Published in final edited form as:
Stat Appl Genet Mol Biol. 2016 December 01; 15(6): 447–471. doi:10.1515/sagmb-2016-0002.

Estimating intrinsic and extrinsic noise from single-cell gene expression measurements

Audrey Qiuyan Fu[a,*] 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 identically regulated gene pairs in single cells. We examine established formulas [Elowitz, M. B., A. J. Levine, E. D. Siggia and P. S. Swain (2002): "Stochastic gene expression in a single cell," Science, 297, 1183–1186.] for the estimation of intrinsic and extrinsic noise and provide interpretations of them in terms of a hierarchical model. This allows us to derive alternative estimators that minimize bias or mean squared error. We provide a geometric interpretation of these results that clarifies the interpretation in [Elowitz, M. B., A. J. Levine, E. D. Siggia and P. S. Swain (2002): "Stochastic gene expression in a single cell," Science, 297, 1183–1186.]. We also demonstrate through simulation and re-analysis of published data that the distribution assumptions underlying the hierarchical model have to be satisfied for the estimators to produce sensible results, which highlights the importance of normalization.

Keywords

gene expression; noise; optimal estimators; single cell

1 Introduction

A gene can have different expression levels in living cells that have the same genetic material and are subject to the same environment (Stegle et al., 2015). During early development of an organism, distinct expression profiles eventually lead to formation of different tissues. Moreover, complex tissues such as brain have many different subtypes of cells with different gene expression profiles. However, variation in expression between cells is reflective not only of distinct biological state, but also of stochasticity underlying many of the processes fundamental to the molecular biology of cell.

In a classic paper on the stochasticity of gene expression in single cells, Elowitz et al. (2002) introduced a clever two-reporter expression assay designed to tease apart "intrinsic" and

[*]**Corresponding authors: Audrey Qiuyan Fu,** Department of Genetics, Stanford University, Stanford, CA 94305, USA; and Department of Human Genetics, University of Chicago, Chicago, IL 60637, USA, audreyf@uidaho.edu; and **Lior Pachter,** Departments of Mathematics, Molecular & Cell Biology and Computer Science, University of California, Berkeley, Berkeley, CA 94720, USA, lpachter@math.berkeley.edu.
[a]**Present address:** Department of Statistical Science, University of Idaho, Moscow, ID 83844, USA

"extrinsic" variation (also called "noise") from the overall variability in gene expression: the intrinsic noise is the variation in the expression of the same gene in identical environment, whereas the extrinsic noise is the variation in gene expression due to cellular environment that impacts all the genes at once. 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 $c_1, \dots c_n$ and $y_1, \dots y_n$, 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 Elowitz et al. (2002), these noise terms are defined as squared coefficients of variation and specific formulas are provided for estimating $\eta^2_{\text{ext}}, \eta^2_{\text{int}}$ and η^2_{tot} (hereafter referred to as the ELSS estimates):

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

$$\eta^2_{\text{ext}} = \frac{\frac{1}{n}\sum_{i=1}^n c_i \cdot y_i - \overline{c} \cdot \bar{y}}{\overline{c} \cdot \bar{y}}, \quad (2)$$

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

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

Hilfinger and Paulsson (2011) later interpreted these estimates in terms of the "law of total variance" (explained in the next section), which sheds light on the statistical basis of the ELSS estimators but does not address questions about their statistical properties. In this paper, we derive the bias and mean squared error of the ELSS estimators and examine their optimality. We also examine the geometric and biological interpretation of the estimators.

The processes that lead to the expression of the reporters (or genes in general) are much more complex than described here, e.g. the models described in the paper ignore the effects of translation. Many studies (e.g. Rausenberger and Kollmann 2008 and Komorowski et al. 2013) have developed detailed mathematical models for these processes. While some of our results may generalize and be relevant in more general settings, we restrict our analysis to the intrinsic and extrinsic noise as examined by Elowitz et al. (2002) and accessible via static reporter expression experiments. Analyses are implemented in the R package noise available on CRAN.

2 A hierarchical model

We begin by introducing a hierarchical model that provides a formal model for the experiments of Elowitz et al. (2002) and that provides insight into the numerators of (1,2,3). They are the key components of the Elowitz et al. (2002) 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 Hilfinger and Paulsson (2011)]. Consistent with Elowitz et al. (2002), we posit that the cellular conditional random variables associated to the two reporters have the same distribution F with mean M_i and variance $\sum_i^2$, both parameters being unique to the i th cell:

$$C_i|Z_i{\sim}F(M_i,\sum\nolimits_i^2) \qquad (4)$$

and

$$Y_i|Z_i{\sim}F(M_i,\sum\nolimits_i^2). \qquad (5)$$

Thinking of a two reporter experiment as “random,” in the sense that the states of cells Z_1, … Z_n are random, across cells we have

$$M_i{\sim}G(\mu,\sigma_\mu^2)$$

and

$$\sum\nolimits_i^2{\sim}H(\sigma^2,\varepsilon),$$

where G is the distribution of all the M_is, with mean μ and variance σ_μ^2, and H that of all the $\sum_i^2$s, 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 $\sum_i^2$ 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:

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

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

$$\mathrm{Var}[C_i]=E_{Z_i}[\mathrm{Var}[C_i|Z_i]]+\mathrm{Var}_{Z_i}[E[C_i|Z_i]]. \quad (7)$$

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*,

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

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

With this notation equation (7) becomes

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

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

$$C_i \sim F'(\mu,\sigma^2+\sigma_\mu^2);$$

$$Y_i \sim F'(\mu,\sigma^2+\sigma_\mu^2),$$

where the marginal distribution F' may or may not be the same as the conditional distribution *F*.

