Poisson regression theory for exponential growth detection

Daniel P. Rice

December 2022

Abstract

This document is an overview of theoretical results for using Poisson regression to detect exponential growth of a virus from a timeseries of k-mer counts in environmental metagenomic samples. It focuses only on the problem of estimating growth rate from the k-mer counts, not on any of the experimental or bioinformatic tasks in getting counts from samples. My objectives are:

1. To define the model and develop a consistent notation that we can use in future investigations.
2. To connect the Poisson regression model to the theory of exponential families, and in so doing to allow us to identify general properties and possible extensions.
3. On the other hand, to be sufficiently concrete that we can connect the results to real-world parameters like the total sequencing effort.
4. To build intuition for how aspects of the problem scale with these parameters.
5. To compare different approaches and lay the groundwork for quantitative performance assessment.
6. To develop our understanding of a simple model as a baseline for adding complications.

I begin with a description of a Poisson timeseries model. Then, I develop maximum likelihood, null-hypothesis testing, and Bayesian approaches to fitting the model to data. Finally, I propose future work to define performance metrics and apply them to these approaches.

Contents

1 Model

We are interested in estimating the abundance over time, $a(t)$, of a virus circulating in the population. Our observations consist of n metagenomic samples from the environment (e.g., wastewater) taken at a discrete set of times $\{t_1, \ldots, t_n\}$. We process the metagenomic samples to identify all k-mers of a given length and count their occurrence in each sample. Here we focus on the counts, $X = \{X_1, \ldots, X_n\}$, of a single k-mer derived from our virus of interest. At each sampling time t_i, we have a known sampling effort

λ_i. The $\{\lambda_i\}$ represent a conversion factor from the abundance of the virus to the count of the focal k-mer so that $\mathbb{E}[X_i] = \lambda_i a(t_i)$, accounting for the combined effects of sequencing depth, extraction and sequencing efficiency, and our bioinformatics pipeline. (Caveat: in practice the conversion factor is not fully known and may fluctuate due to compositional effects in the sample. However, it is reasonable to think of it as proportial to the sequencing depth.)

In the following we will model the counts, conditional on $a(t)$ and λ, as independent Poisson random variables. This assumes that each copy of the virus in the population has a small probability of contributing a read to the metagenomic pool and that they do so independently of one another. Violations of these assumptions will result in overdispersion of the counts relative to the Poisson distribution. Exploring the consequences of overdispersion for our methods will be the subject of another document.

Assuming that $X_i \sim \text{Poisson}(\lambda_i a(t_i))$ and the $\{X_i\}$ are mutually independent, the log-likelihood of our model is:

$$l(a; x, \lambda) = \sum_i \{\log(a(t_i))x_i - \lambda_i a(t_i)\} + \text{const.} \tag{1}$$

where the constant term depends on x and λ but not on a. Eq. 1 is an exponential family with n natural parameters $\{\log(a(t_i))\}$ and sufficient statistics $\{x_i\}$. Fitting this model would amount to estimating the abundance at each time using only the count data at that same time. There are two problems with this. First, scientifically, we are not interested only in the abundances at the observed times. Instead, we'd like to *characterize* the function $a(t)$, answering questions like, "Is the virus spreading?" or, "Does it fluctuate cyclically?" Alternately, we might want to *predict* the abundance in the future. Second, the model overfits to the Poisson counting noise at each time. As long as our sampling times aren't too widely spaced, we expect the abundance at nearby timepoints to be similar. Accordingly, we'd like to borrow information from multiple timepoints to inform our estimate of $a(t)$. This borrowing is the essence of regression.

In order to link our estimates of the abundance at different times, we expand $\log(a(t))$ in terms of a set of basis functions $\{b_r(t)\}$:

$$\log(a(t)) = \sum_r \beta_r b_r(t). \tag{2}$$

For standard Poisson regression, $b_0(t) = 1$ and $b_1(t) = t$. Here, we will work with a general basis as much as possible. This will show us which properties

depend on linearity of $\log(a(t))$ and which do not. It will also allow us to accommodate models with periodic fluctuations. Finally, it will show us how to select the basis with the best statistical properties among a set of equivalent parameterizations.

Substituting Eq. 2 into Eq. 1, our likelihood equation becomes

$$l(\beta; x, \lambda) = \sum_r \beta_r \sum_i b_r(t_i) x_i - \sum_i \lambda_i e^{\sum_r \beta_r b_r(t_i)} + \text{const.} \tag{3}$$

We can now see why we chose to expand $\log(a)$ rather than a: with this choice, our coefficients, β, are the natural parameters of an exponential family with sufficient statistics

$$T_r(X) = \sum_i b_r(t_i) X_i \tag{4}$$

and log-partition function

$$A(\beta) = \sum_i \lambda_i e^{\sum_r \beta_r b_r(t_i)}. \tag{5}$$

(It also lets us avoid the complications of restricting the abundances to be positive.) Note that the sufficient statistics (Eq. 4) are linear combinations of the X_i, namely their projections onto the basis functions $b_r(t)$.

In the following sections, we will derive statistical procedures for inferring the coefficients β, first in a maximum likelihood framework, then in a Bayesian one. We will see how to select basis functions and how the results depend on the sampling times and sequencing effort. Finally, we will analyze an edge case that maximum likelihood does not handle well and show how Bayesian inference regularizes our estimates.

2 Frequentist methods

2.1 Maximum likelihood parameter estimates

For an exponential family with natural parameters ν, sufficient statistics $T(X)$, and log-partition function $A(\nu)$, the log likelihood is

$$l(\nu; x) = \nu \cdot T(x) - A(\nu). \tag{6}$$

Setting $\nabla l = 0$, we find the maximum likelihood parameters $\hat{\nu}$ satisfy the equation

$$T(x) = \nabla A(\hat{\nu}) \tag{7}$$
$$= \mathbb{E}\left[T(X)|\hat{\nu}\right]. \tag{8}$$

That is, $\hat{\nu}$ are the parameters for which the expected values of the sufficient statistics equal their observed values. (The simplicity of this equation is one of the reasons to work with natural parameters.)

