ESTIMATING INTRINSIC AND EXTRINSIC NOISE FROM SINGLE-CELL GENE EXPRESSION MEASUREMENTS

AUDREY QIUYAN FU AND LIOR PACHTER

Abstract. Gene expression is stochastic and displays variation ("noise") both within and between cells. Intracellular (intrinsic) variance can be distinguished from extracellular (extrinsic) variance by applying the law of total variance to data from two-reporter assays that probe expression of identical gene pairs in single-cells. We examine established formulas for the estimation of intrinsic and extrinsic noise and provide interpretations of them in terms of a hierarchical model. This allows us to derive corrections that minimize the mean squared error, an objective that may be important when sample sizes are small. The statistical framework also highlights the need for quantile normalization, and provides justification for the use of the sample correlation between the two reporter expression levels to estimate the percent contribution of extrinsic noise to the total noise. Finally, we provide a geometric interpretation of these results that clarifies the current interpretation.

1. Introduction

In a classic paper on the stochasticity of gene expression, Elowitz *et al.* [2] describe a clever two-reporter expression assay designed to tease apart "intrinsic" and "extrinsic" noise from the overall variability in gene expression. The idea is as follows: two identically regulated reporter genes (cyan fluorescent protein and yellow fluorescent protein) are inserted into individual *E. coli.* cells allowing for comparable expression measurements within and between cells. If n cells are assayed, this leads to expression measurements $c_1, \ldots, c_n$ and $y_1, \ldots, y_n$ where the pair (c_i, y_i) represent the expression measurements for the cyan and yellow reporters in the ith cell. The goal of the experiment is to measure the variance in gene expression from the pairs (c_i, y_i) (denoted by η^2_{tot}) and to ascribe it to two different sources: first, variability due to the different states of cells ("extrinsic noise", denoted by η^2_{ext}), and second, inherent variability that exists even when the state of cells is fixed ("intrinsic noise", denoted by η^2_{int}). In [2], formulas were provided for estimating $\eta^2_{ext}, \eta^2_{int}$ and η^2_{tot} (hereafter referred to as the ELSS estimates) that were later interpreted in terms of the "law of total variance" in [5]:

$$\eta^2_{int} = \frac{\frac{1}{n}\left(\sum_{i=1}^n \frac{1}{2}(c_i - y_i)^2\right)}{\overline{c} \cdot \overline{y}}, \tag{1}$$

$$\eta^2_{ext} = \frac{\frac{1}{n}\sum_{i=1}^n c_i \cdot y_i - \overline{c} \cdot \overline{y}}{\overline{c} \cdot \overline{y}}, \tag{2}$$

$$\eta^2_{tot} = \frac{\frac{1}{n}\sum_{i=1}^n \frac{1}{2}(c_i^2 + y_i^2) - \overline{c} \cdot \overline{y}}{\overline{c} \cdot \overline{y}}, \tag{3}$$

where $\overline{c} = \frac{1}{n}\sum_{i=1}^n c_i$ and $\overline{y} = \frac{1}{n}\sum_{i=1}^n y_i$.

2. A hierarchical model

Although the work of [5] sheds light on the statistical basis of the ELSS estimators, it does not address questions about their statistical properties, such as bias and accuracy. To analyze these aspects of the estimators we introduce a hierarchical model that provides a formal model for the experiments of [2].

In the rest of the paper, we focus on the numerators of (1,2,3). They are the key components of the formulas and can be viewed as estimators of true variances. We note that lower case letters such as c_i and y_i denote observations not only in the ELSS formulas but throughout our paper; we reserve uppercase letters for random variables.

A hierarchical model for expression of the two reporters in a cell emerges naturally from the assumption that reporter expression, conditioned on the same cellular environment, is represented by independent and identically distributed random variables. To allow each cell to be different from the others, we introduce independent identically distributed random variables Z_i, for $i = 1, \ldots, n$ that represent the environments of cells (as in [5]). We posit that the cellular conditional random variables associated to the two reporters have the same distribution F with mean M_i and variance σ_i^2, both parameters being unique to the i-th cell:

$$C_i|Z_i \sim F(M_i, \Sigma_i^2) \text{ and} \tag{4}$$

$$Y_i|Z_i \sim F(M_i, \Sigma_i^2). \tag{5}$$

Thinking of a two reporter experiment as "random", in the sense that the states of cells $Z_1, \ldots, Z_n$ are random, across cells we have

$$M_i \sim G(\mu, \sigma_\mu^2) \text{ and}$$
$$\Sigma_i^2 \sim H(\sigma^2, \epsilon),$$

where G is the distribution of all the M_is, with mean μ and variance σ_μ^2, and H that of all the Σ_i^2s, with mean σ^2 and variance ϵ. In other words, both the mean and variance of reporter expression level is cell specific and the random variable Σ_i^2 and its mean σ^2 represent the "within-cell" variation as distinguished from the parameter σ_μ^2 which represents the "between-cell" variability in the ANOVA setting.

For any i, the mean of C_i or Y_i is μ, according to the following calculation:

$$E[C_i] = E_{Z_i}[E[C_i|Z_i]] = E[M_i] = \mu. \tag{6}$$

The total variance in C_i (or Y_i) can be calculated using the "law of total variance":

$$Var[C_i] = E_{Z_i}[Var[C_i|Z_i]] + Var_{Z_i}[E[C_i|Z_i]]. \tag{7}$$

Using the notation of the hierarchical model described above, and dropping the subscripts for expectation because they are clear by context, we have, for any i,

$$E[Var[C_i|Z_i]] = \sigma^2 \qquad \text{(within-cell variability; intrinsic noise)}, \tag{8}$$

$$Var[E[C_i|Z_i]] = \sigma_\mu^2 \qquad \text{(between-cell variability; extrinsic noise)}. \tag{9}$$

With this notation equation (7) becomes

$$Var[C_i] = E[Var[C_i|Z_i]] + Var[E[C_i|Z_i]] = \sigma^2 + \sigma_\mu^2 \qquad \text{(total noise)}. \tag{10}$$

This means that the marginal (unconditional) distributions of C_i and Y_i are identical:

$$C_i \sim N(\mu, \sigma^2 + \sigma_\mu^2);$$
$$Y_i \sim N(\mu, \sigma^2 + \sigma_\mu^2).$$

In the next sections, we will derive the estimators for intrinsic and extrinsic noise, and examine the bias and mean squared error (MSE) of each estimator. Specifically, for any estimator S, the MSE of S with respect to the true parameter τ is calculated as follows:

$$\begin{aligned} E[(S-\tau)^2] &= E[S - E[S] + E[S] - \tau]^2 \\ &= E\Big[(S-E[S])^2 + (E[S]-\tau)^2 + 2(S-E[S])(E[S]-\tau)\Big] \\ &= E[S-E[S]]^2 + E[E[S]-\tau]^2 \\ &= Var[S] + (E[S]-\tau)^2, \end{aligned}$$

where $E[S] - \tau$ is the bias of S.

3. Intrinsic noise

Starting with the law of total variance, the within-cell variability $E[Var[C_i|Z_i]]$ for cell i can be written as:

$$\begin{aligned} E[Var[C_i|Z_i]] &= Var[C_i] - Var[E[C_i|Z_i]] \\ &= \frac{1}{2}[Var[C_i] + Var[Y_i]] - Cov[C_i, Y_i] \\ &= \frac{1}{2}[Var[C_i] - 2Cov[C_i, Y_i] + Var[Y_i]] \\ &= \frac{1}{2}Var[C_i - Y_i]. \\ &= \frac{1}{2}\left(E[C_i - Y_i]^2 - (E[C_i - Y_i])^2\right) \end{aligned}$$

This leads to the following unbiased estimator for the intrinsic noise:

$$\begin{aligned} S_{int}^* &= \frac{1}{2(n-1)} \sum_{i=1}^{n} \Big[(C_i - Y_i) - (\bar{C} - \bar{Y})\Big]^2 \\ &= \frac{1}{2(n-1)} \sum_{i=1}^{n} (C_i - Y_i)^2 - \frac{n}{2(n-1)}(\bar{C} - \bar{Y})^2. \end{aligned}$$

To find the estimator that minimizes the MSE, we consider estimators of the following general form

$$S_{int} = \frac{1}{2a}\left(\sum_{1}^{n}(C_i - Y_i)^2 - n(\bar{C} - \bar{Y})^2\right). \tag{11}$$

Assuming normality of the distribution G (i.e., cell-specific means M_i follow a normal distribution), as well as $\mu = 0$ and $\epsilon = 0$, the MSE is given by

$$\begin{aligned} E[S_{int} - \sigma^2]^2 &= Var[S_{int}] + (E[S_{int}] - \sigma^2)^2 \\ &= \frac{1}{2a^2}\left[(2n^2 + \frac{6}{n} - 7)\sigma^4 + 2(\frac{2}{n} - 1)\sigma^2\sigma_\mu^2 + \frac{1}{n}\sigma_\mu^4\right] - 2(n-1)\sigma^4\frac{1}{a} + \sigma^4. \end{aligned}$$

The value of a that minimizes this expression is

$$\begin{aligned} a &= \frac{(2n^3 - 7n + 6)\sigma^4 + 2(2-n)\sigma^2\sigma_\mu^2 + \sigma_\mu^4}{2(n^2 - n)\sigma^4} \\ &= \frac{2n^3 - 7n + 6}{2(n^2 - n)} + \frac{2 - n}{n^2 - n}\frac{\sigma_\mu^2}{\sigma^2} + \frac{1}{2(n^2 - n)}\left(\frac{\sigma_\mu^2}{\sigma^2}\right)^2. \end{aligned}$$

See Appendices A and B for the complete derivation.

