

Article

## Pushing for the Extreme: Estimation of Poisson Distribution from Low Count Unreplicated Data—How Close Can We Get?

Peter Tiňo

School of Computer Science, The University of Birmingham, Birmingham, B15 2TT, UK;  
E-Mail: P.Tino@cs.bham.ac.uk

Received: 15 January 2013; in revised form: 21 March 2013 / Accepted: 25 March 2013 /  
Published: 8 April 2013

---

**Abstract:** Studies of learning algorithms typically concentrate on situations where potentially ever growing training sample is available. Yet, there can be situations (e.g., detection of differentially expressed genes on unreplicated data or estimation of time delay in non-stationary gravitationally lensed photon streams) where only extremely small samples can be used in order to perform an inference. On unreplicated data, the inference has to be performed on the smallest sample possible—sample of size 1. We study whether anything useful can be learnt in such extreme situations by concentrating on a Bayesian approach that can account for possible prior information on expected counts. We perform a detailed information theoretic study of such Bayesian estimation and quantify the effect of Bayesian averaging on its first two moments. Finally, to analyze potential benefits of the Bayesian approach, we also consider Maximum Likelihood (ML) estimation as a baseline approach. We show both theoretically and empirically that the Bayesian model averaging can be potentially beneficial.

**Keywords:** Poisson distribution; unreplicated data; Bayesian learning; expected Kullback–Leibler divergence

---

### 1. Introduction

Studies in (computational) learning theory mostly tend to concentrate on situations where potentially ever increasing number of training examples is available. While such results can lead to deep insights into the workings of learning algorithms, e.g., linking together characteristics of the data generating distributions, learning machines and sample sizes, there can be situations where, by very nature of the

problem, only extremely small samples are available. In such situations it is of utmost importance to theoretically analyze exactly what and under what circumstances can be learnt. One example of such a scenario in count data is detection of differentially expressed genes, where even subtle changes in gene expression levels can be indicators of biologically crucial processes [1]. When replicas are costly to obtain one can attempt to use the limited data at one's disposal to make the relevant inferences, as for example in the Audic and Claverie approach [2–6]. Another situation where available count data can be extremely sparse is estimation of time delay in non-stationary gravitationally lensed photon streams. When the scale of variability of the source is of order, say, of more than tens of days and observation gaps are not too long, one can resolve the time delay between lensed images of the same source by working directly with daily measurements of fluxes in the radio, optical or X-ray range [7–10]. However, when the variability scale is of the order of hours, one must turn to photon streams in the lensed images. One possibility of time delay detection in such cases is through comparing counts in relatively short and time-shifted moving time windows in the lensed photon streams.

In this paper we theoretically study what happens in the extreme situation of unreplicated data when the inference has to be performed on the smallest sample possible—sample of size 1. We consider a model-based Bayesian approach that averages over possible Poisson models with weighting determined by the posterior over the models, given the single observation. In fact, such a Bayesian approach has been considered in the bioinformatics literature under the assumption of flat improper prior over the Poisson rate parameter [2–6]. One can, of course, be excused for being highly sceptical about the relevance of such inferences, yet the methodology has apparently been used in a number of successful studies. In an attempt to build theoretical foundations behind such inference schemes, we proved a rather surprising result [11]: The expected Kullback–Leibler divergence from the true unknown Poisson distribution to its *model learnt from a single realization* never exceeds 1/2 bit.

Even though the field of bioinformatics is moving fast and better procedures for detection of differentially expressed genes have been introduced (e.g., not relying on the Poisson assumption, specifically taking into account potential dependencies among the genes, *etc.*), the primary focus of this study is different. Irrespective of the application domain, we theoretically investigate how reliably can a model for count data be build from a single count observation, under the assumption of a Poisson source. There are two issues that need careful consideration:

1. Equal a-priori weighting (flat prior) over possible (unknown) Poisson sources is unrealistic. Typical values of observed counts are usually bounded by the nature of the problem (e.g., gene magnification setting used in the experiments or time window on the photon streams). One may have a good initial (a-priori) guess as to what ranges of typical observed counts might be reasonably expected. In particular, we are interested in the low count regimes. In such cases, it is desirable to incorporate such prior knowledge into the inference mechanism. In this study, we do this in the Bayesian framework through prior distribution over the expected counts.
2. To understand potential benefits of the proposed learning/inference method (in our case Bayesian approach), it is important to compare it with a simple straightforward baseline (here maximum likelihood estimation). We contrast the expected Kullback–Leibler divergences from the true unknown Poisson distribution to its Bayesian and maximum likelihood estimates, inferred from a single realization.

The paper has the following organization. In Section 2 we introduce the maximum likelihood and Bayesian (with flat prior over mean rates) approaches to inferring predictive distribution over counts based on a single count observation. We also briefly review past work on information theoretic properties of the two approaches. Section 3 contains derivation of a more general Bayesian approach with gamma prior on the mean count parameter. In Section 4 we calculate the first two central moments of our generalized model. This enables us to better understand the influence of the prior on the inferred model and highlight the differences with the previous approach using the flat (improper) prior. In Section 5 we perform an information theoretic study of learning capabilities of the generalized model. Empirical investigations are presented in Section 6 and the main findings are discussed and summarized in Section 7.

## 2. Single Count Data—Bayesian and Maximum Likelihood Approaches

In this section we will briefly review the original Audic–Claverie [2] and maximum likelihood approaches outside the bioinformatics context.

### 2.1. Bayesian Averaging in the Audic–Claverie Approach

Let  $x$  be an observed count in an experiment. When repeating the experiment, possibly under different conditions, we observe a (possibly different) count  $y$ . The quantity of interest is the probability of observing  $y$  given that we already observed  $x$ , not knowing the identity of the generating Poisson source

$$P(X = x|\lambda) = e^{-\lambda} \frac{\lambda^x}{x!} \tag{1}$$

where  $\lambda \geq 0$  is the (unknown) parameter representing the mean count value.

Under the null hypothesis (not differentially expressed genes), both counts  $x$  and  $y$  come from the same underlying Poisson distribution  $P(\cdot|\lambda)$ . The key instrument in the Audic–Claverie approach is a distribution  $P_{AC}(y|x)$  over counts  $y$  informed by the observed count  $x$ , under the null hypothesis.  $P_{AC}(y|x)$  is obtained by Bayesian averaging (infinite mixture) of all possible Poisson distributions  $P(y|\lambda')$  with mixing proportions equal to the posteriors  $p(\lambda'|x)$  under the flat prior over  $\lambda$ . Formally, the probability of count  $y$ , given the observed count  $x$  from the same (unknown) Poisson distribution is:

$$\begin{aligned} P(y|x) &= \int_0^\infty p(y, \lambda|x) \, d\lambda \\ &= \int_0^\infty P(y|\lambda, x) p(\lambda|x) \, d\lambda \\ &= \int_0^\infty P(y|\lambda) \frac{P(x|\lambda) p(\lambda)}{\int_0^\infty P(x|\lambda') p(\lambda') \, d\lambda'} \, d\lambda \end{aligned} \tag{2}$$

Imposing the flat (improper) prior  $p(\lambda)$  over the Poisson parameter  $\lambda$  results in

$$P_{AC}(y|x) = \frac{1}{y!} \frac{\int_0^\infty e^{-2\lambda} \lambda^{x+y} \, d\lambda}{\int_0^\infty e^{-\lambda} \lambda^x \, d\lambda}$$

Since Gamma distribution parameterized by  $a, b > 0$  takes the form

$$Gamma(\lambda|a, b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} e^{-b\lambda}$$

where  $\Gamma(a) = \int_0^\infty u^{a-1}e^{-u}du$  is the Gamma function, we have

$$P_{AC}(y|x) = \frac{1}{y!} \frac{\Gamma(x+y+1)}{2^{x+y+1} \Gamma(x+1)} \tag{3}$$

which, since  $x$  and  $y$  are integers (i.e.,  $\Gamma(x) = (x-1)!$ ), can be rewritten as

$$P_{AC}(y|x) = \frac{1}{2^{x+y+1}} \frac{(x+y)!}{x! y!} \tag{4}$$

$$= \frac{1}{2^{x+y+1}} \binom{x+y}{x} \tag{5}$$

$P_{AC}(\cdot|x)$  can then be used, e.g., for principled inferences, construction of confidence intervals or statistical testing.