Applying this result to our Poisson regression model, we have

$$\frac{\partial A}{\partial \beta_r}(\hat{\beta}) = T_r(x) \tag{9}$$
$$\sum_i \lambda_i b_r(t_i) e^{\sum_{r'} \hat{\beta}_{r'} b_{r'}(t_i)} = \sum_i b_r(t_i) x_i. \tag{10}$$

Eq. 10 is an implicit equation for $\hat{\beta}$, so we typically can't solve it in closed form. Instead, we'll use Newton's method, or equivalently iterated reweighted least-squares (IRLS), to find $\hat{\beta}$ as a function of x.

Typically, we will want to compare our full model to a null model in which the abundance is constant over time. In this model $b_0(t) = 1$, $\beta_r = 0$ for $r > 0$, and our task is to find $\hat{\beta}_0$.

$$\sum_i \lambda_i e^{\hat{\beta}_0} = \sum_i x_i \tag{11}$$
$$e^{\hat{\beta}_0} = \frac{\sum_i x_i}{\lambda_{\text{tot}}}, \tag{12}$$

where $\lambda_{\text{tot}} = \sum_i \lambda_i$ is the total sequencing effort. This is reasonable because our null model for the abundance is $a(t) = e^{\hat{\beta}_0}$.

2.2 Quantifying uncertainty: Fisher information

The maximum likelihood estimator $\hat{\beta}$ represents a point estimate of our parameters. We would like to complement this with a measure of the uncertainty of this estimate. A common measure in likelihood theory is the Fisher information matrix, defined as

$$\mathcal{I}(\theta) = -\mathbb{E}\left[H_l(\theta; X)|\theta\right], \tag{13}$$

where H_l is the Hessian matrix of the log likelihood function and θ are parameters. The Fisher information represents the curvature of the log likelihood around the parameter estimate θ. Large Fisher information at the maximum likelihood estimate means that the data give a lot of information about the parameters and the estimate is sharp. Asymptotically, $\mathcal{I}^{-1}$ is the covariance matrix of the maximum likelihood estimator.

For an exponential family with natural parameters ν, the Hessian is independent of the data and we have:

$$\mathcal{I}_{r,s}(\nu) = \frac{\partial^2}{\partial \nu_r \partial \nu_s} A(\nu) \tag{14}$$

$$= \operatorname{cov}\left(T_r(X), T_s(X)\right). \tag{15}$$

In our Poisson regression model, this gives

$$\mathcal{I}_{r,s}(\beta) = \sum_i \lambda_i b_r(t_i) b_s(t_i) e^{\sum_{r'} \beta_{r'} b_{r'}(t_i)}. \tag{16}$$

Under the null model, this reduces to

$$\mathcal{I}_{r,s}(\beta_0) = e^{\beta_0} \sum_i \lambda_i b_r(t_i) b_s(t_i) \tag{17}$$

$$= e^{\beta_0} \lambda_{\text{tot}} \left\langle b_r, b_s \right\rangle_\lambda. \tag{18}$$

Eq. 18 represents the sum in the previous line as an inner product of the basis functions b_r and b_s with respect to the normalized weights $\lambda/\lambda_{\text{tot}}$. Note that the Fisher information increases proportionally to the expected total count.

2.3 Choosing basis functions

Eq. 18 suggests that we should choose our basis functions so that they are orthogonal with respect to the λ weights:

$$\mathcal{I}_{r,s}(\beta_0) = e^{\beta_0} \lambda_{\text{tot}} \|b_r\|_\lambda^2 \delta_{r,s} \tag{19}$$

Such an orthogonal basis will produce a diagonal Fisher information matrix under the null, which will simplify our analytical methods and make our numerical methods better conditioned.

For exponential growth detection, we are interested in a polynomial basis truncated at first order. That is,

$$b_0(t) = 1 \tag{20}$$
$$b_1(t) = t - c \tag{21}$$

where c is a constant to be determined. (We also have a coefficient multiplying t, but it is convenient to set it to be 1 so that β_1 is the growth rate). We choose c to satisfy the orthogonality condition:

$$\langle b_0, b_1 \rangle_\lambda = 0 \tag{22}$$
$$\sum_i \frac{\lambda_i}{\lambda_{\text{tot}}}(t_i - c) = 0 \tag{23}$$
$$c = \sum_i \frac{\lambda_i}{\lambda_{\text{tot}}} t_i. \tag{24}$$

Recognizing that the right-hand side of the last line is the average sampling time, weighted by the sampling intensity, we label it $\bar{t}$, so that

$$b_1(t) = t - \bar{t}. \tag{25}$$

With this choice of basis, our sufficient statistics and log-partition function are:

$$T_0(X) = \sum_i X_i \tag{26}$$
$$T_1(X) = \sum_i (t_i - \bar{t}) X_i \tag{27}$$
$$A(\beta) = \sum_i \lambda_i e^{\beta_0 + \beta_1 (t - \bar{t})}. \tag{28}$$

Eqs. 26–27 show that our sufficient statistics are the total observed counts, and the sum of observed counts weighted by the time they were observed.