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

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

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

3 Extrinsic noise

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

$$
\begin{aligned}
&\mathrm{Var}[E[C_i|Z_i]]=E[E[C_i|Z_i]^2]-(E[E[C_i]Z_i])^2\\
&=E[E[C_i|Z_i]E[Y_i|Z_i]]-(E[E[C_i]Z_i])^2\\
&=E[E[C_iY_i|Z_i]]-E[E[C_i|Z_i]E[E[Y_i|Z_i]]\\
&=E[C_iY_i]-E[C_i]E[Y_i]\\
&=\mathrm{Cov}[C_i,Y_i]. \qquad (11)
\end{aligned}
$$

This connection between the extrinsic noise, the law of total variance and the covariance of C_i and Y_i was noted in Hilfinger and Paulsson (2011).

Formula (11) leads to the following unbiased estimator for the extrinsic noise, as it is an unbiased estimator estimator for the covariance:

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

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:

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

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

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

which is minimized when

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

$$
a=(n-1)\left(1+\frac{1}{n}\right)+\left(\frac{\sigma^2+\sigma_\mu^2}{\sigma_\mu^2}\right)^2=(n-1)\left(1+\frac{1}{n}\right)+\frac{1}{\rho^2}. \qquad (12)
$$

The last step in (12) is due to Equations (9), (10) and (11):

$$
\frac{\sigma_\mu^2}{\sigma^2+\sigma_\mu^2}=\frac{\text{Cov}[C_i,Y_i]}{\text{Var}[C_i]}=\frac{\text{Cov}[C_i,Y_i]}{\sqrt{\text{Var}[C_i]}\sqrt{\text{Var}[Y_i]}}=\rho. \qquad (13)
$$

It is interesting to note that (12) comprises two parts: the first, $(n-1)(1+\frac{1}{n})$ converges to $n-1$ as $n \to \infty$, while the second, $(\frac{\sigma^2+\sigma_\mu^2}{\sigma_\mu^2})^2$ is equal to $\frac{1}{\rho^2}$ where ρ is the correlation between the two reporter expression vectors **C** and **Y**. See Appendices A and B for more details.

4 Intrinsic noise

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

$$
\begin{aligned}
E[\text{Var}[C_i|Z_i]]&=\text{Var}[C_i]-\text{Var}[E[C_i|Z_i]]\\
&=\frac{1}{2}[\text{Var}[C_i]+\text{Var}[Y_i]]-\text{Cov}[C_i,Y_i]\\
&=\frac{1}{2}[\text{Var}[C_i]-2\text{Cov}[C_i,Y_i]+\text{Var}[Y_i]]\\
&=\frac{1}{2}\text{Var}[C_i-Y_i]\\
&=\frac{1}{2}\left(E[C_i-Y_i]^2-(E[C_i-Y_i])^2\right).
\end{aligned} \qquad (14)
$$

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

$$\begin{aligned} S^*_{\text{int}} &= \frac{1}{2(n-1)}\sum_{i=1}^{n}\left[(C_i - Y_i) - \left(\overline{C} - \bar{Y}\right)\right]^2 \\ &= \frac{1}{2(n-1)}\sum_{i=1}^{n}(C_i - Y_i)^2 - \frac{n}{2(n-1)}(\overline{C} - \bar{Y})^2. \end{aligned}$$

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

$$S_{\text{int}} = \frac{1}{2a}\left(\sum_{1}^{n}(C_i - Y_i)^2 - n(\overline{C} - \bar{Y})^2\right). \qquad (15)$$

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

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

The value of *a* that minimizes this expression is

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

See Appendices A and C for the complete derivation.

The analysis above can be simplified with an additional assumption, namely that $\bar{C} = \overline{Y}$. 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 $\bar{C} = \overline{Y}$, the estimator (15) simplifies to

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

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

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

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

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

Assuming again that $\mu = 0$ and $\boldsymbol{\varepsilon} = 0$, the MSE simplifies to

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

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

5 Geometric interpretation

Figure 3A of Elowitz et al. (2002) 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 Elowitz et al. (2002), we have redrawn it in a format that highlights the math (Figure 1). The projection of a point (c_i, y_i) onto the line $y = c$ is the point $((y_i+c_i)/\sqrt{2},(y_i-c_i)/\sqrt{2})$, shown as the red point in Figure 1. Assuming equal means ($\bar{c}=\bar{y}$), the intrinsic noise, as estimated by the unbiased estimator (1), is then the mean squared distance of the points from 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 Figure 1 is based on an alternative formulation of sample covariance (Hayes, 2011):

$$\mathrm{Cov}(\mathbf{c},\mathbf{y})=\frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j>i}^{n}\frac{1}{2}(c_i-c_j)(y_i-y_j). \qquad (16)$$

This formulation of the sample covariance has the interpretation of being an average of the signed area of triangles associated to pairs of points. Figure 1 illustrates these signed triangles using a randomly selected point (the blue point). This formulation is very different from what might be considered at first glance an appropriate analogy to intrinsic noise, namely the sample variance along the line $y = c$.

An alternative estimate for the extrinsic noise based on the sample variance of the projected points along the line $y = c$ (using the projected centroid as the mean, which is shown as the green point in Figure 1) turns out to be biased by an amount equal to the total noise. This sample variance averages the squared distances of the data points from the centroid (green point) after projection onto the line $y = c$; see the distance between the red and green points in Figure 1. Since

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

the bias is

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

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 Elowitz et al. (2002) is completely accurate in stating that "Spread of points perpendicular to the diagonal line on which CFP and YFP intensities are equal corresponds to intrinsic noise, whereas spread parallel to this line is increased by extrinsic noise." However the geometric interpretation of covariance makes it precise *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 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 Elowitz et al. (2002), 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 estimators converge to the ELSS estimate (Table 1). The mean and variance of the estimators are summarized in Table 6 in Appendix E. For intrinsic noise, assuming $\bar{c} = \bar{y}$, the ELSS estimator is optimal for bias and MSE at large n and optimal for bias at small n. Indeed, in Elowitz et al. (2002), 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 the min-MSE estimators when the sample size is small (e.g. much less than 50).