### 2.2. Information Theory of $P_{AC}(y|x)$

Consider a “true” underlying Poisson distribution  $P(y|\lambda)$  (1) over possible counts  $y \geq 0$ . We first use  $P(\cdot|\lambda)$  to generate a count  $x$  and then employ  $P_{AC}(y|x)$  (5) as a model distribution over  $y$ , given the already observed count  $x$ . We ask: If we repeated the process above, how different, in terms of Kullback–Leibler divergence, are on average the two distributions over  $y$ ? One would naturally hope that  $P_{AC}(y|x)$  is sufficiently representative of the true unknown distribution  $P(y|\lambda)$ .

In [11] we proved that, given an underlying Poisson distribution  $P(x|\lambda)$ , if we repeatedly generated a “representative” count  $x$  from  $P(x|\lambda)$ , the average divergence  $\mathcal{E}(\lambda)$  of  $P_{AC}(y|x)$  from the truth  $P(y|\lambda)$  would never exceed 1/2 bit.

**Theorem 1** [11] Consider an underlying Poisson distribution  $P(\cdot|\lambda)$  parameterized by some  $\lambda > 0$ . Then

$$\mathcal{E}(\lambda) = E_{P(x|\lambda)}[D_{KL}[P(y|\lambda)||P_{AC}(y|x)]] = \frac{1}{2} \log 2 + O\left(\frac{1}{\lambda}\right)$$

where  $D_{KL}[P(y|\lambda)||P_{AC}(y|x)]$  is the Kullback–Leibler divergence from  $P(y|\lambda)$  to  $P_{AC}(y|x)$ ,

$$D_{KL}[P(y|\lambda)||P_{AC}(y|x)] = \sum_{y=0}^\infty P(y|\lambda) \log \frac{P(y|\lambda)}{P_{AC}(y|x)}$$

The expected divergence (in bits) can be well-approximated (up to order  $O(\lambda^{-3})$ ) by [11]:

$$\mathcal{E}(\lambda) \approx \frac{1}{2} - \frac{1}{12\lambda} \left(1 - \frac{1}{2}\right) - \frac{1}{24\lambda^2} \left(1 - \frac{1}{2^2}\right) \tag{6}$$

### 2.3. $P_{AC}(y|x)$ vs. Maximum Likelihood

In this section we will briefly recall information theoretic analysis of the maximum likelihood estimate  $P_{ML}(y|x)$  in place of  $P_{AC}(y|x)$  [12]. First note that Poisson distribution  $P(y|\lambda)$  is only defined for positive  $\lambda$ . In the case of observing zero count  $x = 0$ , we cannot directly use the “maximum likelihood estimate”  $P(y|0)$ . One option for dealing with zero observed counts is to allow for some form of model

regularization, e.g., infer a Poisson model  $P(y|\epsilon)$ , for some small  $\epsilon > 0$ . In other words, if a count  $x \geq 1$  is observed, follow the standard maximum likelihood procedure and infer  $P_{ML}(y|x) = P(y|x)$  as the Poisson model; if a zero count is observed,  $x = 0$ , infer  $P_{ML}(y|0) = P(y|\epsilon)$  for some fixed  $\epsilon \in (0, 1]$ . This is the route taken in [12] and adopted in this paper. Only a minimum amount of necessary regularization due to zero observed counts is employed in the otherwise straightforward ML approach.

**Theorem 2** [12] Consider an underlying Poisson distribution  $P(\cdot|\lambda)$  parameterized by some  $\lambda > 0$  and a regularization constant  $\epsilon \in (0, 1]$ . The expected divergence in bits  $\Upsilon(\lambda, \epsilon)$  between the true Poisson source and its (regularized) maximum likelihood estimate based on a single observation,

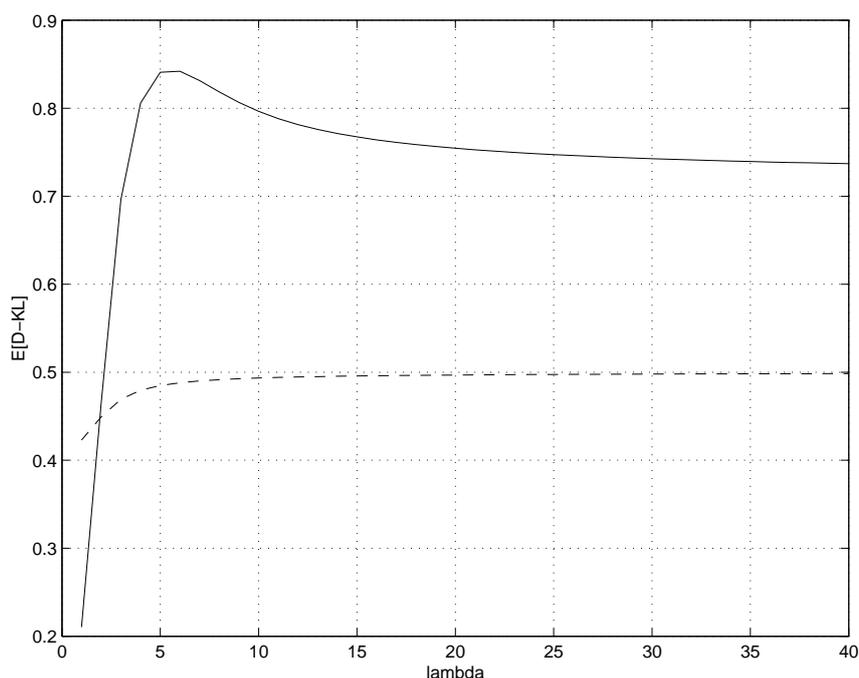
$$\Upsilon(\lambda, \epsilon) = E_{P(x|\lambda)}[ D_{KL}[P(y|\lambda)||P_{ML}(y|x)] ]$$

is equal to

$$\Upsilon(\lambda, \epsilon) = \lambda \left( \log_2 \lambda - \sum_{x=1}^{\infty} P(x|\lambda) \log_2 x \right) + e^{-\lambda} (\epsilon - \lambda \log_2 \epsilon) \tag{7}$$

Note that the expected divergence  $\Upsilon(\lambda, \epsilon)$  can get prohibitively large when regularizing with small  $\epsilon > 0$ . As an illustration, in Figure 1 we show expected divergence  $\Upsilon(\lambda, \epsilon = 1)$  of the ML estimation (zero count regularized with  $\epsilon = 1$ ) for a range of mean parameter values  $\lambda$  of the underlying Poisson source (solid line). Also shown is the expected divergence  $\mathcal{E}(\lambda)$  of  $P_{AC}(y|x)$  (dashed line). Except for very small Poisson source rates  $\lambda$ ,  $P_{AC}(y|x)$  is clearly benefitting from the stabilizing effect of Bayesian averaging, given the extremely small sample size.

**Figure 1.** Expected divergence (in bits)  $\Upsilon(\lambda, \epsilon = 1)$  of the ML estimation (zero count regularized with  $\epsilon = 1$ ) (solid line). Also shown is the expected divergence  $\mathcal{E}(\lambda)$  of  $P_{AC}(y|x)$  (dashed line).



### 3. Generalized $P_{AC}(y|x)$ with Gamma Prior

In this section we will generalize  $P_{AC}(y|x)$  through the use of (conjugate) gamma prior

$$P(\lambda|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

on the Poisson mean parameter  $\lambda$ . The positive parameters  $\alpha, \beta$  determine the overall shape of the prior. Given a single observation  $x$ , the posterior

$$P(\lambda|x, \alpha, \beta) = \frac{P(x|\lambda) P(\lambda|\alpha, \beta)}{\int_0^\infty P(x|\lambda) P(\lambda|\alpha, \beta) d\lambda}$$

is the gamma distribution with parameters  $\alpha + x$  and  $\beta + 1$ ,

$$P(\lambda|x, \alpha, \beta) = \frac{(\beta + 1)^{\alpha+x}}{\Gamma(\alpha + x)} \lambda^{\alpha+x-1} e^{-(\beta+1)\lambda}$$

The mean of  $P(\lambda|x, \alpha, \beta)$  is equal to  $(\alpha + x)/(\beta + 1)$ . A loose intuitive interpretation of the prior parameters  $\alpha, \beta$  (assuming they are integers) is that prior to seeing the current data (in our case only one observation (count)  $x$ ), we have seen  $\beta$  “observations”,  $x'_1, x'_2, \dots, x'_\beta$ , with the total cumulative count  $\alpha = x'_1 + x'_2 + \dots + x'_\beta$ . Hence the mean parameter estimate would shift from  $x$  (ML estimation corresponding to  $\alpha, \beta \rightarrow 0$ ) to  $(x'_1 + x'_2 + \dots + x'_\beta + x)/(\beta + 1)$ .

