hypothesized that in a typical microarray measurement (with a large numbers of genes), virtually all independent information is covered and that most genes, whether included on the array or not, are therefore redundant. Consequently, here the approach is taken to give each pairwise comparison the same weight, independent of the sizes of the sets $Q_{i,j}$ (which, in typical application cases, are similar anyway), by computing an average single-gene probability for each pair (i,j):

$$P(Q_{i,j}|\, s_i, s_j, \sigma_{i,j}) \quad := \quad \left(\prod_{q \in Q_{i,j}} P(q \in Q_{i,j}|\, s_i, s_j, \sigma_{i,j}) \right)^{\frac{1}{|Q_{i,j}|}} \tag{10}$$

Although the equation becomes more complex at this stage, the resulting algorithm will be simplified by this operation, as is shown in the following calculations.

Insertion of Equations 10 and 8 into Equation 6 yields the final likelihood function:

$$P(\mathbf{Q}|\, \mathbf{s}, \Sigma) \quad = \quad \prod_{i,j=1}^{n} \left[\prod_{q \in Q_{i,j}} \frac{1}{\sigma_{i,j}\sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{\log q - \log \frac{s_i}{s_j}}{\sigma_{i,j}} \right)^2 \right) \right]^{\frac{1}{|Q_{i,j}|}} \tag{11}$$

Parameter values $\hat{\mathbf{s}}$ are sought that maximize this likelihood function. Since the logatithm is a strictly monotonic increasing function, it is equivalent to search for parameters that maximize the log-likelihood, which can be written like this:

$$\begin{aligned}
&\log P(\mathbf{Q}|\, \mathbf{s}, \Sigma) \qquad (12)\\
&= \log \prod_{i,j=1}^{n} \left[\prod_{q \in Q_{i,j}} \frac{1}{\sigma_{i,j}\sqrt{2\pi}} \exp\left(-\frac{1}{2} \left(\frac{\log q - \log \frac{s_i}{s_j}}{\sigma_{i,j}} \right)^2 \right) \right]^{\frac{1}{|Q_{i,j}|}} \\
&= \sum_{i,j=1}^{n} \log \left\{ \frac{1}{\sigma_{i,j}\sqrt{2\pi}} \left[\prod_{q \in Q_{i,j}} \exp\left(-\frac{1}{2} \left(\frac{\log q - \log \frac{s_i}{s_j}}{\sigma_{i,j}} \right)^2 \right) \right]^{\frac{1}{|Q_{i,j}|}} \right\} \\
&= \sum_{i,j=1}^{n} \left\{ \log \frac{1}{\sigma_{i,j}\sqrt{2\pi}} + \frac{1}{|Q_{i,j}|} \log \left[\prod_{q \in Q_{i,j}} \exp\left(-\frac{1}{2} \left(\frac{\log q - \log \frac{s_i}{s_j}}{\sigma_{i,j}} \right)^2 \right) \right] \right\} \\
&= \sum_{i,j=1}^{n} \left\{ \log \frac{1}{\sigma_{i,j}\sqrt{2\pi}} + \frac{1}{|Q_{i,j}|} \left[\sum_{q \in Q_{i,j}} -\frac{1}{2} \left(\frac{\log q - \log \frac{s_i}{s_j}}{\sigma_{i,j}} \right)^2 \right] \right\} \\
&= \sum_{i,j=1}^{n} \log \frac{1}{\sigma_{i,j}\sqrt{2\pi}} - \frac{1}{2} \sum_{i,j=1}^{n} \frac{1}{|Q_{i,j}|} \frac{1}{\sigma_{i,j}^2} \sum_{q \in Q_{i,j}} \left(\log q - \log \frac{s_i}{s_j} \right)^2
\end{aligned}$$

In the following auxilliary calculations the rightmost sum of the latter latter formulation of the log-likelihood is expressed in terms of the empirical mean

$$\hat{\mu}_{i,j} := \frac{1}{|Q_{i,j}|} \sum_{q \in Q_{i,j}} \log q \tag{13}$$

and the empirical variance

$$\hat{\sigma}_{i,j}^2 := \frac{1}{|Q_{i,j}|} \sum_{q \in Q_{i,j}} (\log q - \hat{\mu}_{i,j})^2 \tag{14}$$

of the set $\{\log q | q \in Q_{i,j}\}$ of log ratios. This simplification is possible thanks to the term $\frac{1}{|Q_{i,j}|}$ introduced by taking the geometric mean over the $q \in Q_{i,j}$.