Our ML equations (Eq. 10) become:

$$\sum_i \lambda_i e^{\hat{\beta}_0 + \hat{\beta}_1 (t_i - \bar{t})} = \sum_i x_i, \tag{29}$$
$$\sum_i \lambda_i (t_i - \bar{t}) e^{\hat{\beta}_0 + \hat{\beta}_1 (t_i - \bar{t})} = \sum_i (t_i - \bar{t}) x_i. \tag{30}$$

Assuming that we observe at least one count so that $T_0(x) > 0$, we can divide Eq. 30 by Eq. 29 to eliminate $\hat{\beta}_0$ and get an implicit equation for $\hat{\beta}_1$:

$$\frac{\sum_i \lambda_i (t_i - \bar{t}) e^{\hat{\beta}_1 (t_i - \bar{t})}}{\sum_i \lambda_i e^{\hat{\beta}_1 (t_i - \bar{t})}} = \frac{T_1(x)}{T_0(x)}. \tag{31}$$

That is, we can compute the maximum likelihood estimator of the growth rate from the time-weighted sum of counts T_1, normalized by the total count T_0.

To find the Fisher information under the null, all that remains is to find the norms of the basis functions

$$\|b_0\| = 1, \tag{32}$$

$$\|b_1\| = \sum_i \frac{\lambda_i}{\lambda_{\text{tot}}} (t_i - \bar{t})^2 = \sigma_t^2, \tag{33}$$

which we can substitute into Eq. 19 to get:

$$\mathcal{I}(\beta_0) = e^{\beta_0} \lambda_{\text{tot}} \begin{pmatrix} 1 & 0 \\ 0 & \sigma_t^2 \end{pmatrix}. \tag{34}$$

This confirms the intuition that our ability to estimate that non-growing virus are not growing increases with sampling depth and the dispersion of our samples in time.

Note that the orthogonality condition suggests other possible bases as alternative hypotheses. For example, if we are interested in ruling out periodic fluctuations, we might use an orthogonal periodic basis, e.g., a Fourier basis. On the other hand, we could develop methods based on wavelet bases that detect growth while making fewer assumptions about functional form.

2.4 Fisher information under the alternative hypothesis

So far we have only calculated the Fisher information under the null hypothesis $\beta_r = 0$ for $r > 0$. This is a useful baseline, but if we want a better measure of our uncertainty in our estimates, we also need the Fisher information under the alternative hypothesis. In our chosen orthogonal basis,

Eq. 16 becomes

$$\mathcal{I}_{0,0} = \sum_i \lambda_i e^{\hat{\beta}_0 + \hat{\beta}_1(t_i - \bar{t})} = T_0(x) \tag{35}$$

$$\mathcal{I}_{0,1} = \mathcal{I}_{1,0} = \sum_i \lambda_i (t_i - \bar{t}) e^{\hat{\beta}_0 + \hat{\beta}_1(t_i - \bar{t})} = T_1(x) \tag{36}$$

$$\mathcal{I}_{1,1} = \sum_i \lambda_i (t_i - \bar{t})^2 e^{\hat{\beta}_0 + \hat{\beta}_1(t_i - \bar{t})} \tag{37}$$

The second equalities in Eqs. 35 and 36 are due to the maximum likelihood equations (29–30). (They derive from the fact that $\frac{\partial A}{\partial \beta_0} = A$.)

Eq. 37 does not have a simple relationship to the sufficient statistics for general λ. However, we can gain some insight by considering the ratio $\mathcal{I}_{1,1}/T_0(x)$ in the limits of small and large $\hat{\beta}_0$. First, note that this ratio is independent of $\hat{\beta}_0$.

$$\frac{\mathcal{I}_{1,1}}{T_0(x)} = \frac{\sum_i \lambda_i (t_i - \bar{t})^2 e^{\hat{\beta}_0 + \hat{\beta}_1(t_i - \bar{t})}}{\sum_i \lambda_i e^{\hat{\beta}_0 + \hat{\beta}_1(t_i - \bar{t})}} \tag{38}$$

$$= \frac{\sum_i \lambda_i (t_i - \bar{t})^2 e^{\hat{\beta}_1(t_i - \bar{t})}}{\sum_i \lambda_i e^{\hat{\beta}_1(t_i - \bar{t})}} \tag{39}$$

in the limit $\hat{\beta}_1 \to 0$, we can expand in a power series:

$$\frac{\mathcal{I}_{1,1}}{t_0(x)} = \frac{\sum_i \lambda_i (t_i - \bar{t})^2 \left(1 + \hat{\beta}_1(t_i - \bar{t}) + \frac{1}{2}\hat{\beta}_1^2 (t_i - \bar{t})^2 + O(\hat{\beta}_1^3)\right)}{\sum_i \lambda_i \left(1 + \hat{\beta}_1(t_i - \bar{t}) + \frac{1}{2}\hat{\beta}_1^2 (t_i - \bar{t})^2 + O(\hat{\beta}_1^3)\right)} \tag{40}$$

$$= \sigma_t^2 + \mu_t^{(3)} \hat{\beta}_1 + (\mu_t^{(4)} - \sigma_t^4)\hat{\beta}_1^2 + O(\hat{\beta}_1^3), \tag{41}$$

where $\mu_t^{(k)}$ is the k-th central moment of the λ-weighted sampling times.

In the limit $\hat{\beta}_1 \to \infty$, the final sampling time dominates the sum and we have:

$$\frac{\mathcal{I}_{1,1}}{T_0(x)} = \frac{\lambda_n (t_n - \bar{t})^2 e^{\hat{\beta}_1(t_n - \bar{t})} + o(e^{\hat{\beta}_1(t_n - \bar{t})})}{\lambda_n e^{\hat{\beta}_1(t_n - \bar{t})} + o(e^{\hat{\beta}_1(t_n - \bar{t})})} \tag{42}$$

$$\to (t_n - \bar{t})^2. \tag{43}$$

That is, the ratio is bounded for large $\hat{\beta}_1$.