As in the case of  $P_{AC}(y|x)$ , having observed a count  $x$ , we build a predictive distribution over future counts  $y$  by integrating out the mean parameter  $\lambda$  with respect to the posterior  $P(\lambda|x, \alpha, \beta)$ ,

$$\begin{aligned} P_G(y|x, \alpha, \beta) &= \int_0^\infty P(y|\lambda) P(\lambda|x, \alpha, \beta) d\lambda \\ &= \frac{(\beta + 1)^{\alpha+x}}{\Gamma(\alpha + x)} \frac{1}{y!} \int_0^\infty \lambda^{\alpha+x+y-1} e^{-(\beta+2)\lambda} d\lambda \end{aligned} \tag{8}$$

From normalization of the gamma distribution we get

$$\int_0^\infty \lambda^{a-1} e^{-b\lambda} d\lambda = \frac{\Gamma(a)}{b^a}$$

and so

$$\int_0^\infty \lambda^{\alpha+x+y-1} e^{-(\beta+2)\lambda} d\lambda = \frac{\Gamma(x + y + \alpha)}{(\beta + 2)^{x+y+\alpha}}$$

leading to

$$P_G(y|x, \alpha, \beta) = \frac{1}{y!} \frac{\Gamma(x + y + \alpha)}{\Gamma(x + \alpha)} \frac{(\beta + 1)^{x+\alpha}}{(\beta + 2)^{x+y+\alpha}} \tag{9}$$

It can be easily verified that the original  $P_{AC}(y|x)$  is obtained as a special case of  $P_G(y|x, \alpha, \beta)$  when  $\alpha = 1$  and  $\beta \rightarrow 0$ . If Jeffrey’s prior were used instead of the flat prior in  $P_{AC}(y|x)$ , we would obtain  $P_G(y|x, \alpha, \beta)$  with  $\alpha = 1/2$  and  $\beta \rightarrow 0$  etc.

If  $\alpha$  is an integer, we have

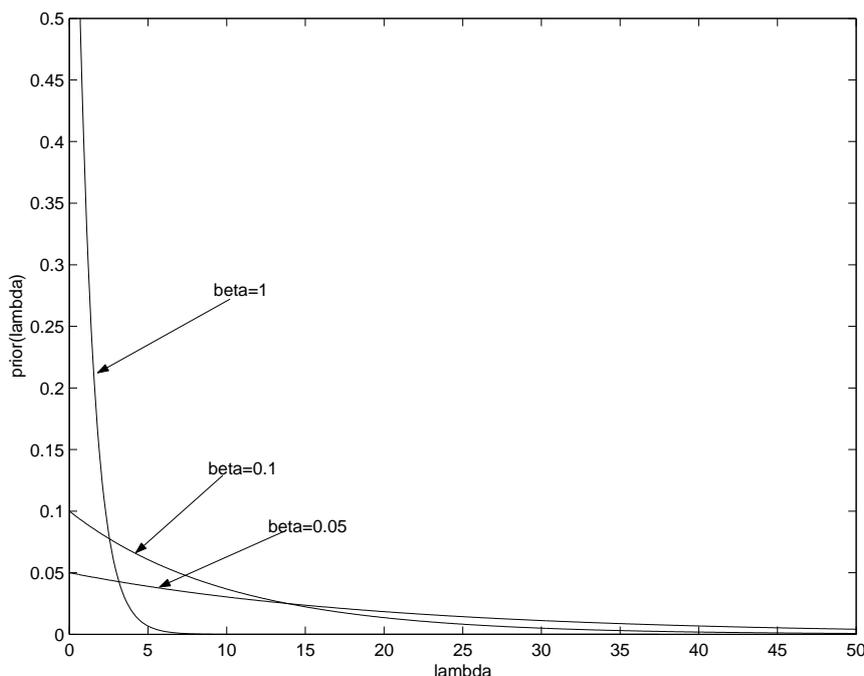
$$P_G(y|x, \alpha, \beta) = \left(\frac{1 + \beta}{2 + \beta}\right)^{x'+1} \left(\frac{1}{2 + \beta}\right)^y \binom{x' + y}{y} \tag{10}$$

where  $x' = x + \alpha - 1$  is the observed count including prior observations. This expression generalizes  $P_{AC}(y|x)$  (5),

$$P_{AC}(y|x) = \left(\frac{1}{2}\right)^{x+1} \left(\frac{1}{2}\right)^y \binom{x+y}{y}$$

While  $P_G(y|x, \alpha, \beta)$  (9) can be used with any appropriate setting of  $\alpha, \beta$  (e.g., given a prior knowledge of the range of counts one may reasonably expect), in this contribution we concentrate on using the gamma prior to mitigate for the unrealistic equal weighting of all  $\lambda > 0$  in the flat prior behind  $P_{AC}(y|x)$ . Indeed, the observed counts are typically bounded by the nature of the problem and one can represent this through setting  $\alpha = 1$  and varying  $\beta > 0$  in the gamma prior  $P(\lambda|\alpha, \beta)$  underlying  $P_G(y|x, \alpha, \beta)$ . Some examples of such priors are shown in Figure 2. Decreasing  $\beta$  leads to weaker emphasis on low  $\lambda$ , eventually recovering the flat (improper) prior for  $\beta = 0$ .

**Figure 2.** Gamma prior  $P(\lambda|\alpha = 1, \beta)$ . Shown are the priors for three possible values of parameter  $\beta, \beta \in \{1, 0.1, 0.05\}$ .



In Section 2.3 maximum likelihood estimation was regularized at zero count by imposing a non-zero “count”  $\epsilon$  instead of the observed zero one. The generalized form of  $P_{AC}(y|x), P_R(y|x, \beta) = P_G(y|x, \alpha = 1, \beta)$  can be also viewed as an alternative “soft” form of regularization of the maximum likelihood approach at zero counts.

Parameter  $\beta$  in the Gamma prior

$$P(\lambda|\alpha = 1, \beta) = \beta e^{-\beta\lambda}$$

can be set in a data driven manner, e.g., using the following strategy: Given the observed count  $x$ , we require that the area up to  $x + 1$  covered by the prior is equal to  $\theta$ , for some threshold  $\theta \in (0, 1)$  (e.g.,

$\theta = 1/4$ ). In other words,  $F(x + 1|\beta) = \theta$ , where  $F(\lambda|\beta) = 1 - e^{-\beta\lambda}$  is the cumulative distribution function of  $P(\lambda|\alpha = 1, \beta)$ . This leads to

$$\beta(x) = -\frac{\ln(1 - \theta)}{x + 1} \tag{11}$$

For zero observed count  $x = 0$ ,  $\beta(0) = -\ln(1 - \theta)$  and the prior gets more concentrated on smaller values of  $\lambda$  as likely candidates for the mean count of the underlying Poisson source. With increasing count values  $x > 0$  the parameter  $\beta(x)$  decreases to 0 and the prior gradually approaches the flat prior of  $P_{AC}(y|x)$ .

