Maximum Likelihood Estimation with Push-Pull Noise Models

Jason W. Rocks and Pankaj Mehta

I. INTRODUCTION

In this set of notes, we describe the maximum likelihood estimation proceudre used to fit our push-pull amplifier models to experimental data.

II. PHOSPHORYLATED SUBSTRATE NOISE MODEL

First, we set up the noise model for the phosphorylated substrate. Unlike other the other species present the experiments, we cannot use GFP to empirically construct a noise model, so we must fit the noise model itself. The end result of this model will be the distribution of concentrations in antibody units conditioned on some underlying "true", but unknown concentration in units of GFP.

First, we assume the GFP and antibody measurements each follow log-normal distributions. We make this assumption based on biophysical arguments and the observed empirical forms of the noise models we have obtained for the other species. We define the means and variances as

$$\mathbf{E}[\log([\text{GFP}])] = \mu_{\text{GFP}} \qquad \text{Var}[\log([\text{GFP}])] = \sigma^2_{\text{GFP}} \tag{1}$$

$$\mathbf{E}[\log([\text{anti}])] = \mu_{\text{anti}} \qquad \text{Var}[\log([\text{GFP}])] = \sigma^2_{\text{anti}} \tag{2}$$

so that the distributions are then given by

$$P(\log([\text{GFP}])) = \frac{1}{\sqrt{2\pi\sigma^2_{\text{GFP}}}} \exp\left(-\frac{(\log([\text{GFP}]) - \mu_{\text{GFP}})^2}{2\sigma^2_{\text{GFP}}}\right) \tag{3}$$

$$P(\log([\text{anti}])) = \frac{1}{\sqrt{2\pi\sigma^2_{\text{anti}}}} \exp\left(-\frac{(\log([\text{anti}]) - \mu_{\text{anti}})^2}{2\sigma^2_{\text{anti}}}\right). \tag{4}$$

Next, we define the Pearson correlation coefficient between the two measurements,

$$\rho = \frac{\text{Cov}[\log([\text{GFP}]), \log([\text{anti}])]}{\sigma_{\text{GFP}}\sigma_{\text{anti}}}, \tag{5}$$

allowing us to define the covariance matrix between the two,

$$\mathbf{K} = \begin{pmatrix} \sigma^2_{\text{anti}} & \rho\sigma_{\text{GFP}}\sigma_{\text{anti}} \\ \rho\sigma_{\text{GFP}}\sigma_{\text{anti}} & \sigma^2_{\text{GFP}} \end{pmatrix}. \tag{6}$$

We also define the vectors

$$\vec{c} = \begin{pmatrix} \log([\text{anti}]) \\ \log([\text{GFP}]) \end{pmatrix} \tag{7}$$

$$\vec{\mu} = \begin{pmatrix} \mu_{\text{anti}} \\ \mu_{\text{GFP}} \end{pmatrix}. \tag{8}$$

Now we define the joint probability of the two measurements (the probability of a particular measurement in units of antibody coinciding with a particular measurement in units of GFP),

$$P(\log([\text{anti}]), \log([\text{GFP}])) = \frac{1}{\sqrt{(2\pi)^2 \det \mathbf{K}}} \exp\left(-\frac{1}{2}(\vec{c} - \vec{\mu})^T \mathbf{K}^{-1} (\vec{c} - \vec{\mu})\right) \tag{9}$$

Our goal is now to calculate the probability of a particular antibody measurement given a particular GFP measurement,

$$P(\log([\text{anti}]) | \log([\text{GFP}])) = \frac{P(\log([\text{anti}]), \log([\text{GFP}]))}{P(\log([\text{GFP}]))} \tag{10}$$

Plugging in all the quantities and simplifying we obtain the formula

$$P(\log([\text{anti}])|\log([\text{GFP}])) = \frac{1}{\sqrt{\sigma_{\text{anti}}^2(1-\rho)}}\exp\left(-\frac{\left[(\log([\text{anti}])-\mu_{\text{anti}})-\rho\frac{\sigma_{\text{anti}}}{\sigma_{\text{GFP}}}(\log([\text{GFP}])-\mu_{\text{GFP}})\right]^2}{2\sigma_{\text{anti}}^2(1-\rho)}\right) \tag{11}$$

$$= \frac{1}{\sqrt{\Sigma^2}}\exp\left(-\frac{[\log([\text{anti}])-A\log([\text{GFP}])-B]^2}{2\Sigma^2}\right) \tag{12}$$

where we have defined the noise parameters,

$$\Sigma^2 = \sigma_{\text{anti}}^2(1-\rho) \tag{13}$$

$$A = \rho\frac{\sigma_{\text{anti}}}{\sigma_{\text{GFP}}} \tag{14}$$

$$B = \mu_{\text{anti}} - \rho\frac{\sigma_{\text{anti}}}{\sigma_{\text{GFP}}}\mu_{\text{GFP}}. \tag{15}$$