The analysis above can be simplified with an additional assumption, namely that $\bar{C} = \bar{Y}$. In some experiments this may be a natural assumption to make, whereas in others the condition is likely to be violated; we comment on this in more detail in the discussion. Here we proceed to note that assuming that $\bar{C} = \bar{Y}$, the estimator (11) simplifies to

$$\tilde{S}_{int} = \frac{1}{2a}\sum_{i=1}^{n}(C_i - Y_i)^2.$$

The unbiased estimator with this form is easily derived by observing that

$$\begin{aligned} E[\tilde{S}_{int}] &= \frac{1}{2a}\sum_{i=1}^{n} E[C_i - Y_i]^2 = \frac{1}{2a}\sum_{i=1}^{n} Var[C_i - Y_i] \\ &= \frac{n}{2a}(2\sigma^2 + 2\sigma_\mu^2 - 2\sigma_\mu^2) = \frac{n}{a}\sigma^2. \end{aligned}$$

Thus, in order for $\tilde{S}_{int}$ to be unbiased the parameter a must be equal to n. The resulting formula is the ELSS formula in (1). This makes clear that the assumption $\bar{C} = \bar{Y}$ underlies the derivation of the ELSS intrinsic noise estimator.

In order to study the mean squared error and derive an estimator that minimizes it, we again assume normality of G. The MSE of S_{int} is then given by

$$\begin{aligned} E[\tilde{S}_{int} - \sigma^2]^2 &= Var[\tilde{S}_{int}] + (E[\tilde{S}_{int}] - \sigma^2)^2 \\ &= \frac{n}{a^2}(3\epsilon + 2\sigma^4) + (\frac{n}{a}\sigma^2 - \sigma^2)^2. \end{aligned}$$

Assuming again that $\mu = 0$ and $\epsilon = 0$, the MSE simplifies to

$$\begin{aligned} E[\tilde{S}_{int} - \sigma^2]^2 &= \frac{2n}{a^2}\sigma^4 + \sigma^4\left(\left(\frac{n}{a}\right)^2 - \frac{2n}{a} + 1\right) \\ &= \frac{n\sigma^4(n+2)}{a^2} - \frac{2n\sigma^4}{a} + \sigma^4, \end{aligned}$$

which is minimized when $a = n+2$ (see Appendices A and C for the complete derivation).

4. Extrinsic noise

To examine estimators for extrinsic noise, we again start with the law of total variance, this time noting that the within-cell variability $Var[E[C_i|Z_i]]$ can be written as:

$$\begin{aligned} Var[E[C_i|Z_i]] &= E[E[C_i|Z_i]^2] - E[E[C_i|Z_i]]^2 \\ &= E[E[C_i|Z_i]E[Y_i|Z_i]] - E[E[C_i|Z_i]]^2 \\ &= E[E[C_iY_i|Z_i]] - E[E[C_i|Z_i]E[E[Y_i|Z_i]] \\ &= E[C_iY_i] - E[C_i]E[Y_i] \\ &= Cov[C_i, Y_i]. \end{aligned} \tag{12}$$

This connection between the extrinsic noise, the law of total variance and the covariance of C_i and Y_i was noted by Hilfinger and Paulsson in [5].

Formula (12) leads to the following unbiased estimator for the extrinsic noise, as it is an unbiased estimator estimator for the covariance:

$$S^*_{ext} = \frac{1}{n-1}\left(\sum_{i=1}^{n} C_iY_i - n\bar{C}\bar{Y}\right).$$

We note that the ELSS estimator (2) uses the scalar $1/n$, which unlike the case of the intrinsic noise estimator (1) leads to a biased estimator in this case.

In order to find the estimator that minimizes the MSE, we consider the following general estimator:

$$S_{ext} = \frac{1}{b}\left(\sum_{i=1}^{n} C_iY_i - n\bar{C}\bar{Y}\right).$$

We again assume that M_i is normal and that $\mu = 0$ and $\epsilon = 0$. The MSE of S_{ext} is

$$\begin{aligned} E[S_{ext} - \sigma^2_\mu]^2 &= \frac{n-1}{b^2}(\sigma^2 + \sigma^2_\mu)^2 + \frac{(n-1)^2}{nb^2}\sigma^4_\mu + \left(\frac{n-1}{b}\sigma^2_\mu - \sigma^2_\mu\right)^2 \\ &= (n-1)(\sigma^2 + \sigma^2_\mu)^2\frac{1}{b^2} + (n-1)^2\left(1 + \frac{1}{n}\right)\sigma^4_\mu\frac{1}{b^2} - 2(n-1)\sigma^4_\mu\frac{1}{b} + \sigma^4_\mu \\ &= \left((n-1)(\sigma^2 + \sigma^2_\mu)^2 + (n-1)^2\left(1 + \frac{1}{n}\right)\sigma^4_\mu\right)\frac{1}{b^2} - 2(n-1)\sigma^4_\mu\frac{1}{b} + \sigma^4_\mu, \end{aligned}$$

which is minimized when

$$\frac{1}{b} = \frac{\sigma^4_\mu}{(\sigma^2 + \sigma^2_\mu)^2 + (n-1)\left(1 + \frac{1}{n}\right)\sigma^4_\mu}, \text{ or equivalently}$$

$$b = (n-1)\left(1 + \frac{1}{n}\right) + \left(\frac{\sigma^2 + \sigma^2_\mu}{\sigma^2_\mu}\right)^2 = (n-1)\left(1 + \frac{1}{n}\right) + \frac{1}{\rho(\mathbf{C}, \mathbf{Y})^2}. \tag{13}$$

It is interesting to note that (13) comprises two parts: the first, $(n-1)(1+\frac{1}{n})$ converges to $n-1$ as $n \to \infty$, while the second, $(\frac{\sigma^2+\sigma^2_\mu}{\sigma^2_\mu})^2$ is equal to $\frac{1}{\rho(\mathbf{C},\mathbf{Y})^2}$ where $\rho(\mathbf{C}, \mathbf{Y})$ is the correlation between vectors $\mathbf{C}$ and $\mathbf{Y}$. See Appendices A and D for more details.

5. Geometric interpretation

Figure 3a of [2] shows a scatterplot of data (c_i, y_i) for an experiment and suggests thinking of intrinsic and extrinsic noise geometrically in terms of projection of the points onto a pair of orthogonal lines. While this geometric interpretation of noise agrees exactly with the ELSS intrinsic noise formula, the interpretation of extrinsic noise is more subtle. Here we complete the picture.

To understand the intuition behind Figure 3a in [2], we have redrawn it in a format that highlights the math (Fig 1). The projection of a point (c_i, y_i) onto the line $y = c$ is the point $(\frac{1}{2}(y_i + c_i), \frac{1}{2}(y_i + c_i))$, shown as the red point in Fig. 1. The intrinsic noise, as estimated by the unbiased estimator (1) is then the mean squared distance from the origin to the points projected onto the line $y = -c$.

The ELSS estimate for the extrinsic noise is the sample covariance. Intuitively, it indicates how the measurements of one reporter track that of the other across cells. The geometric meaning of the sample covariance in Fig. 1 is based on an alternative formulation of sample covariance [3, 4]:

$$\mathrm{Cov}(\mathbf{c}, \mathbf{y}) = \frac{1}{n-1} \sum_{i=1}^{n} (c_i - \bar{c})(y_i - \bar{y}).$$

This formulation of the sample covariance has the interpretation of being an average of the signed area of triangles associated to pairs of points, and is very different from what might be considered at first glance an appropriate anology to intrinsic noise, namely the sample variance along the line $y = c$.

The estimate corresponding to the sample variance of the projected points along the line $y = c$, using as a mean the projected centroid $(\frac{\bar{c}+\bar{y}}{2}, \frac{\bar{c}+\bar{y}}{2})$ which is shown as the green point in Fig. 1, turns out to be biased by an amount equal to the total noise. Using

$$\begin{aligned}
\tilde{S}^*_{ext} &= \frac{1}{n-1} \sum_{i=1}^{n} 2 \left(\frac{1}{2}(Y_i - \bar{Y} + C_i - \bar{C}) \right)^2 \\
&= \frac{1}{2(n-1)} \sum_{i=1}^{n} \left((C_i + Y_i)^2 - (\bar{C} + \bar{Y})^2 \right)
\end{aligned}$$

the bias is

$$\begin{aligned}
E[\tilde{S}^*_{ext}] - \sigma^2_\mu &= \frac{1}{2} Var[C_i + Y_i] - \sigma^2_\mu \\
&= \frac{1}{2} \left(Var[C_i] + Var[Y_i] + 2Cov[C_i, Y_i] \right) - \sigma^2_\mu \\
&= \frac{1}{2} \left(2(\sigma^2 + \sigma^2_\mu) + 2\sigma^2_\mu \right) - \sigma^2_\mu = \sigma^2 + \sigma^2_\mu
\end{aligned}$$

which is the true total noise.