It is worth pointing out that the correction factor $1/a$ in the min-MSE estimators tends to be smaller than that in the unbiased estimators $(1/(n-1))$ and the asymptotic estimators ($1/n$; Table 1). This smaller correction $1/a$ makes the min-MSE estimators "shrinkage" estimators, such that they achieve better MSE despite being biased, just like the Jame-Stein estimator (James and Stein, 1961). Our simulation results confirm this point (Table 2). However, using the sample correlation, instead of the true one, in our min-MSE estimators leads to increased MSE, although the estimates with the sample correlation do not differ much on average from that with the true correlation.

6.2 Data normalization

Our hierarchical model, as well as the ANOVA interpretation, is consistent with the model in Elowitz et al. (2002); 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. We also performed simulations under the hierarchical model, with and without identical distribution for the two reporters, and summarized the results in Table 3. Estimates of intrinsic and extrinsic noise are the same as the truth when the identical distribution assumption applies. When this assumption is not satisfied, the theory breaks down and it is unclear what the estimates mean.

Other studies have adapted this system and used other reporter combinations that may have markedly different distributions. For example, Yang et al. (2014) 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 Figure 3A from their paper. In contrast, Schmiedel et al. (2015) 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 equations 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.

6.3 Assessing the ratio of extrinsic to intrinsic noise from sample correlation

We have seen from (13) 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 $\rho(\mathbf{c}, \mathbf{y})/[1 - \rho(\mathbf{c},\mathbf{y})]$.

7 Re-analysis of published two-reporter experiment data

Michael Elowitz and Peter Swain have kindly shared with us their data published in Elowitz et al. (2002). Here we focus on the data in Figure 3A of their paper, which contain the unnormalized fluorescence intensities of CFP and YFP in the *E. coli.* strain D22 and in strain M22. We normalized the data as follows such that the resulting scatterplots are close to Figure 3A:

$$\text{D22:}\ c_i=(c_i^*-\overline{c^*})/(8s_c^*)+1; y_i=(y_i^*-\overline{y^*})/(8s_y^*)+1;$$

$$\text{M22:}\ c_i=(c_i^*-\overline{c^*})/(12s_c^*)+1; y_i=(y_i^*-\overline{y^*})/(12s_y^*)+1,$$

where c_i^* and y_i^* are the unnormalized intensity of the CFP and YFP, respectively, in the *i*th cell, $\overline{c^*}$ the sample mean, and s_y^* the sample standard deviation. The normalized intensities are close to normal distributions, and all four distributions have mean 1. At a sample size of over 200, the different estimators in Table 4 give essentially the same result. Additionally, the ratio of the estimated extrinsic and total noise is close to the sample correlation, verifying our theoretical result.

Nam Ki Lee and Sora Yang have also kindly shared with us their data published in Yang et al. (2014). Here we analyze the data in Figure 3A of their paper, which are the expression levels (intensities) of two reporters, CFP and mCherry (also see Sec. 6.2). The shared, unnormalized intensities have very different sample means (Table 4). Application of the estimators in Table 1 to these data gives two different estimates of the intrinsic noise, with the ELSS estimate being nearly three times the estimates under the equal mean assumption. To normalize the data, we removed the few negative values, $\log_2$ transformed the data, and quantile normalized between the two reporters (see summary statistics in Table 4). Applying our estimators to the normalized data, all estimates are consistent with one another. This analysis illustrates the importance of the equal mean assumption: when this assumption is not satisfied, the ELSS estimator leads to overestimation of the intrinsic noise.

Additionally, we subsampled from these data sets and assessed the performance of the estimators as the sample size decreased. At each sample size, we repeated the subsampling 1000 times and computed the mean and standard deviation of the noise estimates (Table 5). Whereas the means of the estimates do not differ from those obtained using the entire data sets, the variation (measured by the standard deviation) increases quickly with decreasing sample sizes. For the Elowitz et al. data, the standard deviation in the estimates roughly doubles for both types of noise as the sample size halves. Comparing the standard deviation to the mean suggests that 200 is indeed a reasonable sample size for estimates with small variation (compare with their actual sample sizes of 284 and 250 for the two strains). For the Yang et al. data, the increase in the standard deviation is much less drastic, and 200 also appears a decent sample size for reasonably small variation in the estimates.

8 Conclusions and discussion

Our hierarchical model for Elowitz et al. (2002) provides statistically interpretable parameters representing intrinsic and extrinsic noise, and allows for the derivation of estimators with optimality guarantees. Furthermore, the model highlights experimental assumptions that need to be satisfied for the estimators to be valid, specifically that the two reporters need to have the same distribution (within a cell) and hence normalization may be necessary. Whereas similar hierarchical models have been proposed before to study heterogeneity among single cells (see, e.g. Finkenstädt et al., 2013, and Koeppl et al., 2012), our hierarchical model explicitly parameterize the two types of noise, and reveals their equivalence to other quantities, as indicated by (11) and (14), which enable derivation of closed-form estimators of these parameters (summarized in Table 1). We use bias and MSE to explicitly evaluate the performance of different estimators, and recognize the asymptotic equivalence of multiple estimators.

Other experiments have been set up to explore and assess intrinsic and extrinsic noise, and some of our results may be useful in those settings. For example, Volfson et al. (2006) used a single reporter but two *Saccharomyces cerevisiae* strains, with one strain containing only one copy of the reporter, and the other strain two copies. Assuming no strain effect, which may be thought of as batch effect, the authors applied the following estimators for (unscaled) intrinsic and extrinsic noise (consistent with their notation, and without the denominator of $\bar{C}\bar{Y}$ as used in the ELSS estimators in Table 1):

$$V_i = 2V_1 - V_2/2; \quad (17)$$

$$V_e = V_2/2 - V_1, \quad (18)$$

where V_1 and V_2 are the variance in the 1-copy and 2-copy strains, respectively, and V_i and V_e are intrinsic and extrinsic noise, respectively. These estimators are in fact consistent with (11) and (14) under our hierarchical model:

$$V_1 = \mathrm{Var}[C_1] = V_i + V_e = \mathrm{Var}[C_2]; \quad (19)$$

$$V_2 = \mathrm{Var}[C_1 + C_2] = \mathrm{Var}[C_1] + \mathrm{Var}[C_2] + 2\mathrm{Cov}[C_1, C_2] = 2(V_i + V_e) + 2V_e. \quad (20)$$