Finally, we contrast  $P_G(y|x, \alpha, \beta)$  with the negative binomial distribution

$$P_{NB}(y|r, q) = \frac{1}{y!} \frac{\Gamma(r + y)}{\Gamma(r)} q^r (1 - q)^y \tag{12}$$

with parameters  $r > 0$  and  $q \in [0, 1]$ . One interpretation of the negative binomial distribution  $P_{NB}(y|r, q)$  is that it corresponds to a Gamma–Poisson mixture that one obtains by imposing a Gamma prior  $P(\lambda|r, (1 - q)/q)$  on the mean count parameter  $\lambda$  of the Poisson distribution  $P(y|\lambda)$  and integrating out  $\lambda$ . In our context it is natural to identify  $r$  and  $(1 - q)/q$  with hyperparameters  $\alpha$  and  $\beta$  used in  $P_G(y|x, \alpha, \beta)$ . It follows that  $q = (\beta + 1)^{-1}$ . Hence, we rewrite (12) as

$$P_{NB}(y|\alpha, (\beta + 1)^{-1}) = \frac{1}{y!} \frac{\Gamma(\alpha + y)}{\Gamma(\alpha)} \frac{\beta^y}{(\beta + 1)^{\alpha+y}} \tag{13}$$

Direct comparison of (13) with (9) leads to an intuitive insight: The  $\beta$  prior measurements of total count  $\alpha$  introduced by the gamma prior  $P(\lambda|\alpha, \beta)$  are in the case of  $P_G(y|x, \alpha, \beta)$  extended with a single observation  $x$ , resulting in  $\beta + 1$  observations of total count  $\alpha + x$ . This can be represented by

$$P_{NB}(y|\alpha + x, (\beta + 2)^{-1}) = \frac{1}{y!} \frac{\Gamma(x + y + \alpha)}{\Gamma(x + \alpha)} \frac{(\beta + 1)^y}{(\beta + 2)^{x+y+\alpha}} \tag{14}$$

It follows that

$$\frac{P_G(y|x, \alpha, \beta)}{P_{NB}(y|\alpha + x, (\beta + 2)^{-1})} = (\beta + 1)^{x+\alpha-y}.$$

Bayesian averaging in  $P_G(y|x, \alpha, \beta)$  with respect to the posterior over  $\lambda$ , given a count  $x$ , differs from the corresponding negative binomial distribution  $P_{NB}(y|\alpha + x, (\beta + 2)^{-1})$  by the factor  $(\beta + 1)^{x+\alpha-y}$  that depends on the difference between the prior+observed count  $\alpha + x$  and  $y$ .

#### 4. First and Second Moments of the Generalized $P_{AC}(y|x)$

In [11] we showed that  $P_{AC}(y|x)$  and the underlying Poisson distribution are quite similar in their nature: for any (integer) mean rate  $\lambda \geq 1$ , the Poisson distribution  $P(\cdot|\lambda)$  has two neighboring modes located at  $\lambda$  and  $\lambda - 1$ , with  $P(\lambda|\lambda) = P(\lambda - 1|\lambda)$ . Analogously, given a count  $x \geq 1$ ,  $P_{AC}(\cdot|x)$  has two neighboring modes, one located at  $x$ , the other at  $x - 1$ , with  $P_{AC}(x|x) = P_{AC}(x - 1|x)$ . As in Poisson distribution, the values of  $P_{AC}(y|x)$  decrease as one moves away from the modes in both directions. In this section we derive the first two moments of the generalized  $P_{AC}(y|x)$ ,  $P_G(y|x, \alpha, \beta)$ . As a special case, we will show that as a result of Bayesian averaging, the variance of  $P_{AC}(y|x)$  is double that of the underlying (unobserved) Poisson distribution.

**Theorem 3** Consider a non-negative integer  $x$  and the associated generalized model  $P_G(y|x, \alpha, \beta)$ . Then,

$$E_{P_G(y|x, \alpha, \beta)}[y] = \frac{x + \alpha}{\beta + 1}, \quad Var[y] = \frac{\beta + 2}{\beta + 1} E_{P_G(y|x, \alpha, \beta)}[y]$$

Proof: Let us evaluate

$$\begin{aligned} E_{P_G(y|x, \alpha, \beta)}[y] &= \sum_{y=0}^{\infty} \frac{\Gamma(x + y + \alpha)}{\Gamma(x + \alpha)} \frac{(\beta + 1)^{x+\alpha}}{(\beta + 2)^{x+y+\alpha}} \frac{1}{y!} y \\ &= \sum_{y=1}^{\infty} \frac{\Gamma(x + y + \alpha)}{\Gamma(x + \alpha)} \frac{(\beta + 1)^{x+\alpha}}{(\beta + 2)^{x+y+\alpha}} \frac{1}{(y - 1)!} \\ &= \sum_{y'=0}^{\infty} \frac{\Gamma(x + y' + 1 + \alpha)}{\Gamma(x + \alpha)} \frac{(\beta + 1)^{x+\alpha}}{(\beta + 2)^{x+y'+1+\alpha}} \frac{1}{y'!} \\ &= \sum_{y'=0}^{\infty} \frac{\Gamma(x + y' + \alpha) \cdot (x + y' + \alpha)}{\Gamma(x + \alpha)} \frac{(\beta + 1)^{x+\alpha}}{(\beta + 2) \cdot (\beta + 2)^{x+y'+\alpha}} \frac{1}{y'!} \end{aligned} \tag{15}$$

In the third equality we have used substitution  $y' = y - 1$  and the last equality follows from  $\Gamma(z + 1) = z \cdot \Gamma(z)$ . By (15),

$$E_{P_G(y|x, \alpha, \beta)}[y] = \sum_{y=0}^{\infty} P_G(y|x, \alpha, \beta) \frac{x + \alpha + y}{\beta + 2} \tag{16}$$

$$= \frac{x + \alpha}{\beta + 2} + \frac{1}{\beta + 2} E_{P_G(y|x, \alpha, \beta)}[y] \tag{17}$$

Solving (17) we obtain

$$E_{P_G(y|x, \alpha, \beta)}[y] = \frac{x + \alpha}{\beta + 1} \tag{18}$$

For the variance of  $P_G(y|x, \alpha, \beta)$  we have

$$Var_{P_G(y|x, \alpha, \beta)}[y] = E_{P_G(y|x, \alpha, \beta)}[y^2] - (E_{P_G(y|x, \alpha, \beta)}[y])^2 \tag{19}$$

Now,

$$\begin{aligned} E_{P_G(y|x, \alpha, \beta)}[y^2] &= \sum_{y=0}^{\infty} \frac{\Gamma(x + y + \alpha)}{\Gamma(x + \alpha)} \frac{(\beta + 1)^{x+\alpha}}{(\beta + 2)^{x+y+\alpha}} \frac{1}{y!} y^2 \\ &= \sum_{y=1}^{\infty} \frac{\Gamma(x + y + \alpha)}{\Gamma(x + \alpha)} \frac{(\beta + 1)^{x+\alpha}}{(\beta + 2)^{x+y+\alpha}} \frac{1}{y!} y^2 \\ &= \sum_{y=1}^{\infty} \frac{\Gamma(x + y + \alpha)}{\Gamma(x + \alpha)} \frac{(\beta + 1)^{x+\alpha}}{(\beta + 2)^{x+y+\alpha}} \frac{1}{(y - 1)!} y \\ &= \sum_{y'=0}^{\infty} \frac{\Gamma(x + y' + 1 + \alpha)}{\Gamma(x + \alpha)} \frac{(\beta + 1)^{x+\alpha}}{(\beta + 2)^{x+y'+1+\alpha}} \frac{1}{y'!} (y' + 1) \end{aligned}$$

$$\begin{aligned}
 &= \sum_{y'=0}^{\infty} \frac{(x + y' + \alpha) \Gamma(x + y' + \alpha)}{\Gamma(x + \alpha)} \frac{(\beta + 1)^{x+\alpha}}{(\beta + 2) (\beta + 2)^{x+y'+\alpha}} \frac{1}{y'!} (y' + 1) \\
 &= \frac{1}{\beta + 2} \sum_{y'=0}^{\infty} P_G(y'|x, \alpha, \beta) [(x + y' + \alpha) (y' + 1)] \\
 &= \frac{1}{\beta + 2} \sum_{y'=0}^{\infty} P_G(y'|x, \alpha, \beta) [x + y' + \alpha] \\
 &+ \frac{1}{\beta + 2} \sum_{y'=0}^{\infty} P_G(y'|x, \alpha, \beta) [y' (x + \alpha) + y'^2] \tag{20}
 \end{aligned}$$

Using (16), (18) and (20), we obtain

$$\begin{aligned}
 E_{P_G(y|x,\alpha,\beta)}[y^2] &= E_{P_G(y|x,\alpha,\beta)}[y] + \frac{x + \alpha}{\beta + 2} E_{P_G(y|x,\alpha,\beta)}[y] + \frac{1}{\beta + 2} E_{P_G(y|x,\alpha,\beta)}[y^2] \\
 &= \frac{x + \alpha}{\beta + 1} \left( 1 + \frac{x + \alpha}{\beta + 2} \right) + \frac{1}{\beta + 2} E_{P_G(y|x,\alpha,\beta)}[y^2] \tag{21}
 \end{aligned}$$

which can be solved as

$$E_{P_G(y|x,\alpha,\beta)}[y^2] = \frac{(x + \alpha) (x + \alpha + \beta + 2)}{(\beta + 1)^2} \tag{22}$$