Putting these results together gives us that the Fisher information under the alternative hypothesis has the form

$$\mathcal{I}(\hat{\beta}) = \begin{pmatrix} T_0(x) & T_1(x) \\ T_1(x) & f(\hat{\beta}_1)T_0(x) \end{pmatrix}, \tag{44}$$

$$\tag{45}$$

where

$$f(\hat{\beta}_1) = \sigma_t^2 + \mu_t^{(3)}\hat{\beta}_1 + (\mu_t^{(4)} - \sigma_t^4)\hat{\beta}_1^2 + O(\hat{\beta}_1^3), \qquad \hat{\beta}_1 \to 0, \tag{46}$$

$$\to (t_n - \bar{t})^2, \qquad \hat{\beta}_1 \to \infty. \tag{47}$$

For $T_1(x) = 0$, $\hat{\beta}_1 = 0$ and we recover the null result. For $T_1(x) > 0$, $\hat{\beta}_1 > 0$, we get that the asymptotic variance of $\hat{\beta}_1$ increases with $\hat{\beta}_1$, but plateaus.

2.5 A problematic edge case: the first observation

In this section, we examine an edge case that causes difficulty for maximum likelihood estimation. Consider observing a sequence of zero counts followed by a single non-zero observation, i.e., $x = (0, \ldots, 0, x_n)$. In the context of continuous monitoring, this is the data we'd expect the first time we observed a k-mer.

In this case, the sufficient statistics are

$$T_0(x) = x_n \tag{48}$$

$$T_1(x) = (t_n - \bar{t})x_n \tag{49}$$

Substituting into Eq. 31 gives

$$\frac{\sum_i \lambda_i (t_i - \bar{t}) e^{\hat{\beta}_1 (t_i - \bar{t})}}{\sum_i \lambda_i e^{\hat{\beta}_1 (t_i - \bar{t})}} = \frac{(t_n - \bar{t})x_n}{x_n} \tag{50}$$

$$= t_n - \bar{t}, \tag{51}$$

which is solved by letting $\hat{\beta}_1 \to \infty$. In order to match the finite T_0, $\hat{\beta}_0$ must go to $-\infty$ like

$$\hat{\beta}_1 \sim \log(x_n/\lambda_n) - (t_n - \bar{t})\hat{\beta}_1. \tag{52}$$

As in the logistic regression when there is perfect separation between cases, the model tries to fit this data with an infinitely steep curve that passes through all the points exactly.

The Fisher information matrix is

$$\mathcal{I}(\hat{\beta}) = x_n \begin{pmatrix} 1 & (t_n - \bar{t}) \\ (t_n - \bar{t}) & (t_n - \bar{t})^2 \end{pmatrix}. \tag{53}$$

Note that all terms are finite.

Thus we have a case in which the maximum likelihood estimator takes on an extreme result based on relatively little information. Worse, the Fisher information is a poor guide to our uncertainty about the parameter estimates. Intuitively, there are a wide range of parameter values that are consistent with this data, but the Fisher information suggests that we should be confident that the parameters are infinite. Regression software that outputs approximate p-values based on the Fisher information will be infinitely overconfident that the null hypothesis can be rejected. In the next section, we'll explore hypothesis testing in more detail.

2.6 Null hypothesis testing

One way of framing exponential growth detection is as a test of the null hypothesis that $\hat{\beta}_1 = 0$. In frequentist statistics a null hypothesis test is defined by a choice of *test statistic*, a function of the data that is designed to quantify the degree to which the data would be surprising if the null hypothesis were true. To apply the test, the user computes the test statistic from the observed data and uses it to calculate (or more typically approximate) a *p-value*, the probability of observing a more extreme value of the test statistic under the null hypothesis. If the p-value falls below a pre-specified threshold, the null hypothesis is rejected. For a given problem, there are many possible choices of test statistic, which vary in their ease of calculation and ability to discriminate between models. In this section, we'll examine the properties of three different test statistics.

A particular challenge for null hypothesis testing in the context of exponential growth detection is the multiple hypothesis testing burden. A metagenomic sequence dataset will have a very large number of k-mers. If we were to set a p-value threshold at a typical value of 0.05, we would expect to reject the null hypothesis for one out of every twenty k-mers, even if none were growing at all. This scale of false positives would likely swamp

our post-processing pipeline. To reduce the number of false positives, we could set a higher p-value threshold for our test. However, the asymptotic approximations to the null distribution of the test statistic break down in the extreme tails of the distribution, even in very large samples. Worse, any model misspecification issues are likely to be worse in the tails as well. Therefore, it probably makes more sense to avoid p-values and either to construct a decision procedure based directly on the test statistic, to use Bayesian regularization (see below), and/or to design a loss function based on particular concrete objectives. (Note: any procedure we develop should take into account that the data is arriving as a stream rather than just one time.)

2.6.1 Z-test

The most obvious choice of test statistic in a regression context is the regression coefficient β itself. In particular, we could ask the probability that the maximum likelihood estimator $\hat{\beta}_1$ would be greater than inferred from the data if the true $\hat{\beta}_1 = 0$. This is the intuition behind the Z-test, which uses the test statistic

$$Z = \frac{\hat{\beta}_1}{\sqrt{\mathcal{I}_{1,1}(\hat{\beta})}}. \tag{54}$$

Likehihood theory shows that in the limit of large datasets Z is asymptotically normal. When the `statsmodels` Python package reports Z-scores and p-values, this what is it is using. (Note: we have shown that the Fisher information depends on the parameters, so that the variance of $\hat{\beta}_1$ is different under the null and alternative hypotheses. The software uses the Fisher information evaluated at $\hat{\beta}$, which overestimates the variability in $\hat{\beta}_1$ under the null. It would be better to normalize by $\mathcal{I}_{1,1}(\hat{\beta}_{\text{null}})$.)