The above calculation also shows that if the intrinsic and extrinsic noise are both estimated as variances along the projections to the lines $y = -c$ and $y = c$ respectively, then the total noise will be overestimated by a factor of two.

In summary, the caption to Figure 3a in [2] is completely accurate in stating that "Spread of points perpendicular to the diagonal line on which CFP and YFP intensities

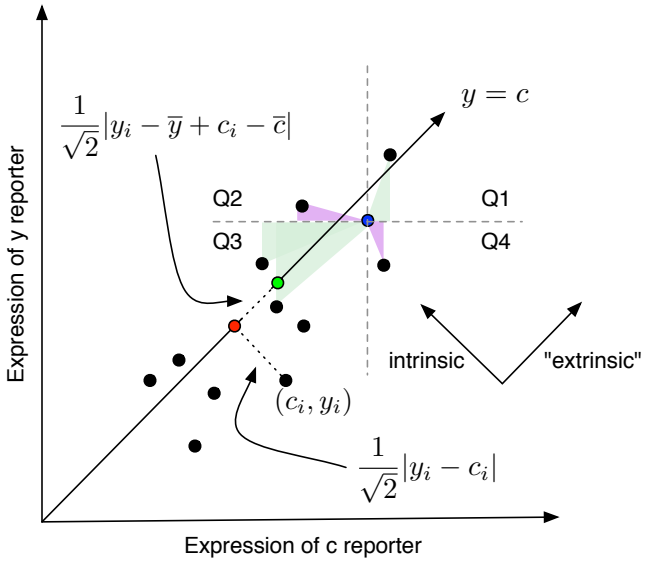

FIGURE 1. Geometric interpretation of intrinsic and extrinsic noise. The intrinsic noise, or the within-cell variability, is the variance of the points projected to the line $y = -c$, which is perpendicular to $y = c$. In other words, it is the average of the squared lengths $\frac{1}{2}(y_i - c_i)^2$. The red point is the projection of point (c_i, y_i) onto the line $y = c$. The green point is the centroid. See the main text for additional detail. The extrinsic noise, or the between-cell variability, is the sample covariance between c_i and y_i. The colored triangles around the blue point illustrate the geometric interpretation of the sample covariance: it is the average (signed) area of triangles formed by pairs of data points: green triangles in Q1 and Q3 (some not shown) represent a positive contribution to the covariance, whereas the magenta triangles in Q2 and Q4 a negative contribution. Since most data points lie in the 1st (Q1) and 3rd (Q3) quadrants relative to the blue point, most of the contribution involving the blue point is positive. Similarly, since most pairs of data points can be connected by a positively signed line, their positive contribution will result in a positive covariance. In [2] the direction along the line $y = c$ is labeled extrinsic, which makes sense in terms of the intuition for positive sample covariance. However we have placed that label "extrinsic" in quotes because the extrinsic noise estimator corresponding directly to the sample variance for points projected onto the line $y = c$ (in analogy with intrinsic noise) is heavily biased and not usable in practice.

are equal corresponds to intrinsic noise, whereas spread parallel to this line is increased by extrinsic noise." However the geometric interpretation of covariance makes precise exactly *how* an increase in extrinsic noise relates to the spread of points in the direction of the line $y = c$.

6. Practical considerations

6.1. **Data normalization.** Our hierarchical model, as well as the ANOVA interpretation, is consistent with the model in Elowitz *et al.* [2]; both models assume that within each cell there are two distributions for the expression of the two reporter genes and that they have the same true mean and true variance. With the normality assumption, this means that the two reporters have identical distributions. Elowitz *et al.* measured the single-color distributions of strains that contained lac-repressible promoter pairs, which verified that this was a reasonable assumption in the case of cyan fluorescent protein (CFP) and yellow fluorescent protein (YFP) in their experiment.

Other studies have adapted this system and used other reporter combinations that may have markedly different distributions. For example, Yang *et al.* [7] used CFP and mCherry with vastly different ranges of intensity values: whereas CFP varied from 0 to 6000 (arbitrary units; i.e., a.u.), mCherry could vary from 0 to 9000 (a.u.); see Fig. 3a from their paper. In contrast, another study [6] normalized the two reporters used in their experiment (ZsGreen and mCherry) to have the same mean. However, the variances, or more generally, the two distributions, also need to be the same. Since the decomposition of the total noise depends on the assumption that both reporters in the same cellular environment have similar variance (see (4) and (5)), we recommend that in general a quantile normalization which normalizes the reporter measurements to identical distributions be performed before the calculations of noise components. Such a normalization procedure is standard in many settings requiring similar assumptions.

6.2. **Optimal estimators for intrinsic and extrinsic noise.** We have derived the estimators that are optimal for minimizing bias or the MSE (summarized in Table 1). The ELSS estimator in (1) is in fact a special case of the general estimator under the assumption that $\bar{C} = \bar{Y}$, and is appropriate for data that are normalized to have the same sample mean (i.e., $\bar{c} = \bar{y}$). In [2], the intensities of the two reporters were normalized to have mean 1. In the case where the assumption of equal reporter means does not hold, the general estimator is more suitable.

Similar to the estimators for the intrinsic noise, we derived two estimators for extrinsic noise, optimized for bias and for MSE respectively (Table 1).

The sample size n is the leading term in the denominator of all the optimal (in either the bias or MSE sense) intrinsic and extrinsic noise estimators. As a result, the unbiased estimator has the same form as the min-MSE estimator for large n (Table 1). For extrinsic noise, the general estimates converge to the ELSS estimate (Table 1). For intrinsic noise, assuming $\bar{c} = \bar{y}$, the ELSS estimate is optimal for bias and MSE for large n and optimal for bias at small n. Indeed, in [2], typical values for n are greater than 100, making the ELSS formulas suitable for the analyses performed (with the assumption of equal mean satisfied). However, our derivations indicate that the two types of noise can be estimated using fewer cells.

As a general rule we recommend computing the inverse squared correlation between the c_i and y_i values and applying a correction if it is comparable (up to a small factor) to the sample size.

TABLE 1. Estimators for intrinsic and extrinsic noise

	Exact Estimator for Small n		Large n
	Minimizing Bias (Unbiased)	Minimizing MSE	
Intrinsic noise			
General	$\frac{1}{2(n-1)}\left[\sum_1^n (C_i-Y_i)^2 - n(\bar{C}-\bar{Y})^2\right]$	$\frac{1}{2a}\left[\sum_1^n (C_i-Y_i)^2 - n(\bar{C}-\bar{Y})^2\right]$, where $a = \frac{2n^3-7n+6}{2(n^2-n)} + \frac{2-n}{n^2-n}\frac{\sigma_\mu^2}{\sigma^2} + \frac{1}{2(n^2-n)}\left(\frac{\sigma_\mu^2}{\sigma^2}\right)^2$	$\frac{1}{2n}\left[\sum_1^n (C_i-Y_i)^2 - n(\bar{C}-\bar{Y})^2\right]$
Assuming $\bar{C} = \bar{Y}$	$\frac{1}{2n}\sum_{i=1}^n (C_i-Y_i)^2$ (ELSS estimator)	$\frac{1}{2(n+2)}\sum_{i=1}^n (C_i-Y_i)^2$	$\frac{1}{2n}\sum_{i=1}^n (C_i-Y_i)^2$ (ELSS estimator)
Extrinsic noise			
General	$\frac{1}{n-1}(\sum_{i=1}^n C_iY_i - n\bar{C}\bar{Y})$	$\frac{\sigma_\mu^4}{(\sigma^2+\sigma_\mu^2)^2+(n-1)\left(1+\frac{1}{n}\right)\sigma_\mu^4}\left(\sum_{i=1}^n C_iY_i - n\bar{C}\bar{Y}\right)$	$\frac{1}{n}(\sum_{i=1}^n C_iY_i - n\bar{C}\bar{Y})$ (ELSS estimator)

6.3. **Assessing the ratio of extrinsic to intrinsic noise from sample correlation.** We have seen that the proportion of the between-cell variability to total variability is the correlation $\rho(\mathbf{C}, \mathbf{Y})$. This leads to a simple approach for estimating the relative magnitude of the two types of noise: one can compute the sample correlation of the expression of the two reporters, $\rho(\mathbf{c}, \mathbf{y})$, and the ratio of extrinsic to intrinsic noise is then estimated by $\rho(\mathbf{c}, \mathbf{y})/[1 - \rho(\mathbf{c}, \mathbf{y})]$. For example, in Elowitz et al [2], the sample correlation $\rho(\mathbf{c}, \mathbf{y})$ is roughly 0.7, which implies that about 70% of the total noise is extrinsic noise and the ratio of extrinsic to intrinsic noise is 2.33.