Plugging (22) into (19) we obtain

$$Var_{P_G(y|x,\alpha,\beta)}[y] = \frac{(x + \alpha) (\beta + 2)}{(\beta + 1)^2} = \frac{\beta + 2}{\beta + 1} E_{P_G(y|x,\alpha,\beta)}[y]$$

□

Given an observation  $x$ , the maximum likelihood estimate of the underlying Poisson distribution is the Poisson distribution with mean  $x$ ,

$$P(y|x) = e^{-x} \frac{x^y}{y!}$$

After observing  $x$ , the mean of the maximum likelihood and  $P_{AC}(\cdot|x)$  estimates is  $x$  and  $x + 1$ , respectively. Hence, Bayesian averaging in  $P_{AC}(\cdot|x)$  induced by the flat improper prior over the mean rate  $\lambda$  results in increased expected value  $x + 1$  of the next count from the same underlying source, given that the current count  $x$ . However, a much more marked consequence of using the flat prior can be seen in the variance of  $P_{AC}(\cdot|x)$ : while variance of the maximum likelihood is  $x$ , it is  $2(x + 1)$  in  $P_{AC}(\cdot|x)$ .

Theorem 3 illustrates the role of more concentrated prior over  $\lambda$  on the generalized model. The mean expected count, after seeing  $x$ , is equal to the mean of the posterior  $P(\lambda|x, \alpha, \beta)$  over  $\lambda$ , namely  $(\alpha + x)/(\beta + 1)$ . As explained earlier, observed single count  $x$  with prior  $\beta$  counts of cumulative value  $\alpha$  results in  $\beta + 1$  counts of cumulative value  $\alpha + x$ . Hence the mean count per observation is  $(\alpha + x)/(\beta + 1)$ . As with Poisson distribution, the variance of the generalized model is closely related to its mean and approaches the mean with increasing number of prior counts  $\beta$ .

As for the soft regularization  $P_R(y|x, \beta) = P_G(y|x, \alpha = 1, \beta)$ , its mean is, as expected, biased towards values smaller than the observed count  $x$ , provided  $\beta > 1/x$ . Increased values of  $\beta$  result in smaller variance of  $P_R(y|x, \beta)$ . But how do such prior parameter modifications manifest themselves

in terms of accuracy of estimation of the underlying source? This question is investigated in the next section.

**5. Expected Divergence of the Generalized  $P_{AC}(y|x)$  from the True Underlying Poisson Distribution**

Consider an underlying Poisson source  $P(x|\lambda)$  generating counts  $x$ . In this section we would like to quantify the average divergence

$$\mathcal{E}_G(\lambda; \beta) = E_{P(x|\lambda)}[ D_{KL}[P(y|\lambda)||P_R(y|x, \beta)] ] \tag{23}$$

of the corresponding generalized  $P_{AC}(y|x)$ ,  $P_R(y|x, \beta) = P_G(y|x, \alpha = 1, \beta)$  (“softly” regularized ML), from the truth  $P(y|\lambda)$ , if we repeatedly generated a “representative” count  $x$  from  $P(x|\lambda)$ . The same question was considered in the context of maximum likelihood estimation in Section 2.3. In particular, we are interested in specifying under what circumstances is the generalized form of  $P_{AC}(y|x)$ ,  $P_R(y|x, \beta) = P_G(y|x, \alpha = 1, \beta)$ , preferable to the original  $P_{AC}(y|x) = P_G(y|x, \alpha = 1, \beta \rightarrow 0)$  and how it fares with the maximum likelihood estimation  $P_{ML}(y|x)$  of Section 2.3.

**Theorem 4** Consider an underlying Poisson distribution  $P(\cdot|\lambda)$  parameterized by some  $\lambda > 0$ . Then for  $\beta \geq 0$ ,

$$\mathcal{E}_G(\lambda; \beta) = \log_2 \left( \frac{\beta + 2}{\beta + 1} \right) - \frac{1}{2} + \lambda \left[ 2 \log_2 \left( \frac{\beta + 2}{2} \right) - \log_2(\beta + 1) \right] + O(\lambda^{-1}) \tag{24}$$

A higher order approximation (up to order  $\lambda^3$ ) reads:

$$\begin{aligned} \mathcal{E}_G(\lambda; \beta) &= \log_2 \left( \frac{\beta + 2}{\beta + 1} \right) - \frac{1}{2} + \lambda \left[ 2 \log_2 \left( \frac{\beta + 2}{2} \right) - \log_2(\beta + 1) \right] \\ &\quad - \frac{1}{12\lambda} \left( 1 - \frac{1}{2} \right) - \frac{1}{24\lambda^2} \left( 1 - \frac{1}{2^2} \right) - \frac{19}{360\lambda^3} \left( 1 - \frac{1}{2^3} \right) + O(\lambda^{-4}) \end{aligned} \tag{25}$$

Proof: Let us first express the divergence  $D_\beta(\lambda, x) = D_{KL}[P(y|\lambda)||P_R(y|x, \beta)]$ . We have

$$D_\beta(\lambda, x) = -H[P(y|\lambda)] - E_{P(y|\lambda)}[\log P_R(y|x, \beta)]$$

where  $H[P(y|\lambda)] = -E_{P(y|\lambda)}[\log P(y|\lambda)]$  is the entropy of the source  $P(y|\lambda)$  and

$$\begin{aligned} E_{P(y|\lambda)}[\log P_R(y|x, \beta)] &= -\log x! \\ &\quad - E_{P(y|\lambda)}[y] \log(\beta + 2) - (x + 1) \log \left( \frac{\beta + 2}{\beta + 1} \right) \\ &\quad - E_{P(y|\lambda)}[\log y!] + E_{P(y|\lambda)}[\log(x + y)!] \end{aligned}$$

Denoting (for integer  $d \geq 0$ )  $E_{P(y|\lambda)}[\log(y + d)!]$  by  $F(\lambda, d)$ , we write

$$\begin{aligned} D_\beta(\lambda, x) &= -H[P(y|\lambda)] + \log x! \\ &\quad + \lambda \log(\beta + 2) + (x + 1) \log \left( \frac{\beta + 2}{\beta + 1} \right) \\ &\quad + F(\lambda, 0) - F(\lambda, x) \end{aligned}$$

We are now ready to calculate the expectation  $\mathcal{E}_G(\lambda; \beta) = E_{P(x|\lambda)}[D_\beta(\lambda, x)]$ .

$$\begin{aligned} \mathcal{E}_G(\lambda; \beta) &= -H[P(y|\lambda)] + F(\lambda, 0) \\ &\quad + \lambda \log(\beta + 2) + (\lambda + 1) \log\left(\frac{\beta + 2}{\beta + 1}\right) \\ &\quad + F(\lambda, 0) - E_{P(x|\lambda)}[F(\lambda, x)] \end{aligned}$$

We have proved in [11] that  $E_{P(x|\lambda)}[F(\lambda, x)] = F(2\lambda, 0)$ , and so

$$\begin{aligned} \mathcal{E}_G(\lambda; \beta) &= -H[P(y|\lambda)] + \log\left(\frac{\beta + 2}{\beta + 1}\right) \\ &\quad + \lambda \log\left(\frac{(\beta + 2)^2}{\beta + 1}\right) \\ &\quad + 2F(\lambda, 0) - F(2\lambda, 0) \end{aligned}$$

Since

$$\begin{aligned} -H[P(y|\lambda)] &= E_{P(y|\lambda)}[\log P(y|\lambda)] \\ &= -\lambda \log e + E_{P(y|\lambda)}[y] \log \lambda - E_{P(y|\lambda)}[\log y!] \\ &= -\lambda \log e + \lambda \log \lambda - F(\lambda, 0) \end{aligned} \tag{26}$$

we have

$$\begin{aligned} \mathcal{E}_G(\lambda; \beta) &= \log\left(\frac{\beta + 2}{\beta + 1}\right) \\ &\quad + \lambda \left[ \log \lambda + \log\left(\frac{(\beta + 2)^2}{\beta + 1}\right) - \log e \right] \\ &\quad + F(\lambda, 0) - F(2\lambda, 0) \end{aligned} \tag{27}$$

Using entropy approximation (see [11]), one obtains

$$F(\lambda, 0) = \lambda(\log \lambda - \log e) + \frac{1}{2} \log(2\pi e \lambda) + O(\lambda^{-1})$$

leading to (in log base 2)

$$F(\lambda, 0) - F(2\lambda, 0) = -\frac{1}{2} + \lambda(\log_2 e - \log_2 \lambda - 2) + O(\lambda^{-1})$$

Finally,

$$\begin{aligned} \mathcal{E}_G(\lambda; \beta) &= \log_2\left(\frac{\beta + 2}{\beta + 1}\right) - \frac{1}{2} \\ &\quad + \lambda \left[ \log_2\left(\frac{(\beta + 2)^2}{\beta + 1}\right) - 2 \right] + O(\lambda^{-1}) \end{aligned}$$

which is equivalent to (24).