There are a few difficulties with working with the Z-test. First, $\hat{\beta}_1$ is a complicated, non-linear transformation of the data, making it relatively slow to calculate and impossible to get exact null distributions for Z. Second, we have seen that in the edge case examined above, $\hat{\beta}_1 \to \infty$, while the Fisher information is finite. The Z-test will thus report infinitely small p-values, even if there is only a single count of the k-mer in the dataset. This suggests that Z does a poor job of summarizing the strength of evidence against the null hypothesis when the data is marginal.

2.6.2 Likelihood ratio test

The likelihood ratio test compares the likelihood of the ML parameters given the data under the null and alternative hypotheses. It uses the test statistic

$$\Lambda = 2\left(l_{\text{alt}}(\hat{\beta}_{\text{alt}}) - l_{\text{null}}(\hat{\beta}_{\text{null}})\right). \tag{55}$$

In our model

$$\Lambda = 2\left((\hat{\beta}_{\text{alt}} - \hat{\beta}_{\text{null}}) \cdot T(x) - A_{\text{alt}}(\hat{\beta}_{\text{alt}}) + A_{\text{null}}(\hat{\beta}_{\text{null}})\right) \tag{56}$$

$$= 2(\hat{\beta}_{\text{alt}} - \hat{\beta}_{\text{null}}) \cdot T(x), \tag{57}$$

where the second equality comes from the fact that $A(\hat{\beta}) = \frac{\partial A}{\partial \beta_0}(\hat{\beta}) = T_0(x)$ under both models. Like Z, Λ is asymptotically standard normal under the null model.

Because it depends on $\hat{\beta}$, the likelihood ratio test shares the downsides of complexity with the Z-test. On the other hand, it handles the first-observation problem more gracefully. To see this, apply Eq. 52, to get

$$\Lambda \to 2\left(\log(x_n/\lambda_n)x_n - \log(x_n/\lambda_{\text{tot}})x_n\right) \tag{58}$$

$$= 2\log(\lambda_{\text{tot}}/\lambda_n)x_n \tag{59}$$

$$\approx 2\log(n)x_n, \tag{60}$$

which is finite.

2.6.3 Score test

The score test is an approximation to the likelihood ratio test that is simpler to calculate. Its test statistic is based on the gradient of the log-likelihood (sometimes known as the 'score'):

$$S^2 = (\nabla l)^T \mathcal{I}^{-1} \nabla l, \tag{61}$$

where all components are evaluated at the parameters of the null hypothesis. Substituting our previous results, gives

$$S^2(x) = \frac{\left(T_0(x) - \lambda_{\text{tot}} e^{\hat{\beta}_0}\right)^2 + T_1^2(x)/\sigma_t^2}{\lambda_{\text{tot}} e^{\hat{\beta}_0}}. \tag{62}$$

$$= \frac{T_1^2(x)}{T_0(x)\sigma_t^2} \tag{63}$$

Under the null hypothesis S^2 is asymptotically χ-squared distributed with one degree of freedom so that S is standard Gaussian.

The score test has the advantage that its test statistic can be computed directly from the count data without fitting the model. This makes it promising as an initial screening step to rule out k-mers that have very little evidence of exponential growth. It also makes it faster to simulate the distribution of the statistic under different hypotheses.

In the first-observation case, we have a finite test statistic:

$$S^2 = \frac{(t_n - \bar{t})^2}{\sigma_t^2} x_n. \tag{64}$$

Note that unlike the likelihood ratio statistic, this does not depend directly on the number of sampled points. In a sense, the score test forgets this information by looking only at the aggregate sufficient statistics. In doing so, it is more conservative about this case than the likelihood ratio test. (S also scales with $\sqrt{x_n}$, unlike Λ, which is proportional to x_n, so the score test is more conservative about larger counts at the first observation.)

2.6.4 Possible follow-up work on hypothesis tests

It may be worthwhile to follow up on the above section with a more thorough (and largely numerical) analysis of the merits of the different tests and their test statistics. Possible issues to investigate:

1. Calibration of asymptotic p-values: how do the true distributions of the test statistics compare to their asymptotic Gaussian distributions? Do any converge to Gaussian faster than the others?

2. Power: How good is the test at detecting growth as a function of growth rate and false-positive rate (e.g., precision-recall curves)?

3. Robustness: How sensitive is the test to model mis-specification (e.g., over-dispersion of counts, periodic fluctuations)?

4. Extensibility: How easy is it to extend the test to deal with complications like the serial acquisition of data?

2.7 Elastic net regularization [TODO]

2.8 Uniform-sampling approximation

Up until this point, we have worked with arbitrary λ and sampling times. It can be useful for intuition to consider the case of evenly spaced samples with equal intensity λ_{tot}/n on the interval $[-\Delta t/2, \Delta t/2]$. We'll further assume that we have dense sampling so that $n \gg 1$. To simplify the expressions we'll neglect terms of order n^{-1} and higher. [Note: we can compute more terms in the asympotic series in n^{-1}, to capture the dependence on n, but I don't think it's worth it.] With these assumptions, we can make the substitutions:

$$\frac{2t_i}{\Delta t} \to t \tag{65}$$

$$\bar{t} \to 0 \tag{66}$$

$$\frac{\beta_1 \Delta t}{2} \to \beta_1 \tag{67}$$

$$\sum_{i=1}^{n} \lambda_i \to \int_{-1}^{1} dt \frac{\lambda_{\text{tot}}}{2} \tag{68}$$