Together, (19) and (20) give rise to (17) and (18). Note that (19) and (20) imply that the extrinsic noise is also the covariance here, except that the covariance is between the 1-copy and 2-copy strains with the same reporter; this is also pointed out by Sherman et al. (2015). Additionally, the total (marginal) noise of the reporter is the sum of intrinsic and extrinsic noise (19). However, consistent with our analysis of the assumptions of the hierarchical model, these estimators hold only when the variance for each single copy in the 2-copy strain is identical to that in the 1-copy strain. This is equivalent to assuming no strain (batch) effect, which can be a rather strong assumption.

We note that during the preparation of this manuscript, Erik van Nimwegen independently examined the Elowitz et al. (2002) paper form a Bayesian point of view (van Nimwegen, 2016).

Acknowledgments

This project began as a result of discussion during a journal club meeting of Prof. Jonathan Pritchard's group that A.F. was attending. We thank Michael Elowitz, Peter Swain, Nam Ki Lee and Sora Yang for sharing their data from Elowitz et al. (2002) and from Yang et al. (2014), respectively. We also thank helpful comments we have received since posting the manuscript online. In particular, we thank Arjun Raj for bringing up the 1- vs 2-copy experiment, and Erik van Nimwegen for helpful discussions. We also thank Editor in Chief Prof. Michael Stumpf and two anonymous reviewers for insightful comments that led to a significantly enriched version. A.F. was partially supported by K99 HG007368 and R00 HG007368 (NIH/NHGRI). L.P. was partially supported by NIH grants R01 HG006129 and R01 DK094699.

Appendix

A Moments of *M*

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 $\sum_i^2 \sim H(\sigma^2,\varepsilon)$, such that

$$E[\sum_i^2]=\sigma^2;$$

$$\mathrm{Var}[\sum_i^2]=\varepsilon,$$

we have

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

Further assuming that μ = 0, i.e. the means are all 0, and that ε = 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

$$E[C_i^3]{=}0$$

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

B Calculating *Var*[*S*_{*ext*}]

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

B.1 Calculating $\mathrm{Var}\left[\sum_{i=1}^{n}C_iY_i\right]$

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

where

$$
\begin{aligned}
E[C_iY_i]^2 &= E\left[E[C_i^2Y_i^2|Z_i]\right]\\
&= E\left[E[C_i^2|Z_i)E(Y_i^2|Z_i]\right]\\
&= E[\sum\nolimits_i^2+M_i^2]^2\\
&= E[\sum\nolimits_i^4+M_i^4+2\sum\nolimits_i^2M_i^2]\\
&= \mathrm{Var}[\sum\nolimits_i^2]+(E[\sum\nolimits_i^2])^2+E[M_i^4]+2E[\sum\nolimits_i^2]E[M_i^2]\\
&= \varepsilon+\sigma^4+E[M_i^4]+2\sigma^2(\sigma_\mu^2+\mu^2);
\end{aligned}
$$

and

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

Therefore,

$$
\mathrm{Var}\left[\sum_{i=1}^{n} C_iY_i\right]=\sum_{i=1}^{n}\left(\varepsilon+\sigma^4+\mathrm{EM}_i^4+2\sigma^2(\sigma_\mu^2+\mu^2)-(\sigma_\mu^2+\mu^2)^2\right).
$$

B.2 Calculating *Var*[*n* $\bar{C}$ $\bar{Y}$]

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

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

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

Also,

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

Under the assumptions made above, we have

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

If $i = k$,

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

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,

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

Additionally, under the normality assumption and with μ = 0 and **ε** = 0,

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

Therefore,

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

B.3 Calculating $\mathrm{Cov}\ [\sum_{i=1}^{n} C_i Y_i, n\overline{C}\bar{Y}]$

$$
\begin{aligned}
\mathrm{Cov}\ \left[\sum_{i=1}^{n} C_i Y_i, n\overline{C}\bar{Y}\right] &= \frac{1}{n}\mathrm{Cov}\ \left[\sum_{i=1}^{n} C_i Y_i, \sum_{k} C_k Y_k + \sum_{i \neq j} C_i Y_j\right] \\
&= \frac{1}{n}\left(\mathrm{Cov}\ \left[\sum_{i=1}^{n} C_i Y_i, \sum_{k} C_k Y_k\right] + \mathrm{Cov}\ \left[\sum_{i=1}^{n} C_i Y_i, \sum_{i \neq j} C_i Y_j\right]\right) \\
&= \frac{1}{n}\left(\mathrm{Var}\ \left[\sum_{i=1}^{n} C_i Y_i\right]\right) \\
&= (\sigma^2 + \sigma_\mu^2)^2 + \sigma_\mu^4.
\end{aligned}
$$

Putting the terms above together, we have

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

C MSE of the general intrinsic noise estimator

The general form of the estimator for intrinsic noise is

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

C.1 Calculating *Var*[*S*]

Thus

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

Below we will assume normality, as well as μ = 0 and ε = 0, to facilitate the derivation. Note that $Var[\Sigma(C_i - Y_i)^2]$ is derived in Appendix D.