The higher order expression (25) is simply obtained by using higher order approximation to  $F(\lambda, 0) - F(2\lambda, 0)$ . □

Note that for  $\beta \rightarrow 0$  we recover our original result [11] that the expected divergence  $\mathcal{E}(\lambda)$  of the original  $P_{AC}(y|x)$  from the “truth”  $P(y|\lambda)$  is (up to terms of order  $\lambda^{-1}$ ) never greater than 1/2 bit. The soft regularization in  $P_R(y|x, \beta)$  (using prior  $P(\lambda|\alpha = 1, \beta)$  with  $\beta > 0$ ) can result in larger expected divergence from the underlying source than is the case for  $P_{AC}(y|x)$  (using improper flat prior over  $\lambda$ ). Moreover, (unlike in  $P_{AC}(y|x)$ ) such a regularization causes linear divergence of  $\mathcal{E}_G(\lambda; \beta)$  for large  $\lambda$ . The next theorem specifies for which underlying Poisson sources the soft regularization approach of  $P_R(y|x, \beta)$  is preferable to the original  $P_{AC}(y|x)$ .

**Theorem 5** For Poisson sources with mean rates

$$\lambda < \kappa(\beta) = \frac{\log\left(1 + \frac{\beta}{\beta+2}\right)}{\log\left(1 + \frac{\beta^2}{4(\beta+1)}\right)} \tag{28}$$

it holds  $\mathcal{E}(\lambda) > \mathcal{E}_G(\lambda; \beta)$  and hence  $P_R(y|x, \beta)$  is on average guaranteed to approximate (in the Kullback–Leibler divergence sense) the underlying source better than the original  $P_{AC}(y|x)$ .

Proof: It was shown in [11] that for the original  $P_{AC}(y|x)$ ,

$$\begin{aligned} \mathcal{E}(\lambda) &= \lambda(\log \lambda - \log e + 2 \log 2) + \log 2 \\ &\quad + F(\lambda, 0) - F(2\lambda, 0). \end{aligned} \tag{29}$$

From (27) and (29) we have that the difference between the expected divergences of the original and generalized forms of  $P_{AC}(y|x)$  is

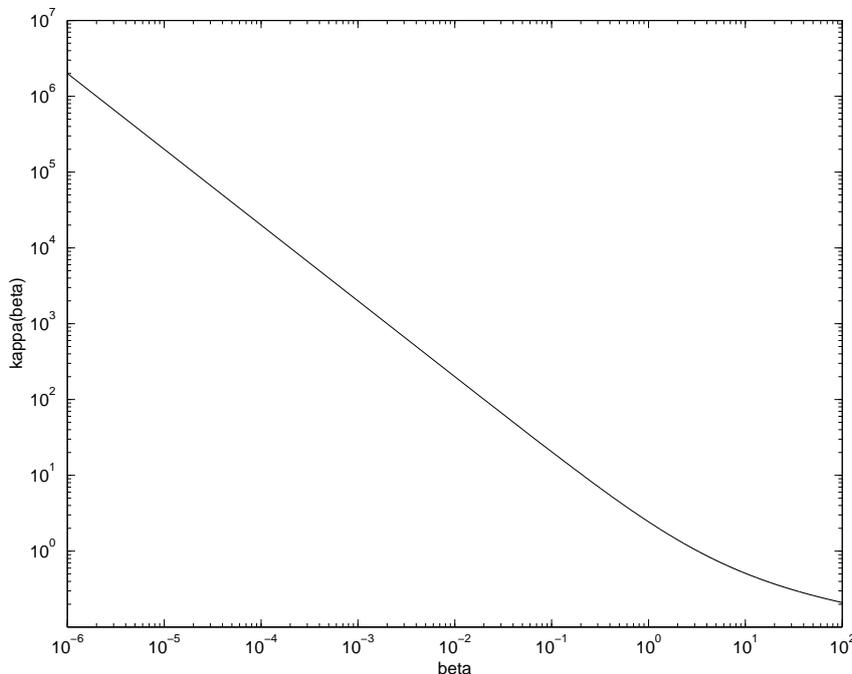
$$\begin{aligned} \mathcal{E}(\lambda) - \mathcal{E}_G(\lambda; \beta) &= \log 2 - \log\left(\frac{\beta + 2}{\beta + 1}\right) \\ &\quad + \lambda \left[ 2 \log 2 - \log\left(\frac{(\beta + 2)^2}{\beta + 1}\right) \right] \\ &= \log \frac{2(\beta + 1)}{\beta + 2} \\ &\quad + \lambda \log \frac{4(\beta + 1)}{(\beta + 2)^2} \end{aligned} \tag{30}$$

The result follows from solving for  $\mathcal{E}(\lambda) > \mathcal{E}_G(\lambda; \beta)$ . □

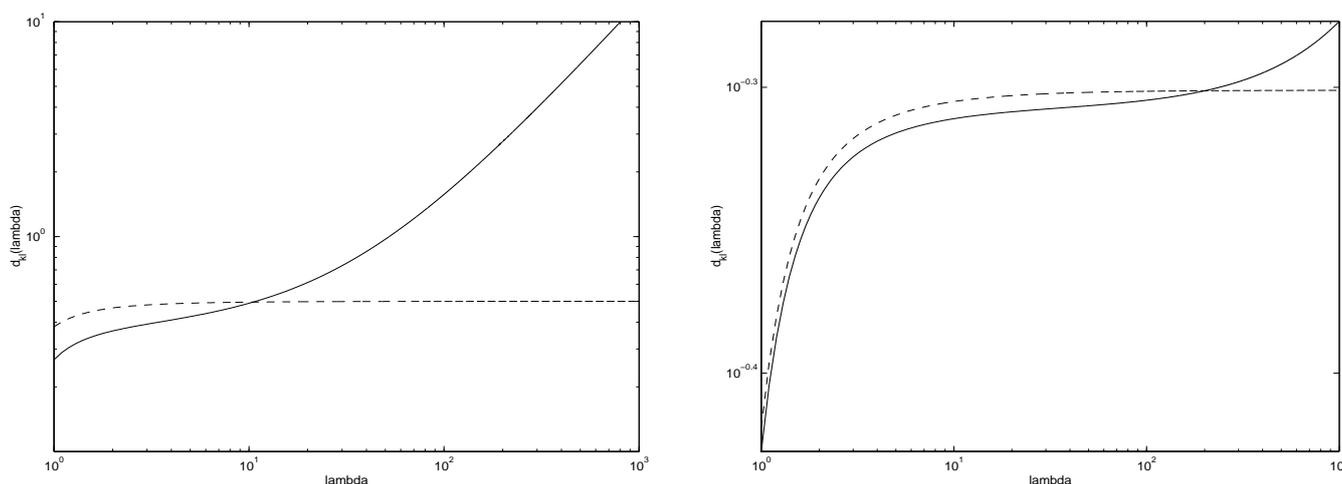
The graph (in log-log scale) of  $\kappa(\beta)$  is shown in Figure 3. An alternative way of data-driven setting of parameter  $\beta$  is suggested by the fact that  $\kappa(\beta)$  is lower bounded by  $\beta^{-1}$ . If the experimental setting is such that most counts are expected not to exceed some  $x_{max}$ ,  $\beta$  can be set to  $\beta = 1/x_{max}$ , so that  $P_R(y|x, \beta)$  is preferable to  $P_{AC}(y|x)$ .