Note that we're now measuring time in terms of the total interval of our sampling, so that β_1 is a dimensionless quantity. These substitutions send the log-partition function to

$$A(\beta) \to \int_{-1}^{1} dt \frac{\lambda_{\text{tot}}}{2} e^{\beta_0 + \beta_1 t} \tag{69}$$

$$= \lambda_{\text{tot}} e^{\beta_0} \frac{\sinh(\beta_1)}{\beta_1}. \tag{70}$$

We now have the gradient:

$$\frac{\partial A}{\partial \beta_0} = A \tag{71}$$

$$\frac{\partial A}{\partial \beta_1} = \lambda_{\text{tot}} e^{\beta_0} \frac{\beta_1 \cosh \beta_1 - \sinh \beta_1}{\beta_1^2}. \tag{72}$$

This yields the normalized max likelihood function for $\hat{\beta}_1$

$$\frac{T_1(x)}{T_0(x)} = \coth \hat{\beta}_1 - \frac{1}{\hat{\beta}_1} \begin{cases} = \frac{\hat{\beta}_1}{3} - \frac{\hat{\beta}_1^3}{45} + O(\hat{\beta}_1^5) \\ \sim 1 - \frac{1}{\hat{\beta}_1}, \hat{\beta}_1 \to \infty. \end{cases} \tag{73}$$

Finally, we have the Fisher information:

$$\mathcal{I}(\beta) = \lambda_{\text{tot}} e^{\beta_0} \begin{pmatrix} \frac{\sinh\beta_1}{\beta_1} & \frac{\beta_1\cosh\beta_1 - \sinh\beta_1}{\beta_1^2} \\ \frac{\beta_1\cosh\beta_1 - \sinh\beta_1}{\beta_1^2} & \frac{(\beta_1^2+2)\sinh\beta_1 - 2\beta_1\cosh\beta_1}{\beta_1^3}. \end{pmatrix} \tag{74}$$

Some algebra will show that these results agree with the general λ results above.

3 Bayesian inference

3.1 The conjugate prior and update rule

In Bayesian inference, the *conjugate prior* for a given likelihood is the family of distributions that is closed under Bayesian updating. That is, if the prior is a member of the conjugate family, the posterior is also a member.

For a likelihood in an exponential family with natural parameters η, sufficient statistics T and log-partition function $A(\eta)$, the conjugate prior is given by

$$\log p_\pi(\eta;\chi,\nu) = \eta\cdot\chi - \nu A(\eta) - \log Z(\chi,\nu), \tag{75}$$

where χ and ν are hyperparameters, and Z is the partition function of the prior. When we make a single observation x, we update according to Bayes rule to get the posterior:

$$\log p_{\text{post}}(\eta;\chi,\nu,x) = l(x;\eta) + \log p_\pi(\eta;\chi,\nu) \tag{76}$$

$$= \eta\cdot(\chi + T(x)) - (\nu+1)A(\eta) - \log Z \tag{77}$$

where Z is the new partition function and does not depend on η. Comparing Eq. 75 and Eq. 77, we see that the posterior indeed belongs to the same family as the prior and that the hyper parameters are updated according to the update rule

$$\chi' = \chi + T(x) \tag{78}$$

$$\nu' = \nu + 1. \tag{79}$$

If we had multiple observations, we could iterate this process to get the final posterior including all the data.

Above, we showed that our Poisson regression model is an exponential family with natural parameters β. Thus, our conjugate prior is

$$\log p_\pi(\beta; \chi, \nu) = \beta_0 \chi_0 + \beta_1 \chi_1 - \nu \lambda_{\text{tot}} e^{\beta_0} f(\beta_1) - \log Z, \tag{80}$$

$$f(\beta_1) = \sum_i \frac{\lambda_i}{\lambda_{\text{tot}}} e^{\beta_1 (t_i - \bar{t})}. \tag{81}$$

By examining the update rules, we can interpret our hyperparameters in terms of pseudocounts. That is, ν can be thought of as an effective number of observations (of the entire timeseries) that contributed to the prior, χ_0 as the total number of counts we saw in those observations, and χ_1 as the total time-weighted counts. Larger values of the hyperparameters will tend to have a stronger influence on the posterior, requiring more data to overcome them. Unless we have a reason to believe that most k-mers tend to increase or decrease on net, symmetry suggests we should choose $\chi_1 = 0$.

The rest of this section will be devoted to examining the properties of this conjugate distribution in order to inform our choice of hyperparameters and interpretation of the posterior. First, we will look at the conditional distribution of β_0, then we will use this result to find the marginal distribution of β_1, which is obviously more important for exponential growth detection.

3.2 Conditional distribution of β_0

To find the conditional distribution of β_0 (that is, the distribution conditioned on a particular value of β_1), we make the substitution $y = e^{\beta_1}$. Applying the change of variables formula, we find:

$$p(y; \chi, \nu, \beta_1) = p_\pi(\log y, \beta_1; \chi, \nu) \frac{d}{dy} \log y \tag{82}$$

$$= \frac{1}{Z} y^{\chi_0 - 1} e^{-\nu \lambda_{\text{tot}} f(\beta_1) y} e^{\beta_1 \chi_1}. \tag{83}$$

We can recognize this as a Gamma distribution with shape parameter χ_0 and scale parameter $\nu \lambda_{\text{tot}} f(\beta_1)$. This is unsurprising because the Gamma is the conjugate prior of the Poisson likelihood and $\lambda_{\text{tot}} f(\beta_1) e^{\beta_0}$ is the rate parameter for the total count, which is Poisson-distributed under our model. (The sum of independent Poissons is itself Poisson.)

We now have some more information about the interpretation of our hyperparameters. The shape parameter χ_0 controls the behavior of the Gamma distribution around zero:

- For $0 < \chi_0 < 1$, $p(y)$ has a power-law singularity at zero. It says that we expect to see y at all scales as we approach zero.
- For $\chi_0 = 1$, $p(y)$ is an exponential distribution.
- For $\chi_1 > 1$, $p(0) = 0$ and there is an internal mode. That is, the likely values of y are clustered around a finite value.