$$\begin{aligned}
&\frac{1}{|Q_{i,j}|} \sum_{q \in Q_{i,j}} \left(\log q - \log \frac{s_i}{s_j} \right)^2 \qquad (15)\\
&= \frac{1}{|Q_{i,j}|} \sum_{q \in Q_{i,j}} \left(\log q - \hat{\mu}_{i,j} + \hat{\mu}_{i,j} - \log \frac{s_i}{s_j} \right)^2 \\
&= \frac{1}{|Q_{i,j}|} \sum_{q \in Q_{i,j}} \left[(\log q - \hat{\mu}_{i,j})^2 + (\log q - \hat{\mu}_{i,j}) \left(\hat{\mu}_{i,j} - \log \frac{s_i}{s_j} \right) + \left(\hat{\mu}_{i,j} - \log \frac{s_i}{s_j} \right)^2 \right] \\
&= \frac{1}{|Q_{i,j}|} \sum_{q \in Q_{i,j}} (\log q - \hat{\mu}_{i,j})^2 \\
&\quad + \left(\hat{\mu}_{i,j} - \log \frac{s_i}{s_j} \right) \left(\frac{1}{|Q_{i,j}|} \sum_{q \in Q_{i,j}} \log q - \hat{\mu}_{i,j} \right) \\
&\quad + \left(\hat{\mu}_{i,j} - \log \frac{s_i}{s_j} \right)^2 \frac{1}{|Q_{i,j}|} \sum_{q \in Q_{i,j}} 1 \\
&= \hat{\sigma}_{i,j}^2 + 0 + \left(\hat{\mu}_{i,j} - \log \frac{s_i}{s_j} \right)^2
\end{aligned}$$

Inserting the result into Equation 12 yields the final version of the log-likelihood function:

$$\begin{aligned}
&\log P(\mathbf{Q} | \mathbf{s}, \Sigma) \qquad (16)\\
&= \sum_{i,j=1}^{n} \log \frac{1}{\sigma_{i,j}\sqrt{2\pi}} - \frac{1}{2} \sum_{i,j=1}^{n} \frac{1}{\sigma_{i,j}^2} \left[\hat{\sigma}_{i,j}^2 + \left(\hat{\mu}_{i,j} - \log \frac{s_i}{s_j} \right)^2 \right] \\
&= \sum_{i,j=1}^{n} \log \frac{1}{\sigma_{i,j}\sqrt{2\pi}} - \frac{1}{2} \sum_{i,j=1}^{n} \frac{\hat{\sigma}_{i,j}^2}{\sigma_{i,j}^2} - \frac{1}{2} \sum_{i,j=1}^{n} \left(\frac{\hat{\mu}_{i,j} - \log \frac{s_i}{s_j}}{\sigma_{i,j}} \right)^2
\end{aligned}$$

Thus, optimal parameter values of the model are defined by:

$$\begin{aligned}
\hat{\mathbf{s}} &\in \arg_{\mathbf{s}} \max_{\mathbf{s},\Sigma} [\log P(\mathbf{Q} | \mathbf{s}, \Sigma)] \qquad (17)\\
&= \arg_{\mathbf{s}} \max_{\mathbf{s},\Sigma} \left[\sum_{i,j=1}^{n} \log \frac{1}{\sigma_{i,j}\sqrt{2\pi}} - \frac{1}{2} \sum_{i,j=1}^{n} \frac{\hat{\sigma}_{i,j}^2}{\sigma_{i,j}^2} - \frac{1}{2} \sum_{i,j=1}^{n} \left(\frac{\hat{\mu}_{i,j} - \log \frac{s_i}{s_j}}{\sigma_{i,j}} \right)^2 \right] \\
&= \arg_{\mathbf{s}} \min_{\mathbf{s},\Sigma} \left[\sum_{i,j=1}^{n} \left(\frac{\hat{\mu}_{i,j} - \log \frac{s_i}{s_j}}{\sigma_{i,j}} \right)^2 + \sum_{i,j=1}^{n} \frac{\hat{\sigma}_{i,j}^2}{\sigma_{i,j}^2} - 2 \sum_{i,j=1}^{n} \log \frac{1}{\sigma_{i,j}\sqrt{2\pi}} \right]
\end{aligned}$$