C.1.1 Calculating $Var[(\bar{C} - \bar{Y})^2]$—First, we note that

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

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

This is because

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

Additionally, from Appendix B, we have

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

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

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

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

$$\mathrm{Cov}[C_i^2, C_i Y_j] = E[C_i^3 Y_j] - E[C_i^2]E[C_i Y_j] = 0$$

and

$$\mathrm{Cov}[C_i^2, C_j Y_k] = E[C_i^2 C_j Y_k] - E[C_i^2]E[C_j Y_k] = 0,$$

we have

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

For $Cov\, [\Sigma_{i\neq j}\, C_i C_j, \Sigma C_l Y_l]$, since

$$\mathrm{Cov}[C_i C_j, C_i Y_i] = E[C_i^2 Y_i C_j] - E[C_i C_j]E[C_i Y_i] = 0$$

and

$$\mathrm{Cov}[C_k C_l, C_i Y_i] = E[C_k C_l C_i Y_i] - E[C_k C_l]E[C_i Y_i] = 0,$$

we have

$$\mathrm{Cov}\ \left[\sum_{i\neq j} C_i C_j, \sum C_l Y_l\right] = 0.$$

Additionally,

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

Therefore,

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

Furthermore,

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

In the expression above,

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

$$
\mathrm{Cov}\left[\sum C_k^2,\sum_{m\neq r} Y_m Y_r\right]=\mathrm{Cov}\left[\sum Y_l^2,\sum_{i\neq j} C_i C_j\right]=0;
$$

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

Then we have

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

Putting the terms together, we have

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

C.1.2 Calculating *Cov* [Σ(C_i − Y_i)^2, ($\bar{C}$ − $\bar{Y}$)^2]—Next, we note that

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

where

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

and

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

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

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

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

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

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

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

Additionally,

$$E\left[Y_i^2\overline{C}\bar{Y}\right]=E\left[C_i^2\overline{C}\bar{Y}\right]=\frac{n+2}{n^2}\sigma_\mu^2(\sigma^2+\sigma_\mu^2);$$

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

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

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

Therefore,

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

So we have

$$\begin{aligned}
&\text{Cov }\left[\sum (C_i-Y_i)^2,(\overline{C}-\bar{Y})^2\right]=\sum \text{Cov }\left[(C_i-Y_i)^2,(\overline{C}-\bar{Y})^2\right]\\
&=\sum \Big(E(C_i^2-2C_iY_i+Y_i^2)(\overline{C}^2-2\overline{C}\bar{Y}+\bar{Y}^2)\\
&\quad -E(C_i^2-2C_iY_i+Y_i^2)E(\overline{C}^2-2\overline{C}\bar{Y}+\bar{Y}^2)\Big)\\
&=n\left(\frac{4(n+2)\sigma^4}{n^2}-2\sigma^2\frac{2\sigma^2}{n}\right)\\
&=\frac{8\sigma^4}{n}.
\end{aligned}$$

The variance of the estimator is then

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

C.2 Calculating *E*[*S*]

The expectation of the estimator is

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

where

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

and

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

Hence,

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

C.3 Calculating the MSE

The MSE of the estimator is then

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

The value of *a* that minimizes this MSE is

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

D Calculating *Var*[$\tilde{S}_{int}$]

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

The individual terms can be computed as follows:

$$
\begin{aligned}
\mathrm{Var}\left[\sum_{i=1}^{n} C_i^2\right] &= \sum_{i=1}^{n} \mathrm{Var}[C_i^2] \\
&= \sum_{i=1}^{n} \left(E[C_i^4] - (E[C_i^2])^2\right) \\
&= \sum_{i=1}^{n} \left(E[C_i^4] - (\mathrm{Var}[C_i] + (E[C_i])^2)^2\right) \\
&= \sum_{i=1}^{n} \left(E[C_i^4] - (\sigma^2 + \sigma_\mu^2 + \mu^2)^2\right) \\
&= \mathrm{nEC}_1^4 - n(\sigma^2 + \sigma_\mu^2 + \mu^2)^2.
\end{aligned}
$$

Assuming normality, we have

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

Assuming additionally that $\mu = 0$ and $\boldsymbol{\varepsilon} = 0$, we have

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

Since C_i and Y_i are symmetrically defined, we have

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

Next, from Appendix B,

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

Assuming normality, we have

$$
E[C_iY_i]^2 = \varepsilon + \sigma^4 + 3\sigma_\mu^4 + 6\mu^2\sigma_\mu^2 + \mu^4 + 2\sigma^2\sigma_\mu^2 + 2\sigma^2\mu^2;
$$

$$
E[C_iY_i] = \sigma_\mu^2 + \mu^2;
$$

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

Assuming additionally that μ = 0 and **ε** = 0, we have

$$E[C_iY_i]^2=(\sigma^2+\sigma_\mu^2)^2+2\sigma_\mu^4;$$

$$E[C_iY_i]=\sigma_\mu^2;$$

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

The covariance terms are computed as follows:

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

Assuming normality, we have

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

Assuming additionally that μ = 0 and **ε** = 0, we have

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

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

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

where

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

Assuming normality, we have

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

$$E[C_i^2]=\sigma^2+\sigma_\mu^2+\mu^2;$$

$$E[C_iY_i]=\sigma_\mu^2+\mu^2;$$

and therefore,

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

Assuming additionally that μ = 0 and ε = 0, we have

$$E[C_i^3Y_i]=3\sigma^2\sigma_\mu^2+3\sigma_\mu^4;$$

$$E[C_i^2]=\sigma^2+\sigma_\mu^2;$$

$$E[C_iY_i]=\sigma_\mu^2;$$

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

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

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

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

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

E Summary of mean and variance of the estimators

We summarize the mean and variance of the estimators in Table 6.

References

Elowitz MB, Levine AJ, Siggia ED, Swain PS. Stochastic gene expression in a single cell. Science. 2002; 297:1183–1186. [PubMed: 12183631]

Finkenstädt B, Woodcock DJ, Komorowski M, Harper CV, Davis JR, White MR, Rand DA. Quantifying intrinsic and extrinsic noise in gene transcription using the linear noise approximation: an application to single cell data. Ann. Appl. Stat. 2013; 7:1960–1982.

Hayes K. A geometrical interpretation of an alternative formula for the sample covariance. Am. Stat. 2011; 65:110–112.

Hilfinger A, Paulsson J. Separating intrinsic from extrinsic fluctuations in dynamic biological systems. Proc. Natl. Acad. Sci. USA. 2011; 108:12167–12172. [PubMed: 21730172]

James W, Stein C. Estimation with quadratic loss. Proc. Fourth Berkeley Symp. Math. Stat. Prob. 1961; 1:361–379.