Conditional on zero growth, $\beta_1 = 0$, we have $f(0) = 1$, so

$$p(y; \chi, \nu, 0) = \frac{1}{Z} y^{\chi_0 - 1} e^{\nu \lambda_{\rm tot} y} \tag{84}$$

So the expected total count for steady-state counts is χ_0/ν. Together, these facts could be used in an empirical Bayes setting to choose hyperparameters, although we will see below some other considerations that may be more important.

3.3 Marginal distribution of β_1

The fact that $y = e^{\beta_0}$ is conditionally Gamma-distributed is convenient, because it allows us to integrate out β_0 to find the marginal distribution of β_1. This is useful because our ultimate decision rule may depend only on β_1, so that β_0 is a nuisance parameter.

Integrating Eq. 83 over y and using the definition of the Gamma function, we get the marginal distribution:

$$p_\pi(\beta_1; \chi, \nu) = \frac{1}{Z(\chi, \nu)} \int_0^\infty dy y^{\chi_0 - 1} e^{-\nu \lambda_{\rm tot} f(\beta_1) y} e^{\beta_1 \chi_1} \tag{85}$$

$$= \frac{1}{Z(\chi)} \exp\left(-\chi_0 \log f(\beta_1) + \chi_1 \beta_1\right). \tag{86}$$

Interestingly, the marginal distribution of β_1 is independent of the hyperparameter ν.

Because of the complicated function f, this is not a named distribution. (Although in the uniform sampling limit—see above—it is tantalizingly similar to a skewed generalized hyperbolic secant distribution.) Fortunately, we can say quite a bit about it, including characterizing its mode, its tail behavior, and its convergence to a Gaussian distribution.

3.3.1 Bayesian regularization: the posterior mode

A common Bayesian point estimate is the posterior mode of a parameter. To find the mode, set the derivative of the log posterior to zero and solve for β_1:

$$\frac{d}{d\beta_1}\log p = -\chi_0\frac{f'(\beta_1^*)}{f(\beta_1^*)} + \chi_1 = 0 \tag{87}$$

$$\frac{\chi_1}{\chi_0} = \frac{f'(\beta_1^*)}{f(\beta_1^*)} \tag{88}$$

$$= \frac{\sum_i \frac{\lambda_i}{\lambda_{\text{tot}}}(t_i - \bar{t})e^{\beta_1^*(t_i-\bar{t})}}{\sum_i \frac{\lambda_i}{\lambda_{\text{tot}}}e^{\beta_1^*(t_i-\bar{t})}}. \tag{89}$$

We recognize this as the maximum likelihood estimator (Eq. 31) with χ playing the role of $T(x)$.

Using our update rule, we find that the posterior mode is determined by

$$\frac{\chi_1'}{\chi_0'} = \frac{\chi_1 + T_1(x)}{\chi_0 + T_0(x)}. \tag{90}$$

This is our first illustration of Bayesian inference as regularization. When the evidence is weak compared to the prior, the prior dominates and sets posterior mode. In contrast, when the evidence is much stronger than the prior, the effect of the prior is minimal. In intermediate cases, the mode is pulled toward the prior, but not all the way.

For example, consider the first-observation problem described above. In the maximum-likelihood framework, the ratio $\frac{T_1}{T_0}$ achieves its maximum possible value and $\hat{\beta}_1 \to \infty$, even with only one observed count. In Bayesian model, the prior, through χ_0, pulls ratio downward. The effect of χ_0 on the posterior mode is the same as the effect of χ_0 counts at the midpoint of the timeseries on the maximum likelihood estimator. The larger χ_0 is, the more observations at the final timepoint are necessary to get a large value of β_1^*.

3.3.2 Small β_1 approximation

When β_1 is small, we can simplify Eq. 86 by expanding $\log f$ about zero. This is particularly relevant when χ_1 is small so that the posterior mode is

near zero. To second order in β_1, we have:

$$\log p_\pi(\beta_1;\chi) = -\chi_0 \sum_i \frac{1}{2}(t_i - \bar{t})^2\beta_1^2 + \chi_1\beta_1 + \text{const.} \tag{91}$$

$$= -\frac{\chi_0\sigma_t^2}{2}\beta_1^2 + \chi_1\beta_1 + \text{const.} \tag{92}$$

$$= -\frac{\chi_0\sigma_t^2}{2}(\beta_1 - \frac{\chi_1}{\chi_0\sigma_t^2})^2 - \log Z. \tag{93}$$

That is, for small β_1, $p_\pi(\beta_1)$ is approximately Gaussian with mean $\frac{\chi_1}{\chi_0\sigma_t^2}$ and variance $\frac{1}{\chi_0\sigma_t^2}$.

If we apply the update rule with the hyperparameter $\chi_1 = 0$, we, get an approximately Gaussian posterior with mean $\frac{T_1(x)}{(\chi_0+T_0(x))\sigma_t^2}$ and variance $\frac{1}{(\chi_0+T_0(x))\sigma_t^2}$. Comparing to Eq. 63, we can see that the posterior probability that $\beta_1 \leq 0$ under this approximation is equivalent to the p-value of the score test but with T_0 replaced with χ_0+T_0. This gives us a Bayesian interpretation of the score test.