7. Acknowledgments

This project began as a result of discussion during a journal club meeting of Jonathan Pritchard's group that A.F. was attending. We thank Michael Elowitz and Peter Swain for facilitating reanalysis of the data from [2]. A.F. was partially supported by K99HG007368 (NIH/NHGRI). L.P. was partially supported by NIH grants R01 HG006129 and R01 DK094699.

References

[1] Clive G. Bowsher and Peter S. Swain. Identifying sources of variation and the flow of information in biochemical networks. *Proceedings of the National Academy of Sciences*, 109, 20, E1320-E1328, 2012.

[2] Michael B. Elowitz, Arnold J. Levine, Eric D. Siggia and Peter S. Swain. Stochastic gene expression in a single cell. *Science*, 297, 1183-1186, 2002.

[3] Kevin Hayes. A geometrical interpretation of an alternative formula for the sample covariance. *The American Statistician*, 65, 2, 110-112, 2011.

[4] Peter M. Heffernan. New measures of spread and a simpler formula for the normal distribution. *The American Statistician*, 42, 2, 100-102, 1988.

[5] Andreas Hilfinger and Johan Paulsson. Separating intrinsic from extrinsic fluctuations in dynamic biological systems. *Proceedings of the National Academy of Sciences*, 108, 29, 12167-12172, 2011.

[6] Jörn M. Schmiedel, Sandy L. Klemm, Yannan Zheng, Apratim Sahay, Nils Blüthgen, Debora S. Marks, Alexander van Oudenaarden. MicroRNA control of protein expression noise. *Science*, 348, 128-132, 2015.

[7] Sora Yang, Seunghyeon Kim, Yu Rim Lim, Cheolhee Kim, Hyeong Jeon An, Ji-Hyun Kim, Jaeyoung Sung and Nam Ki Lee. Contribution of RNA polymerase concentration variation to protein expression noise. *Nature Communications*, 5, 4761, 2014.

Appendix A. Moments of M_i and C_i under normality

Assuming that $M_i \sim N(\mu, \sigma_\mu^2)$, we have

$$E[M_i - \mu]^3 = 0;$$
$$E[M_i - \mu]^4 = 3\sigma_\mu^4.$$

We can compute the third and fourth moments of M_i as follows:

$$\begin{aligned} E[M_i - \mu]^3 &= E[M_i^2 + \mu^2 - 2M_i\mu)(M_i - \mu] \\ &= E[M_i^3 - 2M_i^2\mu + M_i\mu^2 - M_i^2\mu - \mu^3 + 2M_i\mu^2] \\ &= E[M_i^3 - 3M_i^2\mu + 3M_i\mu^2 - \mu^3] \\ &= E[M_i^3] - 3\mu(\sigma_\mu^2 + \mu^2) + 3\mu^3 - \mu^3 \\ &= E[M_i^3] - 3\mu\sigma_\mu^2 - \mu^3, \end{aligned}$$

which gives

$$E[M_i^3] = 3\mu\sigma_\mu^2 + \mu^3.$$

$$\begin{aligned} E[M_i - \mu]^4 &= E[M_i^2 - 2M_i\mu + \mu^2]^2 \\ &= E[M_i^4 + \mu^4 + 4M_i^2\mu^2 + 2M_i^2\mu^2 - 4M_i^3\mu - 4M_i\mu^3] \\ &= E[M_i^4 + \mu^4 + 6M_i^2\mu^2 - 4M_i^3\mu - 4M_i\mu^3] \\ &= E[M_i^4] + \mu^4 + 6\mu^2(\sigma_\mu^2 + \mu^2) - 4\mu(3\mu\sigma_\mu^2 + \mu^3) - 4\mu^4 \\ &= E[M_i^4] + \mu^4 + 6\mu^2\sigma_\mu^2 + 6\mu^4 - 12\mu^2\sigma_\mu^2 - 4\mu^4 - 4\mu^4 \\ &= E[M_i^4] - 6\mu^2\sigma_\mu^2 - \mu^4, \end{aligned}$$

which gives

$$E[M_i^4] = 3\sigma_\mu^4 + 6\mu^2\sigma_\mu^2 + \mu^4.$$

For the random variable C_i, since $\Sigma_i^2 \sim H(\sigma^2, \epsilon)$, such that

$$E[\Sigma_i^2] = \sigma^2;$$
$$\mathrm{Var}[\Sigma_i^2] = \epsilon,$$

we have

$$\begin{aligned} E[C_i^4] &= E[E[C_i^4|Z_i]] \\ &= E[3\Sigma_i^4 + 6M_i^2\Sigma_i^2 + M_i^4] \\ &= 3(\epsilon + \sigma^4) + 6(\sigma_\mu^2 + \mu^2)\sigma^2 + 3\sigma_\mu^4 + 6\mu^2\sigma_\mu^2 + \mu^4 \\ &= 3\epsilon + 3\sigma^4 + 6\sigma_\mu^2\sigma^2 + 6\mu^2\sigma^2 + 3\sigma_\mu^4 + 6\mu^2\sigma_\mu^2 + \mu^4. \end{aligned}$$

Further assuming that $\mu = 0$, i.e., the means are all 0, and that $\epsilon = 0$, which means that the variability is the same across cells, we have

$$E[M_i^3] = 0$$
$$E[M_i^4] = 3\sigma_\mu^4;$$

and

$$E[C_i^3] = 0$$
$$E[C_i^4] = 3(\sigma^2 + \sigma_\mu^2)^2.$$

Appendix B. MSE of the general intrinsic noise estimator

The general form of the estimator for intrinsic noise is

$$S = \frac{1}{2a}\left(\sum_1^n (C_i - Y_i)^2 - n(\bar{C} - \bar{Y})^2\right).$$

Thus

$$Var[S] = \frac{1}{4a^2}\left(Var\left[\sum (C_i - Y_i)^2\right] + n^2 Var[(\bar{C} - \bar{Y})^2\right] - 2nCov\left[\sum (C_i - Y_i)^2, (\bar{C} - \bar{Y})^2\right]\right).$$

Below we will assume normality, as well as $\mu = 0$ and $\epsilon = 0$, to facilitate the derivation.

First, we note that

$$\begin{aligned} Var[(\bar{C} - \bar{Y})^2] &= Var[\bar{C}^2 - 2\bar{C}\bar{Y} + \bar{Y}^2] \\ &= Var[\bar{C}^2] + 4Var[\bar{C}\bar{Y}] + Var[\bar{Y}^2] - 4Cov[\bar{C}^2, \bar{C}\bar{Y}] - 4Cov[\bar{Y}^2, \bar{C}\bar{Y}] + 2Cov[\bar{C}^2, \bar{Y}^2]. \end{aligned}$$

$$\begin{aligned} Var[\bar{C}^2] &= Var\left[\frac{C_1 + \cdots + C_n}{n} \cdot \frac{C_1 + \cdots + C_n}{n}\right] \\ &= \frac{1}{n^4} Var\left[\sum C_k^2 + \sum_{i \neq j} C_i C_j\right] \\ &= \frac{1}{n^4}\left(Var \sum C_k^2 + Var[\sum_{i \neq j} C_i C_j] + 2Cov[\sum C_k^2, \sum_{i \neq j} C_i C_j]\right) \\ &= \frac{1}{n^4}\left(2n(\sigma^2 + \sigma_\mu^2)^2 + n(n-1)(\sigma^2 + \sigma_\mu^2)^2 + 0\right) \\ &= \frac{n+1}{n^3}(\sigma^2 + \sigma_\mu^2)^2. \end{aligned}$$

This is because

$$\begin{aligned} Var\left[\sum_{i \neq j} C_i C_j\right] &= \sum_{i \neq j} Var[C_i C_j] \\ &= \sum_{i \neq j}\left(EC_i^2 C_j^2 - (EC_i C_j)^2\right) \\ &= \sum_{i \neq j}\left((\sigma^2 + \sigma_\mu^2)^2 - 0\right) \\ &= n(n-1)(\sigma^2 + \sigma_\mu^2)^2. \end{aligned}$$