Koeppl H, Zechner C, Ganguly A, Pelet S, Peter M. Accounting for extrinsic variability in the estimation of stochastic rate constants. Int. J. Robust Nonlin. 2012; 22:1103–1119.

Komorowski M, Miękisz J, Stumpf MP. Decomposing noise in biochemical signaling systems highlights the role of protein degradation. Biophys. J. 2013; 104:1783–1793. [PubMed: 23601325]

Rausenberger J, Kollmann M. Quantifying origins of cell-to-cell variations in gene expression. Biophys. J. 2008; 95:4523–4528. [PubMed: 18689455]

Schmiedel JM, Klemm SL, Zheng Y, Sahay A, Blüthgen N, Marks DS, van Oudenaarden A. MicroRNA control of protein expression noise. Science. 2015; 348:128–232. [PubMed: 25838385]

Sherman MS, Lorenz K, Lanier MH, Cohen BA. Cell-to-cell variability in the propensity to transcribe explains correlated fluctuations in gene expression. Cell Syst. 2015; 1:315–325. [PubMed: 26623441]

Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. Nat. Rev. Genet. 2015; 16:133–145. [PubMed: 25628217]

van Nimwegen, E. Inferring intrinsic and extrinsic noise from a dual fluorescent reporter. bioRxiv 049486. 2016. doi: http://dx.doi.org/10.1101/049486

Volfson D, Marciniak J, Blake WJ, Ostroff N, Tsimring LS, Hasty J. Origins of extrinsic variability in eukaryotic gene expression. Nature. 2006; 439:861–864. [PubMed: 16372021]

Yang S, Kim S, Lim YR, Kim C, An HJ, Kim J-H, Sung J, Lee NK. Contribution of RNA polymerase concentration variation to protein expression noise. Nat. Commun. 2014; 5:4761. [PubMed: 25175593]

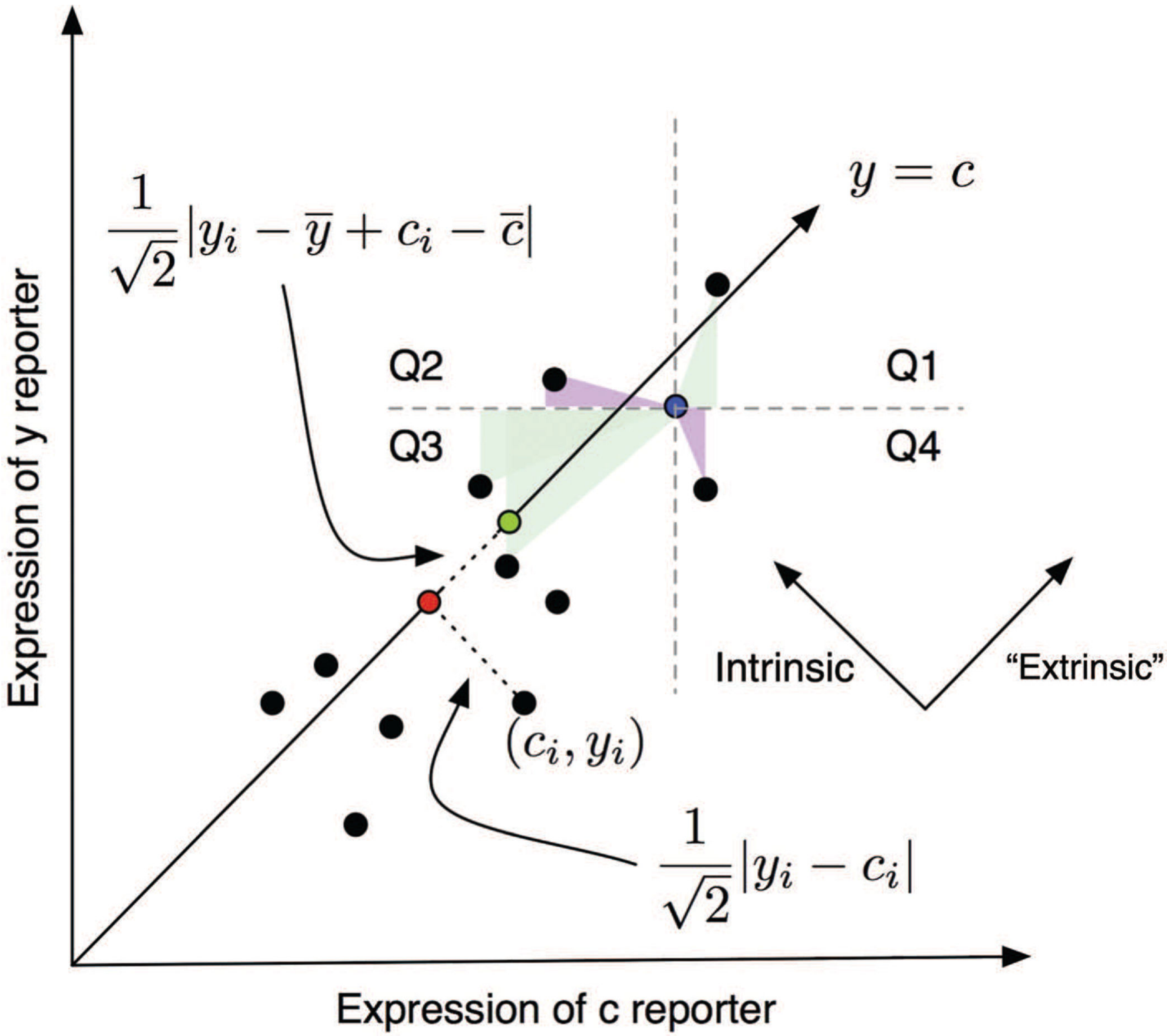

Figure 1.