In Figure 4 we present the expected divergences  $\mathcal{E}_G(\lambda; \beta)$  (solid line) and  $\mathcal{E}(\lambda)$  (dashed line) for  $\beta = 0.2$  (left) and  $\beta = 0.01$  (right). As expected, for underlying sources with small mean counts  $\lambda$  the advantage of using the regularized form  $P_R(y|x, \beta)$  (as opposed to the original  $P_{AC}(y|x)$ ) is more pronounced. However, for larger  $\lambda$  there is a heavy price to be paid in terms of inaccurate modelling by  $P_R(y|x, \beta)$ .

**Figure 3.** Graph of  $\kappa(\beta)$ . For Poisson sources with mean rates  $\lambda < \kappa(\beta)$ ,  $\mathcal{E}(\lambda) > \mathcal{E}_G(\lambda; \beta)$  and hence  $P_R(y|x, \beta)$  is on average guaranteed to approximate the underlying source better than the original  $P_{AC}(y|x)$ .



**Figure 4.** Expected divergences  $\mathcal{E}_G(\lambda; \beta)$  (solid line) and  $\mathcal{E}(\lambda)$  (dashed line) for  $\beta = 0.2$  (left) and  $\beta = 0.01$  (right).



### 6. Empirical Investigations

To investigate potential value of the more sophisticated Bayesian approach in the original and the generalized Audic–Claverie frameworks (Sections 2.1 and 3, respectively) against the baseline of simple (regularized) maximum likelihood estimation (Section 2.3), we conducted a series of simple illustrative experiments. In the generalized Audic–Claverie framework developed in this study, we used the two schemes for setting the regularization parameter  $\beta$  suggested in Sections 3 and 5. In the regularized

maximum likelihood approach  $P_{ML}(y|x)$  we set  $\epsilon = 1$ . From Figure 1, it appears that the biggest difference between the expected divergences from the true underlying Poisson source  $P(x|\lambda)$  to the original  $P_{AC}(\cdot|x)$  and the maximum likelihood estimate occurs for small mean rates  $\lambda$  roughly around  $\lambda = 5$ . We therefore run the experiments with  $\lambda = 5$ .

For illustration purposes, we follow the data generation mechanism used in [13] to compare methods for distinguishing between differential expression of genes associated with two treatment regimes. We stress that in no way we suggest that our experiments have strong relevance for bioinformatics, nor do we claim that the framework of [13] is the best test bed for assessing differential gene expression detection algorithms. We use the framework of [13] merely to illustrate whether the sophistication of the Bayesian approach (as opposed to simple (regularized) maximum likelihood) can bring benefits in a practical situation with low-count data.

Gene counts are simulated across the two treatment groups  $T_1$  and  $T_2$ . The tests are assessed by comparing false positive and true positive rates. In each experiment 10,000 gene pair counts  $(x_{1,j}, x_{2,j})$ ,  $j = 1, 2, \dots, 10,000$ , were produced, counts  $x_{1,j}$  and  $x_{2,j}$  associated with regimes  $T_1$  and  $T_2$ , respectively. As specified above, the sampling rate for  $T_1$  was fixed at  $\lambda_1 = 5$  throughout the experiment. We varied the mean  $\log_2$  fold change (LFC) between  $T_1$  and  $T_2$  from  $-2$  to  $2$ . Each gene pair count  $(x_{1,j}, x_{2,j})$ ,  $j = 1, 2, \dots, 10,000$ , was obtained through a generative process specified in [13] and described in detail in Appendix A.

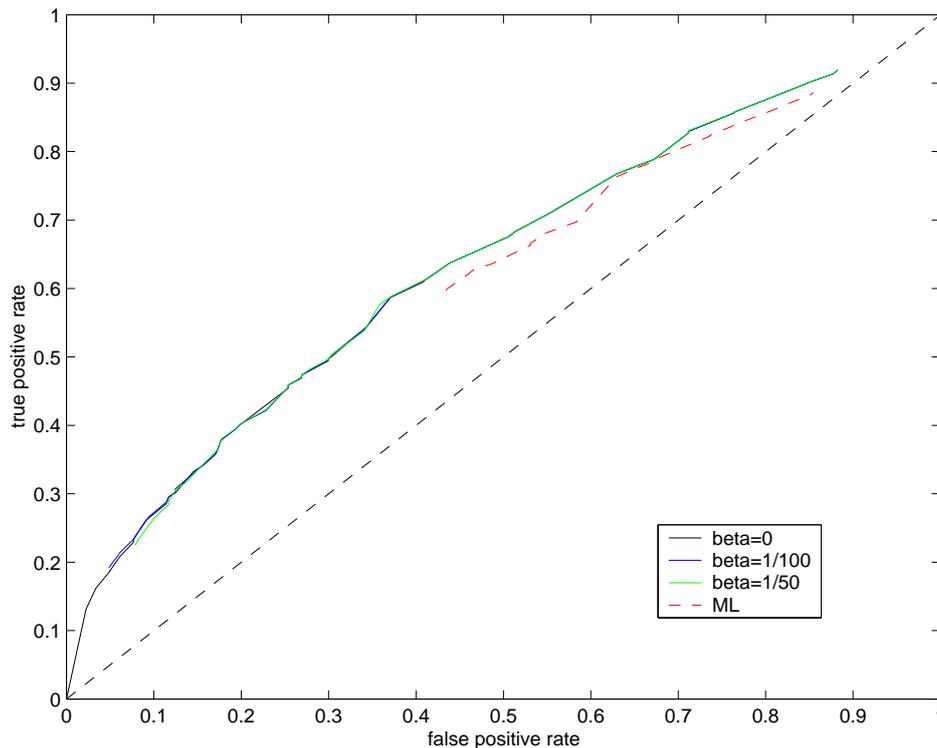
Having generated the gene pair counts, we used methods considered in this study to make a decision for each  $j = 1, 2, \dots, 10,000$ , whether the counts  $x_{1,j}, x_{2,j}$  originated from the same underlying source, *i.e.*, whether when generating  $y_{1,j}$  and  $y_{2,j}$ , the mean rates in the two regimes  $T_1$  and  $T_2$  were identical ( $LFC_j = 0$ ). Given the “test distribution”  $Q(y|x)$  and a confidence level  $\vartheta \in [0, 1]$ , we guess that  $x_{1,j}, x_{2,j}$  originated from the same source if the  $(1 - \vartheta)$ -quantile around the mean of  $Q(y|x_{1,j})$  contains  $x_{2,j}$  and vice-versa, *i.e.*, if the  $(1 - \vartheta)$ -quantile around the mean of  $Q(y|x_{2,j})$  contains  $x_{1,j}$ . In place of  $Q(y|x)$  we used  $P_{AC}(y|x)$ , its regularized form  $P_R(y|x, \beta)$  and the regularized maximum likelihood estimate  $P_{ML}(y|x)$  with  $\epsilon = 1$ .

For a given confidence level  $\vartheta \in [0, 1]$  and test statistic  $Q(y|x)$  we calculate the *false positive rate* (type I error rate) as the proportion of times a gene count pair  $(x_{1,j}, x_{2,j})$  was declared to have originated from two different underlying sources (differentially expressed gene) when in fact  $LFC_j$  was zero. The *true positive rate* (statistical power) was determined as the proportion of times a gene was correctly declared differentially expressed -  $(x_{1,j}, x_{2,j})$  declared to have originated come from two different underlying sources and  $LFC_j \neq 0$ .

Plot of false positive rate vs. true positive rate obtained for different values of  $\vartheta$  constitutes a *receiver operating characteristic* (ROC) curve. If the ROC curve for one test distribution is always above another, this suggests its superiority in classifying genes as differentially expressed. Trivial classification of genes as differentially expressed using a completely random guess would yield the identity (diagonal) ROC curve. ROC curves for the maximum likelihood method ( $\epsilon = 1$ , red dashed line) and the soft regularization model  $P_R(y|x, \beta)$ ,  $\beta = 1/50, 1/100$  (solid lines) are plotted in Figure 5. Not surprisingly, the Bayesian approach (solid lines) outperforms the penalized maximum likelihood one (red dashed line). However, the original  $P_{AC}(y|x)$  ( $\beta = 0$ , black line) and the soft regularization model (color solid lines) achieve almost identical performances. In this challenging setting (single observations at

low mean rate with additional noise), the scheme for setting the regularization parameter  $\beta$  suggested in Section 5 has little effect on the resulting classification performance. We also ran experiments to test the “dynamic” scheme for setting  $\beta$  introduced in Section 3, but no significant performance improvements were achieved.