Additionally,

$$\begin{aligned} Var[\bar{C}\bar{Y}] &= \frac{1}{n^2} Var[n\bar{C}\bar{Y}] \\ &= \frac{1}{n^2}\left((\sigma^2+\sigma_\mu^2)^2 + \frac{\sigma_\mu^4}{n}\right) \\ &= \frac{1}{n^2}(\sigma^2+\sigma_\mu^2)^2 + \frac{\sigma_\mu^4}{n^3}. \end{aligned}$$

$$\begin{aligned} Cov[\bar{C}^2, \bar{C}\bar{Y}] &= \frac{1}{n^4} Cov\left[\sum C_k^2 + \sum_{i\neq j} C_i C_j, \sum C_l Y_l + \sum_{m\neq r} C_m C_r\right] \\ &= \frac{1}{n^4}\Bigg(Cov\left[\sum C_k^2, \sum C_l Y_l\right] + Cov\left[\sum C_k^2, \sum_{m\neq r} C_m C_r\right] \\ &\quad + Cov\left[\sum_{i\neq j} C_i C_j, \sum C_l Y_l\right] + Cov\left[\sum_{i\neq j} C_i C_j, \sum_{m\neq r} C_m C_r\right]\Bigg). \end{aligned}$$

$$\begin{aligned} Cov\left[\sum C_k^2, \sum C_l Y_l\right] &= Cov\left[\sum C_k^2, \sum C_k Y_k\right] \\ &= \sum (E[C_k^3 Y_k] - E[C_k^2]E[C_k Y_k]) \\ &= \sum\left[3\sigma^2\sigma_\mu^2 + 3\sigma_\mu^4 - (\sigma^2+\sigma_\mu^2)\sigma_\mu^2\right] \\ &= 2n\sigma_\mu^2(\sigma^2+\sigma_\mu^2). \end{aligned}$$

For $Cov\left[\sum C_k^2, \sum_{m\neq r} C_m C_r\right]$, since

$$Cov[C_i^2, C_i Y_j] = E[C_i^3 Y_j] - E[C_i^2]E[C_i Y_j] = 0$$

and

$$Cov[C_i^2, C_j Y_k] = E[C_i^2 C_j Y_k] - E[C_i^2]E[C_j Y_k] = 0,$$

we have

$$Cov\left[\sum C_k^2, \sum_{m\neq r} C_m C_r\right] = 0.$$

For $Cov\left[\sum_{i\neq j} C_i C_j, \sum C_l Y_l\right]$, since

$$Cov[C_i C_j, C_i Y_i] = E[C_i^2 Y_i C_j] - E[C_i C_j]E[C_i Y_i] = 0$$

and

$$Cov[C_k C_l, C_i Y_i] = E[C_k C_l C_i Y_i] - E[C_k C_l]E[C_i Y_i] = 0,$$

we have

$$Cov\left[\sum_{i\neq j} C_iC_j, \sum C_lY_l\right] = 0.$$

Additionally,

$$\begin{aligned} Cov\left[\sum_{i\neq j} C_iC_j, \sum_{m\neq r} C_mC_r\right] &= \sum_{i,j,m,r} Cov[C_iC_j, C_mC_r] \\ &= \sum_{i\neq j} Cov[C_iC_j, C_iC_j] \\ &= \sum_{i\neq j} Var[C_iC_j] \\ &= n(n-1)(\sigma^2+\sigma_\mu^2)^2. \end{aligned}$$

Therefore,

$$Cov[\bar{C}^2, \bar{C}\bar{Y}] = \frac{2}{n^3}\sigma_\mu^2(\sigma^2+\sigma_\mu^2) + \frac{n-1}{n^3}(\sigma^2+\sigma_\mu^2)^2.$$

Furthermore,

$$\begin{aligned} Cov[\bar{C}^2, \bar{Y}^2] &= \frac{1}{n^4} Cov\left[\sum C_k^2 + \sum_{i\neq j} C_iC_j, \sum Y_l^2 + \sum_{m\neq r} Y_mY_r\right] \\ &= \frac{1}{n^4}\Bigg(Cov\left[\sum C_k^2, \sum Y_l^2\right] + Cov\left[\sum C_k^2, \sum_{m\neq r} Y_mY_r\right] \\ &\quad + Cov\left[\sum Y_l^2, \sum_{i\neq j} C_iC_j\right] + Cov\left[\sum_{i\neq j} C_iC_j, \sum_{m\neq r} Y_mY_r\right]\Bigg). \end{aligned}$$

In the expression above,

$$Cov\left[\sum C_k^2, \sum Y_l^2\right] = 2n\sigma_\mu^4;$$

$$Cov\left[\sum C_k^2, \sum_{m\neq r} Y_mY_r\right] = Cov\left[\sum Y_l^2, \sum_{i\neq j} C_iC_j\right] = 0;$$

$$\begin{aligned} Cov\left[\sum_{i\neq j} C_iC_j, \sum_{m\neq r} Y_mY_r\right] &= \sum_{i\neq j} Cov[C_iC_j, Y_iY_j] \\ &= \sum_{i\neq j}(E[C_iC_jY_iY_j] - E[C_iC_j]E[Y_iY_j]) \\ &= \sum_{i\neq j}(E[C_iY_i]E[C_jY_j] - 0) \\ &= n(n-1)\sigma_\mu^4. \end{aligned}$$

Then we have

$$\begin{aligned} Cov[\bar{C}^2, \bar{Y}^2] &= \frac{1}{n^4}\bigg(2n\sigma_\mu^4 + n(n-1)\sigma_\mu^4\bigg) \\ &= \frac{n+1}{n^3}\sigma_\mu^4. \end{aligned}$$

Putting the terms together, we have

$$\begin{aligned} Var[\bar{C}-\bar{Y}]^2 &= Var[\bar{C}^2] + 4Var[\bar{C}\bar{Y}] + Var[\bar{Y}^2] - 4Cov[\bar{C}^2, \bar{C}\bar{Y}] - 4Cov[\bar{Y}^2, \bar{C}\bar{Y}] + 2Cov[\bar{C}^2, \bar{Y}^2] \\ &= \frac{2(n+1)}{n^3}(\sigma^2+\sigma_\mu^2)^2 + \frac{4}{n^2}(\sigma^2+\sigma_\mu^2)^2 + \frac{4\sigma_\mu^4}{n^3} - \frac{16}{n^3}\sigma_\mu^2(\sigma^2+\sigma_\mu^2) - \frac{8(n-1)}{n^3}(\sigma^2+\sigma_\mu^2)^2 \\ &\quad + \frac{2(n+1)}{n^3}\sigma_\mu^4 \\ &= \frac{2}{n^3}\bigg((6-n)(\sigma^2+\sigma_\mu^2)^2 - 8\sigma_\mu^2(\sigma^2+\sigma_\mu^2) + (n+3)\sigma_\mu^4\bigg) \\ &= \frac{2}{n^3}\left((6-n)\sigma^4 + (4-2n)\sigma^2\sigma_\mu^2 + \sigma_\mu^4\right). \end{aligned}$$

Next, we note that

$$\begin{aligned} Cov\left[\sum(C_i-Y_i)^2, (\bar{C}-\bar{Y})^2\right] &= \sum Cov\left[(C_i-Y_i)^2, (\bar{C}-\bar{Y})^2\right] \\ &= \sum\bigg(E[(C_i^2 - 2C_iY_i + Y_i^2)(\bar{C}^2 - 2\bar{C}\bar{Y} + \bar{Y}^2)] \\ &\quad - E[(C_i^2 - 2C_iY_i + Y_i^2)]E[(\bar{C}^2 - 2\bar{C}\bar{Y} + \bar{Y}^2)]\bigg), \end{aligned}$$

where

$$\begin{aligned} &E[(C_i^2 - 2C_iY_i + Y_i^2)(\bar{C}^2 - 2\bar{C}\bar{Y} + \bar{Y}^2)] \\ &= E\bigg[C_i^2\bar{C}^2 - 2C_iY_i\bar{C}^2 + Y_i^2\bar{C}^2 - 2C_i^2\bar{C}\bar{Y} + 4C_iY_i\bar{C}\bar{Y} - 2Y_i^2\bar{C}\bar{Y} + C_i^2\bar{Y}^2 - 2C_iY_i\bar{Y}^2 + Y_i^2\bar{Y}^2\bigg], \end{aligned}$$

and

$$E[C_i^2 - 2C_iY_i + Y_i^2] = 2(\sigma^2+\sigma_\mu^2) - 2\sigma_\mu^2 = 2\sigma^2,$$

$$E[\bar{C}^2 - 2\bar{C}\bar{Y} + \bar{Y}^2] = \frac{2}{n}(\sigma^2+\sigma_\mu^2) - \frac{2}{n}\sigma_\mu^2 = \frac{2\sigma^2}{n}.$$

$$
\begin{aligned}
E[C_i^2\bar{C}^2] &= \frac{1}{n^2}E[C_i^2\left(\sum C_k^2 + \sum_{i\neq j} C_iC_j\right)] \\
&= \frac{1}{n^2}\left(E[C_i^4] + \sum_{k\neq i} E[C_i^2]E[C_k^2] + \sum_{i\neq j} E[C_k^2C_iC_j]\right) \\
&= \frac{1}{n^2}\left[3(\sigma^2+\sigma_\mu^2)^2 + (n-1)(\sigma^2+\sigma_\mu^2)^2 + 0\right] \\
&= \frac{n+2}{n^2}(\sigma^2+\sigma_\mu^2)^2.
\end{aligned}
$$