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 $(\bar{c}, \bar{y})$ (or ($(\bar{c}+\bar{y})/\sqrt{2}, 0$) after projection) under the assumption that the two means are equal. See the main text for 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 (a randomly selected data 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 Elowitz et al. (2002) 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.

Table 1

Estimators for intrinsic and extrinsic noise. ρ is the correlation between the two reporters, and can be estimated by the sample correlation.

	Exact estimator for small *n*		**Large *n***
	Minimizing bias (Unbiased)	**Minimizing MSE**	
Intrinsic noise			
General	$\frac{1}{2(n-1)}\left[\sum_1^n (C_i - Y_i)^2 - n(\overline{C} - \bar{Y})^2\right]/(\overline{C}\bar{Y})$	$\frac{1}{2a}\left[\sum_1^n (C_i - Y_i)^2 - n(\overline{C} - \bar{Y})^2\right]/(\overline{C}\bar{Y})$, where $a=\frac{2n^3-7n+6}{2(n^2-n)}+\frac{2-n}{n^2-n}\frac{\rho}{1-\rho}+\frac{1}{2(n^2-n)}\left(\frac{\rho}{1-\rho}\right)^2$	$\frac{1}{2n}\left[\sum_1^n (C_i - Y_i)^2 - n(\overline{C} - \bar{Y})^2\right]/(\overline{C}\bar{Y})$
Assuming $\bar{C} = \bar{Y}$	$\frac{1}{2n}\sum_{i=1}^n (C_i - Y_i)^2/(\overline{C}\bar{Y})$ (ELSS estimator)	$\frac{1}{2(n+2)}\sum_{i=1}^n (C_i - Y_i)^2/(\overline{C}\bar{Y})$	$\frac{1}{2n}\sum_{i=1}^n (C_i - Y_i)^2/(\overline{C}\bar{Y})$ (ELSS estimator)
Extrinsic noise			
General	$\frac{1}{n-1}(\sum_{i=1}^n C_iY_i - n\overline{C}\bar{Y})/(\overline{C}\bar{Y})$	$\frac{1}{a}(\sum_{i=1}^n C_iY_i - n\overline{C}\bar{Y})/(\overline{C}\bar{Y})$, where $a = 1/\rho^2 + (n-1)(1 + 1/n)$	$\frac{1}{n}(\sum_{i=1}^n C_iY_i - n\overline{C}\bar{Y})/(\overline{C}\bar{Y})$ (ELSS estimator)

Table 2

Estimates of extrinsic noise in simulated data. Data were simulated under the hierarchical model, where the conditional distributions of the two reporters are identical. Two min-MSE estimators are applied, one using the true correlation, and the other the sample correlation. Mean estimates (standard deviation in parentheses) of intrinsic and extrinsic noise are summarized. Note that in order to compare the estimates with the true parameters, the estimates are unscaled (i.e. not divided by $\bar{c}\bar{y}$).

Simulation parameters	
Sample size (n)	50
Intrinsic noise (σ^2)	0.7
Extrinsic noise (σ_μ^2)	0.8
Distribution of means (G)	$\mathcal{N}(1, 0.8)$
Distribution of vars (H)	Constant: $\sum_i^2 = 0.7$
Distribution of $C_i \| Z_i$	$\mathcal{N}(M_i, 0.7)$
Distribution of $Y_i \| Z_i$	$\mathcal{N}(M_i, 0.7)$
No. of data sets	500
Extrinsic noise estimate	
Unbiased	0.80 (0.25; 0.0604)
minMSE (true corr)	0.73 (0.23; 0.0552)
minMSE (sample corr)	0.73 (0.24; 0.0634)
Asymptotic/ELSS	0.78 (0.06; 0.0582)

Table 3

Estimates of intrinsic and extrinsic noise in simulated data. Data were simulated under two schemes. The first scheme is consistent with the hierarchical model, where the conditional distributions of the two reporters are identical. Under the second scheme, the conditional distributions are different. Intrinsic and extrinsic noise are in fact not defined under the second scheme. Mean estimates (standard deviation in parentheses) of intrinsic and extrinsic noise are summarized. Note that in order to compare the estimates with the true parameters, the estimates are unscaled (i.e. not divided by $\bar{c}\bar{y}$).

	Identical distribution	**Different distributions**
Simulation parameters		
Sample size (*n*)	1000	1000
Intrinsic noise (σ^2)	0.7	0.7
Extrinsic noise (σ_μ^2)	0.8	0.8
Distribution of means (*G*)	$N(1, 0.8)$	$N(1, 0.8)$
Distribution of vars (*H*)	Constant: $\sum_i^2=0.7$	Constant: $\sum_i^2=0.7$
Distribution of $C_i\|Z_i$	$N(M_i, 0.7)$	$N(M_i, 0.7)$
Distribution of $Y_i\|Z_i$	$N(M_i, 0.7)$	$N(2M_i, 1.5 \times 0.7)$
No. of data sets	500	500
Sample correlation	0.53 (0.02)	0.60 (0.02)
Intrinsic noise $(\widehat{\sigma^2})$		
General		
Unbiased	0.70 (0.03)	1.54 (0.07)
minMSE	0.70 (0.03)	1.54 (0.07)
Asymptotic	0.70 (0.03)	1.54 (0.07)
Equal mean		
Unbiased/ELSS	0.70 (0.03)	2.04 (0.08)
minMSE	0.70 (0.03)	2.04 (0.08)
Asymptotic/ELSS	0.70 (0.03)	2.04 (0.08)
Extrinsic noise $(\widehat{\sigma_\mu^2})$		
Unbiased	0.80 (0.06)	1.60 (0.10)
minMSE	0.80 (0.06)	1.59 (0.10)
Asymptotic/ELSS	0.80 (0.06)	1.60 (0.10)
$\widehat{\sigma_\mu^2}/(\widehat{\sigma_\mu^2}+\widehat{\sigma^2})$		
General	0.53	0.51
Equal mean	0.53	0.44

Table 4

Re-analysis of published two-reporter experiment data. Summary statistics and estimates (×10^{-2}) of intrinsic and extrinsic noise are listed, using the estimators from Table 1.