3.3.3 Tail behavior

Now, we consider the opposite limit, $|\beta_1| \to \infty$. In that case, the sum in f is dominated by either the first or last term so that

$$\log p \sim \begin{cases} -|\beta_1|(\chi_0(\bar{t} - t_i) + \chi_1^*) & \text{as } \beta_1 \to -\infty \\ -\beta_1(\chi_0(t_n - \bar{t}) - \chi_1) & \text{as } \beta_1 \to \infty. \end{cases} \tag{94}$$

Thus, while our small-β_1 approximation was Gaussian, the tails of the distribution are actually exponential. The Gaussian approximation will underestimate the extreme tail probabilities oft he posterior.

Incidentally, Eq. 94 places bounds on our hyperpriors. We must have $\chi_0(t_i - \bar{t}) < \chi_1 < \chi_0(t_n - \bar{t})$ in order for the tails to go to zero. Otherwise, we will have an improper prior. Examining the sufficient statistics T will show that if these conditions are met for the prior, they will be met for the posterior as well.

3.3.4 Gaussian approximation for large χ_0

In the previous two sections, we looked at the shape of the marginal distribution of β_1 for particular regions $\mathbb{R}$, namely near the origin and around

infinity. In this section, we will look at the behavior of the distribution for $\chi_0 \gg 1$. This condition could come about either from a strong prior, a large observation of $T_0(x)$, or some combination.

To explore this limit, we will show how to calculate an asymptotic series for the partition function $Z(\chi)$ as $\chi_0 \to \infty$. We focus on Z for two reasons. First, the partition function is related to the moment-generating function, and we can derive all moments of p_π from Z. Second, Z is the normalizing factor for p_π. If we are interested in calculating extreme tail probabilities in the exponential regime (Eq. 94), we will need to divide by Z, which depends on the entire distribution.

Integrating Eq. 86 over β_1 and setting the integral to 1, we have

$$Z(\chi) = \int_{-\infty}^{\infty} e^{\chi_0(-\log f(\beta_1) + \frac{\chi_1}{\chi_0}\beta_1)} d\beta_1. \tag{95}$$

The integrand has a peak at β_1^*, as we showed above, and as $\chi_0 \to \infty$ it becomes more sharply peaked around this value. We can thus use Laplace's method to approximate the integral by expanding the integrand about β_1^*.

Making the substitutions $u = \beta_1 - \beta_1^*$, $\phi = \log f(\beta_1) - \frac{\chi_1}{\chi_0}\beta_1$,

$$Z(\chi) = \int_{-\infty}^{\infty} e^{-\chi_0\left(\phi(\beta_1^*) + \frac{u^2}{2}\phi''(\beta_1^*) + \sum_{k=3}^{\infty}\frac{u^k}{k!}\phi^{(k)}(\beta_1^*)\right)} du \tag{96}$$

$$= e^{-\chi_0\phi(\beta_1^*)} \int_{-\infty}^{\infty} e^{-\frac{u^2}{2}\chi_0\phi''(\beta_1^*)} \exp\left[-\sum_{k=3}^{\infty}\frac{u^k}{k!}\chi_0\phi^{(k)}(\beta_1^*)\right] du \tag{97}$$

$$= e^{-\chi_0\phi(\beta_1^*)} \sqrt{\frac{1}{\phi''(\beta_1^*)\chi_0}} \int_{-\infty}^{\infty} e^{-\frac{u^2}{2}} \exp\left[-\sum_{k=3}^{\infty}\frac{u^k}{k!}\chi_0^{-\frac{k}{2}+1}\frac{\phi^{(k)}(\beta_1^*)}{\phi''(\beta_1^*)^{k/2}}\right] du \tag{98}$$

$$\sim e^{-\chi_0\phi(\beta_1^*)} \sqrt{\frac{2\pi}{\phi''(\beta_1^*)\chi_0}}. \tag{99}$$

To leading order, this is the partition function of a Gaussian distribution with variance $\frac{1}{\chi_0\phi''(\beta_1^*)}$.

Higher-order corrections can be obtained by expanding the exp term in Eq. 98 in powers of u and integrating each term with respect to $e^{-\frac{u^2}{2}}$, the standard Gaussian measure. Each integral will be proportional to the moments of the standard Gaussian of order k, so only even terms will contribute to

the sum. Thus, the lowest-order correction will be $k = 4$, followed by $k = 6$, etc., so that we obtain a power series in χ_0^{-1}.

We can use this information in two ways. First, we can use the posterior variance as a summary of our posterior uncertainty about β_1^*. Second, we can use our approximation of Z to calculate the tail probabilities using a combination of Gaussian and exponential approximation as appropriate.

4 Future directions: assessing performance

The results in this document represent an overview of the basic theory of frequentist and Bayesian Poisson regression. The next steps in the development of our statistical procedure for exponential growth detection will be to assess the performance of various options in the face of the complications we'll see in real data. We will be faced with choices of methods, and of hyperparameters for a given method. To make these choices, we'll need to define the metrics by which we want to measure performance. Some possibilities include:

1. Computational cost: how expensive is it to perform the inference procedure on a realistic dataset? Are there cheap-but-coarse screens that we can apply to cut down the number of candidates for more expensive analysis?

2. Sensitivity: for a given set of conditions, resource expenditure, and hyperparameters, how likely are we to detect exponential growth before a certain threshold abundance (or other criterion) is reached?

3. False discovery rate: for a given set of conditions, expenditure, and hyperparameters, how many false-positives do we expect to get? There will be a tradeoff between sensitivity and FDR, which can be measured by, e.g., a precision-recall curve.

4. Predictive accuracy: given a fit to the data we have, how accurately does the model predict future count data? This could be used for cross-validation of hyperparameters.

5. Robustness: how well does the model perform (according to the above considerations) when the model is mis-specified? What sorts of mis-specifications are we most concerned about?

Combining these considerations will require decision theory. For example, in a Bayesian context, we could define a loss function that takes into account the costs of different types of errors, and choose hyperparameters that minimize the expected loss.