$$
\begin{aligned}
E[C_iY_i\bar{C}^2] &= E\left[C_iY_i\frac{\sum C_j^2 + \sum_{k\neq l} C_kC_l}{n^2}\right] \\
&= \frac{1}{n^2}\left(E[C_iY_iC_i^2] + \sum_{j\neq i} E[C_iY_iC_j^2] + \sum_{k\neq l} E[C_iY_iC_kC_l]\right) \\
&= \frac{1}{n^2}\left(3(\sigma^2\sigma_\mu^2+\sigma_\mu^4) + (n-1)(\sigma^2\sigma_\mu^2+\sigma_\mu^4) + 0\right) \\
&= \frac{n+2}{n^2}\sigma_\mu^2(\sigma^2+\sigma_\mu^2).
\end{aligned}
$$

$$
\begin{aligned}
E[Y_i^2\bar{C}^2] &= E\left[Y_i^2\frac{\sum C_j^2 + \sum_{k\neq l} C_kC_l}{n^2}\right] \\
&= \frac{1}{n^2}\left(E[Y_i^2C_i^2] + \sum_{j\neq i} E[Y_i^2C_j^2] + \sum_{k\neq l} E[Y_i^2C_kC_l]\right) \\
&= \frac{1}{n^2}\left((\sigma^2+\sigma_\mu^2)^2 + 2\sigma_\mu^4 + (n-1)(\sigma^2+\sigma_\mu^2)^2 + 0\right) \\
&= \frac{1}{n^2}\left(n(\sigma^2+\sigma_\mu^2)^2 + 2\sigma_\mu^4\right).
\end{aligned}
$$

$$
\begin{aligned}
E[C_i^2\bar{C}\bar{Y}] &= \frac{1}{n^2}\left(E[C_i^2C_iY_i] + \sum_{j\neq i} E[C_i^2C_jY_j] + \sum_{k\neq l} E[C_i^2C_kY_l]\right) \\
&= \frac{1}{n^2}\left(3(\sigma^2\sigma_\mu^2+\sigma_\mu^4) + (n-1)(\sigma^2\sigma_\mu^2+\sigma_\mu^4) + 0\right) \\
&= \frac{n+2}{n^2}\sigma_\mu^2(\sigma^2+\sigma_\mu^2).
\end{aligned}
$$

$$E[C_iY_i\bar{C}\bar{Y}] = \frac{1}{n^2}\bigg(E[C_i^2Y_i^2] + \sum_{j\neq i}E[C_iY_iC_jY_j] + \sum_{k\neq l}E[C_iY_iC_kY_l]\bigg)$$
$$= \frac{1}{n^2}\bigg((\sigma^2+\sigma_\mu^2)^2 + 2\sigma_\mu^4 + (n-1)\sigma_\mu^4 + 0\bigg)$$
$$= \frac{1}{n^2}\bigg((\sigma^2+\sigma_\mu^2)^2 + (n+1)\sigma_\mu^4\bigg).$$

Additionally,

$$E[Y_i^2\bar{C}\bar{Y}] = E[C_i^2\bar{C}\bar{Y}] = \frac{n+2}{n^2}\sigma_\mu^2(\sigma^2+\sigma_\mu^2);$$
$$E[C_i^2\bar{Y}^2] = E[Y_i^2\bar{C}^2] = \frac{1}{n^2}\bigg(n(\sigma^2+\sigma_\mu^2)^2 + 2\sigma_\mu^4\bigg);$$
$$E[C_iY_i\bar{Y}^2] = E[C_iY_i\bar{C}^2] = \frac{n+2}{n^2}\sigma_\mu^2(\sigma^2+\sigma_\mu^2);$$
$$E[Y_i^2\bar{Y}^2] = E[C_i^2\bar{C}^2] = \frac{n+2}{n^2}(\sigma^2+\sigma_\mu^2)^2.$$

Therefore,

$$E[(C_i^2 - 2C_iY_i + Y_i^2)(\bar{C}^2 - 2\bar{C}\bar{Y} + \bar{Y}^2)]$$
$$= E\bigg[C_i^2\bar{C}^2 - 2C_iY_i\bar{C}^2 + Y_i^2\bar{C}^2 - 2C_i^2\bar{C}\bar{Y} + 4C_iY_i\bar{C}\bar{Y} - 2Y_i^2\bar{C}\bar{Y} + C_i^2\bar{Y}^2 - 2C_iY_i\bar{Y}^2 + Y_i^2\bar{Y}^2\bigg]$$
$$= \frac{2(n+2)}{n^2}(\sigma^2+\sigma_\mu^2)^2 - \frac{4(n+2)}{n^2}\sigma_\mu^2(\sigma^2+\sigma_\mu^2) + \frac{2}{n^2}\bigg(n(\sigma^2+\sigma_\mu^2)^2 + 2\sigma_\mu^4\bigg)$$
$$- \frac{4(n+2)}{n^2}\sigma_\mu^2(\sigma^2+\sigma_\mu^2) + \frac{4}{n^2}\bigg((\sigma^2+\sigma_\mu^2)^2 + (n+1)\sigma_\mu^4\bigg)$$
$$= \frac{4(n+2)\sigma^4}{n^2}.$$

So we have

$$Cov\bigg[\sum(C_i-Y_i)^2, (\bar{C}-\bar{Y})^2\bigg] = \sum Cov\bigg[(C_i-Y_i)^2, (\bar{C}-\bar{Y})^2\bigg]$$
$$= \sum\bigg(E(C_i^2 - 2C_iY_i + Y_i^2)(\bar{C}^2 - 2\bar{C}\bar{Y} + \bar{Y}^2)$$
$$- E(C_i^2 - 2C_iY_i + Y_i^2)E(\bar{C}^2 - 2\bar{C}\bar{Y} + \bar{Y}^2)\bigg)$$
$$= n\bigg(\frac{4(n+2)\sigma^4}{n^2} - 2\sigma^2\frac{2\sigma^2}{n}\bigg)$$
$$= \frac{8\sigma^4}{n}.$$

The variance of the estimator is then

$$\begin{aligned}Var[S] &= \frac{1}{4a^2}\left(Var\left[\sum(C_i - Y_i)^2\right] + n^2 Var[\bar{C} - \bar{Y}]^2 - 2nCov\left[\sum(C_i - Y_i)^2, (\bar{C} - \bar{Y})^2\right]\right) \\ &= \frac{1}{4a^2}\left(8n\sigma^4 + \frac{2}{n}\left((6-n)\sigma^4 + (4-2n)\sigma^2\sigma_\mu^2 + \sigma_\mu^4\right) - 16\sigma^4\right) \\ &= \frac{1}{2a^2}\left(4n\sigma^4 + \frac{1}{n}\left((6-n)\sigma^4 + (4-2n)\sigma^2\sigma_\mu^2 + \sigma_\mu^4\right) - 8\sigma^4\right).\end{aligned}$$

The expectation of the estimator is

$$E[S] = \frac{1}{2a}\left(\sum E[C_i - Y_i]^2 - nE[\bar{C} - \bar{Y}]^2\right),$$

where

$$\begin{aligned}E[(C_i - Y_i)^2] &= Var[C_i - Y_i] \\ &= Var[C_i] + Var[Y_i] - 2Cov[C_i, Y_i] \\ &= 2(\sigma^2 + \sigma_\mu^2) - 2\sigma_\mu^2 = 2\sigma^2,\end{aligned}$$

and

$$\begin{aligned}E[(\bar{C} - \bar{Y})^2] &= Var[\bar{C} - \bar{Y}] \\ &= Var[\bar{C}] + Var[\bar{Y}] - 2Cov[\bar{C}, \bar{Y}] \\ &= \frac{2}{n}(\sigma^2 + \sigma_\mu^2) - \frac{2}{n}\sigma_\mu^2 = \frac{2\sigma^2}{n}.\end{aligned}$$

Hence,

$$E[S] = \frac{1}{2a}(2n\sigma^2 - 2\sigma^2) = \frac{n-1}{a}\sigma^2.$$

The MSE of the estimator is then

$$\begin{aligned}E[(S - \sigma^2)^2] &= Var[S] + (E[S] - \sigma^2)^2 \\ &= \frac{1}{2a^2}\left(4n\sigma^4 + \frac{1}{n}\left((6-n)\sigma^4 + (4-2n)\sigma^2\sigma_\mu^2 + \sigma_\mu^4\right) - 8\sigma^4\right) \\ &\quad + \left(\frac{n-1}{a} - 1\right)^2\sigma^4 \\ &= \frac{1}{2a^2}\left(4n\sigma^4 + \frac{1}{n}\left((6-n)\sigma^4 + (4-2n)\sigma^2\sigma_\mu^2 + \sigma_\mu^4\right) - 8\sigma^4 + 2(n-1)^2\sigma^4\right) \\ &\quad - 2(n-1)\sigma^4\frac{1}{a} + \sigma^4 \\ &= \frac{1}{2a^2}\left((2n^2 + \frac{6}{n} - 7)\sigma^4 + 2(\frac{2}{n} - 1)\sigma^2\sigma_\mu^2 + \frac{1}{n}\sigma_\mu^4\right) - 2(n-1)\sigma^4\frac{1}{a} + \sigma^4.\end{aligned}$$