III. MAXIMUM LIKELIHOOD ESTIMATION

Next, we construct our MLE loss function. First, we define vectors of concentrations for each species (total writer $[W_T]$, total eraser $[E_T]$, total substrate $[S_T]$, total phosphorylated substrate $[S_T^p]$, etc.) in both antibody and GFP units,

$$[\vec{X}]_{\text{anti}} = \begin{pmatrix}[W_T]_{\text{anti}}\\ [E_T]_{\text{anti}}\\ [S_T]_{\text{anti}}\\ [S_T^p]_{\text{anti}}\\ \vdots\end{pmatrix} \tag{16}$$

$$[\vec{X}]_{\text{GFP}} = \begin{pmatrix}[W_T]_{\text{GFP}}\\ [E_T]_{\text{GFP}}\\ [S_T]_{\text{GFP}}\\ [S_T^p]_{\text{GFP}}\\ \vdots\end{pmatrix} \tag{17}$$

We define S as the total number of species.

For each species we define a separate noise model $P[\log([X_j]_{\text{anti}})|\log([X_j]_{\text{GFP}})])$. In addition, we define our thermodynamic model in the following form:

$$F([\vec{X}]_{\text{GFP}};\Theta) = 0 \tag{18}$$

where we have define Θ as our fit parameters. A valid set of GFP values for each species will always satisfy this relation.

Now we can define the probability of a particular data point (in antibody units) given a set of fit parameters for the model as the following integral.

$$P([\vec{X}]_{\text{anti}}|\Theta) = \left(\prod_{j=1}^{S}\int d\log([X_j]_{\text{GFP}})P[\log([X_j]_{\text{anti}})|\log([X_j]_{\text{GFP}})]\right)\delta\Big[F([\vec{X}]_{\text{GFP}};\Theta)\Big]. \tag{19}$$

Note that we have incorporated the thermodynamic model as a delta function, so that only values of GFP are used that satisfy the model.

Here we make an approximation that will greatly simplify our fitting algorithm. Rather than integrating over the noise model for each species which can be very numerically costly, we will instead resample the concentrations of the species with empirical noise models from their joint probability distribution. In essence, we will convert species with empirical noise models from their antibody values to possible GFP values. By resampling many times, we obtain a new dataset that is much larger than the original where each of the original data points is copied many times, but

with different possible GFP values. The only species that we will continue to express in units of antibody will be the phosphorylated substrate $[S_p^T]$. The result is a simplified probability,

$$P([\vec{X}]_{\text{anti}}|\Theta) \approx \int d\log([S_T^p]_{\text{GFP}}) P[\log([S_T^p]_{\text{anti}})|\log([S_T^p]_{\text{GFP}})]\delta\Big[F([\vec{X}]_{\text{GFP}};\Theta)\Big] \tag{20}$$

Next, we express the thermodynamic model in the form

$$F([\vec{X}]_{\text{GFP}};\Theta) = [S_T^p]_{\text{GFP}} - S([\vec{X}]_{\text{GFP}};\Theta) \tag{21}$$

where S is the amount of phosphorylated substrated predicted by the model. Incorporating this we get

$$P([\vec{X}]_{\text{anti}}|\Theta) \approx \int d\log([S_T^p]_{\text{GFP}}) P[\log([S_T^p]_{\text{anti}})|\log([S_T^p]_{\text{GFP}})]\delta\Big[[S_T^p]_{\text{GFP}} - S([\vec{X}]_{\text{GFP}};\Theta)\Big] \tag{22}$$

$$= P\Big[\log([S_T^p]_{\text{anti}})|\log\Big(S([\vec{X}]_{\text{GFP}};\Theta)\Big)\Big] \tag{23}$$

Finally, we take the log and sum over all the resampled data point. Substituting our noise model, we obtain the loss function

$$\mathcal{L}(\times) = -\frac{1}{N_{\text{res}}}\sum_{i=1}^{N_{\text{res}}} \log P\Big[\log([S_T^p]_{\text{anti}})|\log\Big(S([\vec{X}]_{\text{GFP}};\Theta)\Big)\Big] \tag{24}$$

$$= \frac{1}{2\Sigma^2 N_{\text{res}}}\sum_{i=1}^{N_{\text{res}}}\Big[\log([S_T^p]_{\text{anti},i}) - A\log\Big(S([\vec{X}]_{\text{GFP},i};\Theta)\Big) - B\Big]^2 + \frac{1}{2}\log\left(\Sigma^2\right). \tag{25}$$

We will need to fit both the fit parameters for the thermodynamic model Θ as well as the parameters for the noise model Σ^2, A, and B.