Note: all parameters, including ϕ, g, are on log scale.

Suppose that ϕ_{ij} follows a normal distribution with mean ϕ_i and standard deviation σ_ϕ: $\phi_{ij} \sim N(\phi_i, \sigma_\phi^2)$, and ϕ_i's are i.i.d. from $N(\mu, \sigma_\mu^2)$, where $1 \leq i \leq n$ and $1 \leq j \leq m$. (In our case, $n = 106$ and $m = 3$.) Suppose there is a linear relationship between g and ϕ that holds for every gene: $g_i = a + b\phi_i$, then it follows that $g_i + \phi_i = a + (b+1)\phi_i + \epsilon_i$, for $1 \leq i \leq n$ where ϵ_i is the noise in the estimators/observations for gene i, and $\epsilon_i \sim N(0, \sigma_{g\phi}^2)$. Based on these assumptions we have the probability of observing $g\phi_i$ and ϕ_{ij}:

$$\begin{aligned} & P(g\phi_i, \phi_{ij} | \mu, \sigma_\phi, \sigma_\mu, \sigma_{g\phi}, a, b) \\ = & \int P(g_i + \phi_i | a, b, \phi_i, \sigma_{g\phi}) P(\phi_{ij}, \phi_i | \mu, \sigma_\mu) \, \mathrm{d}F(\phi_i) \\ = & \int_{-\infty}^{\infty} \mathbf{\Phi}(\phi_{ij} | \phi_i, \sigma_\phi) \mathbf{\Phi}(\phi_i | \mu, \sigma_\mu) \mathbf{\Phi}(g_i + \phi_i | a + (b+1)\phi_i, \sigma_{g\phi}) \, \mathrm{d}\phi_i \end{aligned}$$

where $\mathbf{\Phi}(x | \mu, \sigma)$ is the probability density function for normal distribution with mean μ and variance σ^2. With the likelihood function for the 6 parameters, and the observed $g_i + \phi_i$ and ϕ_{ij} we can find the MLE(maximum likelihood estimates) for the parameters.

Following is a simulation with parameters $\mu = -2.7$, $a = 3$, $b = 1$, $\sigma_\mu = 0.1$, $\sigma_\phi = 1$ (bigger noise in the estimates of $g\phi$), $\sigma_{g\phi} = 0.1$, and the following plot is produced: Figure 1 shows that, even though the true relationship between g and ϕ is $g = \phi + 3$, which is linear function with positive slope, the relationship shown by the estimated data with noise is

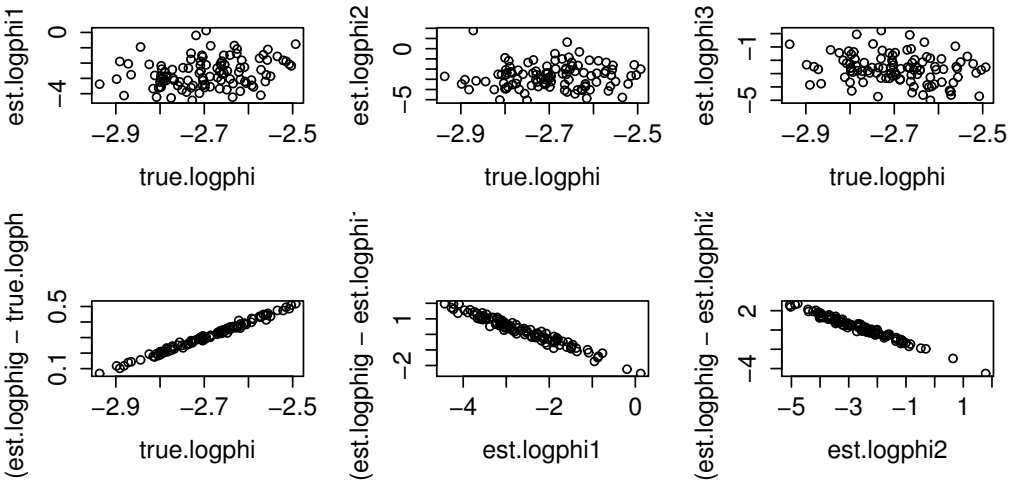

Figure 1: Top row: scatterplots of estimated and true values of $\log(\phi)$; Bottom row: scatterplots of g and ϕ.

negatively correlated, exactly opposite of what is used in the simulation. Whereas the ML estimates for the parameters are not consistent with the true parameters, in that likelihood of true parameters is much smaller than that of some other parameters far from it. Possible reason for this is the impreciseness of the integral in the likelihood function caused by product of very small numbers. - results from R and Mathematica are substantially different.

Explanation of why this happens: Because of the large noise in the estimated values of $\log(\phi)$, the line describing the true relationship between g and $\log(\phi)$ is going to be a line with positive slope but very big width (very thick line with positive slope). When the thickness exceeds the length of the

line itself (which depends on the range of the g and $\log(\phi)$), this line will look like that it has a negative slope. As the noise increases to a point when the width of the line is much bigger than the length of the line, it looks like a thin line with a negative slope.In this case, $\log(\phi g) - \log(\phi)$ is very close to $-\log(\phi) + \text{constant}$ because of small variance in $\log(\phi g)$, therefore the slope of -1 in the figure.