The value of a that minimizes this MSE is

$$\begin{aligned} a &= \frac{(2n^3-7n+6)\sigma^4+2(2-n)\sigma^2\sigma_\mu^2+\sigma_\mu^4}{2(n^2-n)\sigma^4} \\ &= \frac{2n^3-7n+6}{2(n^2-n)}+\frac{2-n}{n^2-n}\frac{\sigma_\mu^2}{\sigma^2}+\frac{1}{2(n^2-n)}\left(\frac{\sigma_\mu^2}{\sigma^2}\right)^2. \end{aligned}$$

Appendix C. Calculating $Var[\tilde{S}_{int}]$

$$\begin{aligned} Var[\tilde{S}_{int}] &= \frac{1}{4a^2}Var\left[\sum_{i=1}^n (C_i-Y_i)^2\right] \\ &= \frac{1}{4a^2}Var\left[\sum_{i=1}^n \left(C_i^2+Y_i^2-2C_iY_i\right)\right] \\ &= \frac{1}{4a^2}Var\left[\sum_{i=1}^n C_i^2+\sum_{i=1}^n Y_i^2-2\sum_{i=1}^n C_i,Y_i\right] \\ &= \frac{1}{4a^2}\Bigg(Var\left[\sum_{i=1}^n C_i^2\right]+Var\left[\sum_{i=1}^n Y_i^2\right]+4Var\left[\sum_{i=1}^n C_iY_i\right]+2Cov\left[\sum_{i=1}^n C_i^2,\sum_{i=1}^n Y_i^2\right] \\ &\quad -4Cov\left[\sum_{i=1}^n C_i^2,\sum_{i=1}^n C_iY_i\right]-4Cov\left[\sum_{i=1}^n Y_i^2,\sum_{i=1}^n C_iY_i\right]\Bigg). \end{aligned}$$

The individual terms can be computed as follows:

$$\begin{aligned} Var\left[\sum_{i=1}^n C_i^2\right] &= \sum_{i=1}^n Var[C_i^2] \\ &= \sum_{i=1}^n \left(E[C_i^4]-(E[C_i^2])^2\right) \\ &= \sum_{i=1}^n \left(E[C_i^4]-(Var[C_i]+(E[C_i])^2)^2\right) \\ &= \sum_{i=1}^n \left(E[C_i^4]-(\sigma^2+\sigma_\mu^2+\mu^2)^2\right) \\ &= nEC_1^4-n(\sigma^2+\sigma_\mu^2+\mu^2)^2. \end{aligned}$$

Assuming normality, we have

$$\begin{aligned} Var\left[\sum_{i=1}^n C_i^2\right] &= n\Big(3\epsilon+3\sigma^4+6\sigma_\mu^2\sigma^2+6\mu^2\sigma^2+3\sigma_\mu^4+6\mu^2\sigma_\mu^2+\mu^4-(\sigma^2+\sigma_\mu^2+\mu^2)^2\Big) \\ &= n(3\epsilon+2\sigma^4+2\sigma_\mu^4+4\sigma^2\sigma_\mu^2+4\mu^2\sigma^2+4\mu^2\sigma_\mu^2). \end{aligned}$$

Assuming additionally that $\mu = 0$ and $\epsilon = 0$, we have

$$Var\left[\sum_{i=1}^{n} C_i^2\right] = 2n(\sigma^2 + \sigma_\mu^2)^2.$$

Since C_i and Y_i are symmetrically defined, we have

$$Var\left[\sum_{i=1}^{n} Y_i^2\right] = Var\left[\sum_{i=1}^{n} C_i^2\right].$$

Next,

$$\begin{aligned} Var\left[\sum_{i=1}^{n} C_i Y_i\right] &= \sum_{i=1}^{n} Var[C_i Y_i] \\ &= \sum_{i=1}^{n} \left(E[C_i^2 Y_i^2] - (E[C_i Y_i])^2 \right) \end{aligned}$$

where

$$\begin{aligned} E[C_i Y_i]^2 &= E\Big[E[C_i^2 Y_i^2 | Z_i]\Big] \\ &= E[E[C_i^2|Z_i)E(Y_i^2|Z_i]] \\ &= E[\Sigma_i^2 + M_i^2]^2 \\ &= E[\Sigma_i^4 + M_i^4 + 2\Sigma_i^2 M_i^2] \\ &= Var[\Sigma_i^2] + (E[\Sigma_i^2])^2 + E[M_i^4] + 2E[\Sigma_i^2]E[M_i^2] \\ &= \epsilon + \sigma^4 + E[M_i^4] + 2\sigma^2(\sigma_\mu^2 + \mu^2); \end{aligned}$$

and

$$\begin{aligned} E[C_i Y_i] &= Cov[C_i, Y_i] + E[C_i]E[Y_i] \\ &= \sigma_\mu^2 + \mu^2. \end{aligned}$$

Therefore,

$$Var\left[\sum_{i=1}^{n} C_i Y_i\right] = \sum_{i=1}^{n} \left(\epsilon + \sigma^4 + EM_i^4 + 2\sigma^2(\sigma_\mu^2 + \mu^2) - (\sigma_\mu^2 + \mu^2)^2 \right).$$

Assuming normality, we have

$$\begin{aligned} E[C_i Y_i]^2 &= \epsilon + \sigma^4 + 3\sigma_\mu^4 + 6\mu^2\sigma_\mu^2 + \mu^4 + 2\sigma^2\sigma_\mu^2 + 2\sigma^2\mu^2; \\ E[C_i Y_i] &= \sigma_\mu^2 + \mu^2; \\ Var\left[\sum_{i=1}^{n} C_i Y_i\right] &= n(\epsilon + \sigma^4 + 2\sigma_\mu^4 + 2\sigma^2\sigma_\mu^2 + 2\mu^2\sigma^2 + 4\mu^2\sigma_\mu^2). \end{aligned}$$

Assuming additionally that $\mu = 0$ and $\epsilon = 0$, we have

$$\begin{aligned} E[C_i Y_i]^2 &= (\sigma^2 + \sigma_\mu^2)^2 + 2\sigma_\mu^4; \\ E[C_i Y_i] &= \sigma_\mu^2; \\ Var\left[\sum_{i=1}^{n} C_i Y_i\right] &= n\Big[(\sigma^2 + \sigma_\mu^2)^2 + \sigma_\mu^4\Big]. \end{aligned}$$

The covariance terms are computed as follows:

$$Cov\left[\sum_{i=1}^{n} C_i^2, \sum_{i=1}^{n} Y_i^2\right] = \sum_{i=1}^{n} Cov[C_i^2, Y_i^2] = \sum_{i=1}^{n} (E[C_i^2 Y_i^2] - E[C_i^2]E[Y_i^2]).$$

Assuming normality, we have

$$\begin{aligned} Cov\left[\sum_{i=1}^{n} C_i^2, \sum_{i=1}^{n} Y_i^2\right] &= n\Big(\epsilon + \sigma^4 + 3\sigma_\mu^4 + 6\mu^2\sigma_\mu^2 + \mu^4 + 2\sigma^2\sigma_\mu^2 + 2\sigma^2\mu^2 - (\sigma^2 + \sigma_\mu^2 + \mu^2)^2\Big) \\ &= n(\epsilon + 2\sigma_\mu^4 + 4\mu^2\sigma_\mu^2). \end{aligned}$$

Assuming additionally that $\mu = 0$ and $\epsilon = 0$, we have

$$Cov\left[\sum_{i=1}^{n} C_i^2, \sum_{i=1}^{n} Y_i^2\right] = 2n\sigma_\mu^4.$$

Finally, since C_i and Y_i are symmetrically defined, we have

$$\begin{aligned} Cov\left[\sum_{i=1}^{n} C_i^2, \sum_{i=1}^{n} C_i Y_i\right] &= Cov\left[\sum_{i=1}^{n} Y_i^2, \sum_{i=1}^{n} C_i Y_i\right] \\ &= \sum_{i=1}^{n} Cov[C_i^2, C_i Y_i] \\ &= \sum_{i=1}^{n} \Big(E[C_i^3 Y_i] - E[C_i^2]E[C_i Y_i]\Big), \end{aligned}$$

where

$$E[C_i^3 Y_i] = E\Big[E[C_i^3 Y_i | Z_i]\Big] = E\Big[E[C_i^3 | Z_i]E[Y_i | Z_i]\Big].$$