	Elowitz et al. data		**Yang et al. data**	
	D22	**M22**	**Figure 3A**	**Normalized on log$_2$**
Sample means	CFP: 1	CFP: 1	CFP: 2660	CFP: 11
	YFP: 1	YFP: 1	mCherry: 3986	mCherry: 11
Sample correlation	0.50	0.49	0.86	0.86
Intrinsic noise				
General				
Unbiased	0.79	0.36	5.44	0.11
minMSE	0.78	0.35	5.44	0.11
Asymptotic	0.78	0.35	5.44	0.11
Equal mean				
Unbiased/ELSS	0.78	0.35	13.72	0.11
minMSE	0.78	0.35	13.72	0.11
Asymptotic/ELSS	0.78	0.35	13.72	0.11
Extrinsic noise				
Unbiased	0.78	0.34	30.29	0.68
minMSE	0.76	0.33	30.29	0.68
Asymptotic/ELSS	0.77	0.34	30.29	0.68
$\widehat{\sigma_\mu^2}/(\widehat{\sigma_\mu^2}+\widehat{\sigma^2})$				
General	0.50	0.49	0.85	0.86
Equal mean	0.50	0.49	0.69	0.86

Table 5

Noise estimates (×10^{-2}) based on subsets of published data. Similar to Table 4, we used the estimators from Table 1.

		Elowitz et al. data		Yang et al. data
		D22	**M22**	**Normalized on $\log_2$**
	Original sample size	284	250	40658
***n* = 200**				
Intrinsic noise				
General	Unbiased	0.79 (0.06)	0.36 (0.02)	0.11 (0.02)
	minMSE	0.78 (0.06)	0.35 (0.02)	0.11 (0.02)
	Asymptotic	0.78 (0.06)	0.35 (0.02)	0.11 (0.02)
Equal mean	Unbiased/ELSS	0.78 (0.06)	0.35 (0.02)	0.11 (0.02)
	minMSE	0.78 (0.06)	0.35 (0.02)	0.11 (0.02)
	Asymptotic/ELSS	0.78 (0.06)	0.35 (0.02)	0.11 (0.02)
Extrinsic noise	Unbiased	0.78 (0.07)	0.34 (0.02)	0.68 (0.09)
	minMSE	0.76 (0.07)	0.33 (0.02)	0.67 (0.08)
	Asymptotic/ELSS	0.78 (0.07)	0.34 (0.02)	0.68 (0.08)
***n* = 100**				
Intrinsic noise				
General	Unbiased	0.79 (0.13)	0.36 (0.04)	0.11 (0.03)
	minMSE	0.77 (0.12)	0.35 (0.04)	0.11 (0.03)
	Asymptotic	0.78 (0.12)	0.35 (0.04)	0.11 (0.03)
Equal mean	Unbiased/ELSS	0.78 (0.12)	0.35 (0.04)	0.11 (0.03)
	minMSE	0.77 (0.12)	0.35 (0.04)	0.11 (0.03)
	Asymptotic/ELSS	0.78 (0.12)	0.35 (0.04)	0.11 (0.03)
Extrinsic noise	Unbiased	0.77 (0.14)	0.34 (0.05)	0.69 (0.12)
	minMSE	0.73 (0.14)	0.32 (0.05)	0.67 (0.12)
	Asymptotic/ELSS	0.76 (0.14)	0.34 (0.05)	0.68 (0.12)
***n* = 50**				
Intrinsic noise				
General	Unbiased	0.78 (0.21)	0.36 (0.07)	0.11 (0.04)
	minMSE	0.75 (0.20)	0.35 (0.07)	0.11 (0.04)
	Asymptotic	0.77 (0.20)	0.35 (0.07)	0.11 (0.04)
Equal mean	Unbiased/ELSS	0.78 (0.21)	0.36 (0.07)	0.11 (0.04)
	minMSE	0.75 (0.20)	0.34 (0.07)	0.11 (0.04)
	Asymptotic/ELSS	0.78 (0.21)	0.36 (0.07)	0.11 (0.04)
Extrinsic noise	Unbiased	0.78 (0.24)	0.34 (0.09)	0.68 (0.16)

	Elowitz et al. data		Yang et al. data
	D22	**M22**	**Normalized on $\log_2$**
minMSE	0.70 (0.24)	0.30 (0.09)	0.65 (0.15)
Asymptotic/ELSS	0.76 (0.23)	0.33 (0.09)	0.66 (0.16)

Table 6

Mean and variance of the estimators in Table 1. Note that only the numerators of the estimators in the general forms are considered here; that is, scalar *a* can take different values depending on which specific estimator is of interest. Values of *a* can be found in Table 1. As in the main text, we assume normality of all distributions, and that μ = 0 and ε = 0, when deriving the mean and variance.

Estimator	Mean	Variance
Intrinsic noise		
General		
$\frac{1}{2a}(\sum_1^n (C_i - Y_i)^2 - n(\overline{C} - \bar{Y})^2)$	$\frac{n-1}{a}\sigma^2$	$\frac{1}{2a^2}\left(4n\sigma^4 + \frac{1}{n}\left((6-n)\sigma^4 + (4-2n)\sigma^2\sigma_\mu^2 + \sigma_\mu^4\right) - 8\sigma^4\right)$
Equal mean		
$\frac{1}{2a}\sum_{i=1}^n (C_i - Y_i)^2$	$\frac{n}{a}\sigma^2$	$\frac{2n}{a^2}\sigma^4$
Extrinsic noise		
$\frac{1}{a}(\sum_{i=1}^n C_i Y_i - n\overline{C}\bar{Y})$	$\frac{n-1}{a}\sigma_\mu^2$	$\frac{n-1}{a^2}(\sigma^2 + \sigma_\mu^2)^2 + \frac{(n-1)^2}{na^2}\sigma_\mu^4$