In order to slightly simplify the following calculations, the substitution variable $t_l := \log \hat{s}_l$ is introduced. This leads to the following objective function to be minimized:

$$\sum_{i,j=1}^{n} \left(\frac{\hat{\mu}_{i,j} - (t_i - t_j)}{\sigma_{i,j}} \right)^2 + \sum_{i,j=1}^{n} \frac{\hat{\sigma}_{i,j}^2}{\sigma_{i,j}^2} - 2 \sum_{i,j=1}^{n} \log \frac{1}{\sigma_{i,j}\sqrt{2\pi}}. \tag{18}$$

A minimum of the function in Equation 18 can be found by equating all partial derivatives with zero. First, the derivatives with respect to the variables Σ are calculated. For each $\sigma_{k,l}$ with $k, l \in \{1, \ldots, n\}$

this yields the equation:

$$\begin{aligned} 0 &\stackrel{!}{=} \frac{\delta}{\delta\sigma_{k,l}}\left[\sum_{i,j=1}^{n}\left(\frac{\hat{\mu}_{i,j}-t_i+t_j}{\sigma_{i,j}}\right)^2+\sum_{i,j=1}^{n}\frac{\hat{\sigma}_{i,j}^2}{\sigma_{i,j}^2}-2\sum_{i,j=1}^{n}\log\frac{1}{\sigma_{i,j}\sqrt{2\pi}}\right] \\ &= (\hat{\mu}_{k,l}-t_k+t_l)\frac{-2}{\sigma_{k,l}^3}+\frac{-2\hat{\sigma}_{k,l}^2}{\sigma_{k,l}^3}-2\left(\frac{-1}{\sigma_{k,l}}\right) \\ &= -2\left((\hat{\mu}_{k,l}-t_k+t_l)\frac{1}{\sigma_{k,l}^3}+\frac{\hat{\sigma}_{k,l}^2}{\sigma_{k,l}^3}-\frac{1}{\sigma_{k,l}}\right). \end{aligned} \tag{19}$$

Using $\hat{\sigma}_{k,l} > 0$, this can be equivalently reformulated as:

$$\begin{aligned} 0 &= (\hat{\mu}_{k,l}-t_k+t_l)+\hat{\sigma}_{k,l}^2-\sigma_{k,l}^2 \\ \Leftrightarrow\ \sigma_{k,l}^2 &= \hat{\sigma}_{k,l}^2+(\hat{\mu}_{k,l}-t_k+t_l) \end{aligned} \tag{20}$$

At this place it is useful to go back to one original assumption of centralization: that the central tendency of gene expression ratios between two microarrays represents a reasonable estimate of their pairwise relative rescaling: $\hat{\mu}_{k,l} \approx \log\frac{s_i}{s_j}$. This implies that the amount of inconsistency in the estimates of the pairwise relative rescaling of the microarrays is moderate. With this assumption, it follows from Equation 20 that the empirical standard deviations are reasonable approximations of the true standard deviations,

$$\sigma_{k,l}^2 \approx \hat{\sigma}_{k,l}^2, \tag{21}$$

which may also be expected by intuition.

Second, the derivatives of the objective function Equation 18 with respect to the variables $\mathbf{s}$ are calculated and equated to zero for each $k = 1, \ldots, n$:

$$\begin{aligned} 0 &\stackrel{!}{=} \frac{\delta}{\delta t_k}\left[\sum_{i,j=1}^{n}\left(\frac{\hat{\mu}_{i,j}-t_i+t_j}{\sigma_{i,j}}\right)^2+\sum_{i,j=1}^{n}\frac{\hat{\sigma}_{i,j}^2}{\sigma_{i,j}^2}-\frac{2}{\sqrt{2\pi}}\sum_{i,j=1}^{n}\log\frac{1}{\sigma_{i,j}}\right] \\ &= \frac{\delta}{\delta t_k}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\frac{\hat{\mu}_{i,j}-t_i+t_j}{\sigma_{i,j}}\right)^2 \\ &= \sum_{j\neq k}\frac{\delta}{\delta t_k}\left(\frac{\hat{\mu}_{k,j}-t_k+t_j}{\sigma_{k,j}}\right)^2+\sum_{i\neq k}\frac{\delta}{\delta t_k}\left(\frac{\hat{\mu}_{i,k}-t_i+t_k}{\sigma_{i,k}}\right)^2+\frac{\delta}{\delta t_k}\left(\frac{\hat{\mu}_{k,k}-t_k+t_k}{\sigma_{k,k}}\right)^2 \\ &= \sum_{j\neq k}\frac{-1}{\sigma_{k,j}}\cdot 2\left(\frac{\hat{\mu}_{k,j}-t_k+t_j}{\sigma_{k,j}}\right)+\sum_{i\neq k}\frac{1}{\sigma_{i,k}}\cdot 2\left(\frac{\hat{\mu}_{i,k}-t_i+t_k}{\sigma_{i,k}}\right) \\ &= \sum_{l\neq k}\frac{-2}{\sigma_{k,l}^2}(\hat{\mu}_{k,l}-t_k+t_l)+\sum_{l\neq k}\frac{-2}{\sigma_{l,k}^2}(-\hat{\mu}_{l,k}+t_l-t_k) \\ &= -2\sum_{l\neq k}\frac{1}{\sigma_{k,l}^2}(\hat{\mu}_{k,l}-t_k+t_l)-2\sum_{l\neq k}\frac{1}{\sigma_{k,l}^2}(\hat{\mu}_{k,l}-t_k+t_l) \\ &= -4\sum_{l\neq k}\frac{\hat{\mu}_{k,l}-t_k+t_l}{\sigma_{k,l}^2} \\ &= -4\left[\sum_{l\neq k}\frac{\hat{\mu}_{k,l}}{\sigma_{k,l}^2}-\left(\sum_{l\neq k}\frac{1}{\sigma_{k,l}^2}\right)t_k+\sum_{l\neq k}\frac{1}{\sigma_{k,l}^2}t_l\right]. \end{aligned} \tag{22}$$

A simultaneous solution of these n equations together with the $\frac{1}{2}(n^2-n)$ non-redundant instances of Equation 20 is difficult. Solving those equations for $\sigma_{k,l}^2$ and substituting the result into Equation 22 introduces the variables t_l into the denominators. Since the least common denominator contains the product of all variables t_l for $l = 1, \ldots, n$, it becomes difficult or even impossible to find a closed form solution of the problem.

Fortunately, the problem simplifies considerably when the approximation given by Equation 21 is used. The resulting system of equations

$$\sum_{l \neq k} \frac{\hat{\mu}_{k,l}}{\hat{\sigma}^2_{k,l}} \quad = \quad \left(\sum_{l \neq k} \frac{1}{\hat{\sigma}^2_{k,l}} \right) t_k + \sum_{l \neq k} \left(\frac{-1}{\hat{\sigma}^2_{k,l}} \right) t_l \tag{23}$$

(again, for each k) is linear in t and can thus easily be solved in $\mathcal{O}\left(n^3\right)$ time by Gauss elimination.

However, this system of equations has no unique solution. The corresponding matrix has rank $n - 1$, since there is one degree of freedom left that relates the measurement values to mRNA molecule numbers. In this equation, the array-dependent scaling factors s_a can be multiplied by an arbitrary positive constant, when the gene-dependent scaling factors d_g are simultaneously divided by the same value. Since there is no way to determine the gene-dependent factors from microarray measurement data alone, the array-dependent scaling can only be determined modulo a common factor.

By simply adding any constraint on the absolute values of t to any row of the matrix, the full rank can be restored to the matrix and thereby a unique solution enforced.[1] Here, the constraint

$$\sum_{k=1}^{n} t_k \quad = \quad 0 \tag{24}$$

is chosen, corresponding to the claim for unbiased (over the set of arrays) array-wise multiplicative errors s_a. This keeps the normalized expression levels as close to the raw intensities as possible. After solving the equation system, the normalization factors $\hat{s}_a$ are obtained by reversal of the substitution:

$$\hat{s}_a \quad = \quad \exp(t_a). \tag{25}$$

The total runtime of both steps of centralization sums up to $\mathcal{O}\left(n^2 m + n^3\right)$. In practice, the application of a Java implementation to the data sets of this study takes a couple of seconds on a current workstation. A simple software tool has been made available through the Internet[2].

References

[1] Alexander Zien, Thomas Aigner, Ralf Zimmer, and Thomas Lengauer. Centralization: A new method for the normalization of gene expression data. *Bioinformatics*, 17:S323–S331, June 2001. Supplement 1.

[1] In general, there may still be several different sets of parameter values that maximize the probability. For centralization, this is not the case except for degenerate problems, that have not yet been observed in practical application.

[2] at `http://cartan.gmd.de/~zien/centralization/`