Assuming normality, we have

$$\begin{aligned}
E[C_i^3 Y_i] &= E\Big[(3M_i\Sigma_i^2 + M_i^3)M_i\Big] \\
&= E[3M_i^2\Sigma_i^2 + M_i^4] \\
&= 3E[M_i^2]E[\Sigma_i^2] + E[M_i^4] \\
&= 3(\sigma_\mu^2 + \mu^2)\sigma^2 + 3\sigma_\mu^4 + 6\mu^2\sigma_\mu^2 + \mu^4 \\
&= \mu^4 + 3\sigma_\mu^4 + 3\sigma^2\sigma_\mu^2 + 3\mu^2\sigma^2 + 6\mu^2\sigma_\mu^2; \\
E[C_i^2] &= \sigma^2 + \sigma_\mu^2 + \mu^2; \\
E[C_i Y_i] &= \sigma_\mu^2 + \mu^2;
\end{aligned}$$

and therefore,

$$\begin{aligned}
Cov\left[\sum_{i=1}^n C_i^2, \sum_{i=1}^n C_i Y_i\right] &= n\Big(\mu^4 + 3\sigma_\mu^4 + 3\sigma^2\sigma_\mu^2 + 3\mu^2\sigma^2 + 6\mu^2\sigma_\mu^2 - (\sigma^2 + \sigma_\mu^2 + \mu^2)(\sigma_\mu^2 + \mu^2)\Big) \\
&= n\Big(\mu^4 + 3\sigma_\mu^4 + 3\sigma^2\sigma_\mu^2 + 3\mu^2\sigma^2 + 6\mu^2\sigma_\mu^2 - (\mu^4 + \sigma_\mu^4 + \sigma^2\sigma_\mu^2 + \mu^2\sigma^2 + 2\mu^2\sigma_\mu^2)\Big) \\
&= 2n(\sigma_\mu^4 + \sigma^2\sigma_\mu^2 + \mu^2\sigma^2 + 2\mu^2\sigma_\mu^2).
\end{aligned}$$

Assuming additionally that $\mu = 0$ and $\epsilon = 0$, we have

$$\begin{aligned}
E[C_i^3 Y_i] &= 3\sigma^2\sigma_\mu^2 + 3\sigma_\mu^4; \\
E[C_i^2] &= \sigma^2 + \sigma_\mu^2; \\
E[C_i Y_i] &= \sigma_\mu^2; \\
Cov\left[\sum_{i=1}^n C_i^2, \sum_{i=1}^n C_i Y_i\right] &= 2n\sigma_\mu^2(\sigma^2 + \sigma_\mu^2).
\end{aligned}$$

Putting the terms together, we derive the variance as follows, assuming that M_i follows a normal distribution,

$$\begin{aligned}
Var[\tilde{S}_{int}] = \frac{1}{4a^2}\Big\{ & 2n(3\epsilon + 2\sigma^4 + 2\sigma_\mu^4 + 4\sigma^2\sigma_\mu^2 + 4\mu^2\sigma^2 + 4\mu^2\sigma_\mu^2) \\
& + 4n(\epsilon + \sigma^4 + 2\sigma_\mu^4 + 2\sigma^2\sigma_\mu^2 + 2\mu^2\sigma^2 + 4\mu^2\sigma_\mu^2) + 2n(\epsilon + 2\sigma_\mu^4 + 4\mu^2\sigma_\mu^2) \\
& - 16n(\sigma_\mu^4 + \sigma^2\sigma_\mu^2 + \mu^2\sigma^2 + 2\mu^2\sigma_\mu^2)\Big\} \\
= \frac{n}{a^2}(3\epsilon + 2\sigma^4).
\end{aligned}$$

Assuming additionally that $\mu = 0$ and $\epsilon = 0$, we have

$$Var[\tilde{S}_{int}] = \frac{2n}{a^2}\sigma^4.$$

Appendix D. Calculating $Var[S_{ext}]$

$$\begin{aligned} Var[S_{ext}] &= Var\left[\frac{1}{a}(\sum_{i=1}^{n} C_iY_i - n\bar{C}\bar{Y})\right] \\ &= \frac{1}{a^2}Var\left[\sum_{i=1}^{n} C_iY_i - n\bar{C}\bar{Y}\right] \\ &= \frac{1}{a^2}\bigg(Var\left[\sum_{i=1}^{n} C_iY_i\right] + Var[n\bar{C}\bar{Y}] - 2Cov\left[\sum_{i=1}^{n} C_iY_i, n\bar{C}\bar{Y}\right]\bigg). \end{aligned}$$

Here,

$$Var\left[\sum_{i=1}^{n} C_iY_i\right] = \sum_{i=1}^{n}\bigg(\epsilon + \sigma^4 + E[M_i^4] + 2\sigma^2(\sigma_\mu^2 + \mu^2) - (\sigma_\mu^2 + \mu^2)^2\bigg).$$

Also,

$$\begin{aligned} Var[n\bar{C}\bar{Y}] &= n^2Var\left[\frac{C_1 + \cdots + C_n}{n} \cdot \frac{Y_1 + \cdots + Y_n}{n}\right] \\ &= \frac{n^2}{n^4}Var\left[\sum_k C_kY_k + \sum_{i\neq j} C_iY_j\right] \\ &= \frac{1}{n^2}\bigg(Var\left[\sum_k C_kY_k\right] + Var\left[\sum_{i\neq j} C_iY_j\right] + 2Cov\left[\sum_k C_kY_k, \sum_{i\neq j} C_iY_j\right]\bigg). \end{aligned}$$

Assuming normality on M_i and assuming that $\mu = 0$ and $\epsilon = 0$ (constant variance across cells), we have

$$\begin{aligned} Var\left[\sum_k C_kY_k\right] &= n(\sigma^4 + 3\sigma_\mu^4 + 2\sigma^2\sigma_\mu^2 - \sigma_\mu^4) \\ &= n(\sigma^2 + \sigma_\mu^2)^2 + n\sigma_\mu^4. \end{aligned}$$

Also,

$$Var\left[\sum_{i\neq j} C_iY_j\right] = \sum_{i\neq j} Var[C_iY_j] + 2\sum_{i=k \text{ or } j=l} Cov[C_iY_j, C_kY_l] + 2\sum_{i\neq k \text{ and } j\neq l} Cov[C_iY_j, C_kY_l].$$

Under the assumptions made above, we have

$$\begin{aligned} Var[C_iY_j] &= E[C_i^2Y_j^2] - (E[C_iY_j])^2 \\ &= E[C_i^2]E[Y_j^2] - (E[C_i]E[Y_j])^2 \\ &= (\sigma^2 + \sigma_\mu^2)^2. \end{aligned}$$

If $i = k$,

$$\begin{aligned} Cov[C_iY_j, C_kY_l] &= E[C_iY_jC_kY_l] - E[C_iY_j]E[C_kY_l] \\ &= E[C_i^2]E[Y_j]E[Y_l] - (E[C_i])^2E[Y_j]E[Y_l] \\ &= 0. \end{aligned}$$

Similarly, we can derive that the covariance is 0 for other cases where $j = l$ or where $i \neq k$ and $j \neq l$. Hence,

$$Var\left[\sum_{i\neq j} C_iY_j\right] = n(n-1)(\sigma^2 + \sigma_\mu^2)^2.$$

Additionally, under the normality assumption and with $\mu = 0$ and $\epsilon = 0$,

$$Cov\left[\sum_k C_kY_k, \sum_{i\neq j} C_iY_j\right] = 0.$$

Therefore,

$$\begin{aligned} Var[n\bar{C}\bar{Y}] &= \frac{1}{n^2}\left(n(\sigma^2 + \sigma_\mu^2)^2 + n\sigma_\mu^4 + n(n-1)(\sigma^2 + \sigma_\mu^2)^2\right) \\ &= \frac{1}{n^2}\left(n^2(\sigma^2 + \sigma_\mu^2)^2 + n\sigma_\mu^4\right) \\ &= (\sigma^2 + \sigma_\mu^2)^2 + \frac{\sigma_\mu^4}{n}. \end{aligned}$$

Furthermore,

$$\begin{aligned} Cov\left[\sum_{i=1}^n C_iY_i, n\bar{C}\bar{Y}\right] &= \frac{1}{n}Cov\left[\sum_{i=1}^n C_iY_i, \sum_k C_kY_k + \sum_{i\neq j} C_iY_j\right] \\ &= \frac{1}{n}\left(Cov\left[\sum_{i=1}^n C_iY_i, \sum_k C_kY_k\right] + Cov\left[\sum_{i=1}^n C_iY_i, \sum_{i\neq j} C_iY_j\right]\right) \\ &= \frac{1}{n}\left(Var\left[\sum_{i=1}^n C_iY_i\right]\right) \\ &= (\sigma^2 + \sigma_\mu^2)^2 + \sigma_\mu^4. \end{aligned}$$

$$\begin{aligned} Var[S_{ext}] &= \frac{1}{a^2}\left(n(\sigma^2 + \sigma_\mu^2)^2 + n\sigma_\mu^4 + (\sigma^2 + \sigma_\mu^2)^2 + \frac{\sigma_\mu^4}{n} - 2(\sigma^2 + \sigma_\mu^2)^2 - 2\sigma_\mu^4\right) \\ &= \frac{n-1}{a^2}(\sigma^2 + \sigma_\mu^2)^2 + \frac{(n-1)^2}{na^2}\sigma_\mu^4. \end{aligned}$$

Department of Genetics, Stanford University; Department of Human Genetics, University of Chicago; Current Address: Department of Statistical Science, University of Idaho

Departments of Mathematics, Molecular & Cell Biology and Computer Science, UC Berkeley