**Figure 5.** ROC curves for test distributions  $P_{AC}(y|x) = P_R(y|x, \beta \rightarrow 0)$  (solid black line),  $P_R(y|x, \beta = 1/100)$  (solid blue line),  $P_R(y|x, \beta = 1/50)$  (solid green line) and  $P_{ML}(y|x)$  with  $\epsilon = 1$  (dashed red line). Mean rate of the underlying Poisson source was fixed at  $\lambda = 5$ .

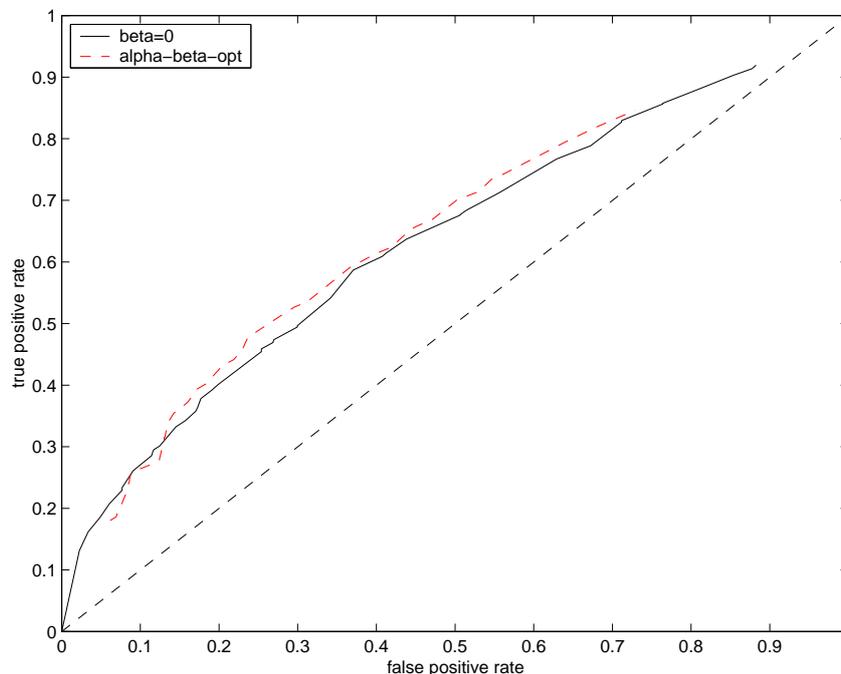


Finally, we devised yet another scheme for determining the hyper-parameters  $\alpha$  and  $\beta$  of the prior  $P(\lambda|\alpha, \beta)$  from the data. In the spirit of type II maximum likelihood, we find the most likely values of  $\alpha, \beta$ , given the observed counts  $\mathcal{C} = \{x_1, x_2, \dots, x_n\}$ , using  $P(\mathcal{C}|\alpha, \beta) = \prod_{i=1}^n P(x_i|\alpha, \beta)$ , where

$$P(x_i|\alpha, \beta) = \int_0^\infty P(x_i|\lambda) p(\lambda|\alpha, \beta) d\lambda \tag{31}$$

Using this method, we first optimize the prior hyperparameters on the observed data. The “optimized” prior  $P(x_i|\alpha_*, \beta_*)$  now reflects the possible ranges of mean counts  $\lambda$  one can expect given the data. We then repeated the experiments using the generalized model  $P_G(y|x, \alpha_*, \beta_*)$  derived from the optimized prior. In this way we can assess to what degree the relatively minor performance differences between the generalized and maximum likelihood models in Figure 5 are due to constraining  $\alpha$  to  $\alpha = 1$  (in  $P_R(y|x, \beta)$ ), or due to inherent difficulty of learning from single counts. The resulting ROC analysis is shown in Figure 6. The data driven setting of hyperparameters  $\alpha, \beta$  leads to slight improvement over  $P_{AC}(y|x)$  and  $P_R(y|x, \beta)$ .

**Figure 6.** ROC curves for test distributions  $P_{AC}(y|x) = P_R(y|x, \beta \rightarrow 0)$  (solid black line) and  $P_G(y|x, \alpha_*, \beta_*)$  (dashed red line). Mean rate of the underlying Poisson source was fixed at  $\lambda = 5$ .



### 7. Discussion and Conclusion

Studies of learning algorithms traditionally concentrate on situations where potentially ever increasing number of training examples is available. However, there are situations where only extremely small samples can be used in order to perform an inference. In this contribution we concentrated on extreme case of low count data governed by Poisson distribution, where only a single observation is available. We performed a rigorous theoretical investigation of the appropriateness of various model estimators, based on the single observation. We considered a Bayesian approach along the lines of [2], where the model built on the basis of a single observed count is no longer Poisson, even though we know that the generating source is Poisson (but do not know the mean rate).

We showed that the Bayesian approach is more optimal than the regularized maximum likelihood, in the sense that the expected Kullback–Leibler divergence from the source to the model is smaller for the Bayesian approach. Furthermore, we generalized the original model of [2] to account for possible prior information on expected expression counts. Detailed information theoretic study of learning capabilities of such a generalized model was conducted for the case of low count data. We also quantified the effect of Bayesian averaging on its first two moments.

We demonstrated both theoretically and empirically that the Bayesian model averaging on the generalized model can be potentially beneficial. For large  $\lambda$ , the expected divergence  $\Upsilon(\lambda, \epsilon)$  of the maximum likelihood estimator from the true Poisson source is dominated by the term

$$\lambda \left( \log \lambda - \sum_{x=1}^{\infty} P(x|\lambda) \log x \right)$$

since  $\lim_{\lambda \rightarrow \infty} e^{-\lambda} (\epsilon - \lambda \log \epsilon) = 0$ . We empirically determined that for  $\lambda \geq 10$ ,  $\Upsilon(\lambda, \epsilon = 1)$  expressed in bits is bounded by  $0.7 < \Upsilon(\lambda, \epsilon = 1) < 0.8$ . Hence, for mean Poisson rates  $\lambda \geq 10$ , the difference between the expected divergences of the Audic–Claverie and ML estimates from the true source is never less than 0.2 bits and never more than 0.3 bits. In other words,

$$0.2 < \Upsilon(\lambda, \epsilon = 1) - \mathcal{E}(\lambda) < 0.3, \quad \lambda \geq 10$$

**Acknowledgements**

This work was supported by a BBSRC grant (no. BB/H012508/1).

**Appendix A**

In the generative process of [13], each gene pair count  $(x_{1,j}, x_{2,j})$ ,  $j = 1, 2, \dots, 10,000$ , was obtained as follows:

1. The sampling rate  $\lambda_{2,j}$  for the treatment group  $T_2$  is obtained as

$$\begin{aligned} \lambda_{2,j} &= 2^{(\log_2 \lambda_1) - LFC_j} \\ LFC_j &\sim \text{Uniform}\{-2.0, -1.5, -1.0, \dots, 1.5, 2.0\} \end{aligned}$$

2. A pair of gene counts  $(y_{1,j}, y_{2,j})$  is sampled with respect to  $Poisson(\lambda_1)$  and  $Poisson(\lambda_{2,j})$ ,

$$y_{1,j} \sim Poisson(\lambda_1), \quad y_{2,j} \sim Poisson(\lambda_{2,j})$$

3. Zero mean Gaussian noise is then added to each gene count (rounding to the nearest integer using the rounding operator  $[\cdot]$ ):

$$\begin{aligned} y'_{i,j} &= y_{i,j} + [\eta_j], \quad i = 1, 2 \\ \eta_j &\sim N\left(0, \sigma_j = \frac{v_j}{\psi}\right) \\ v_j &= \frac{\lambda_1 + \lambda_{2,j}}{2} \end{aligned}$$

where  $\psi = 10$ .

4. The batch and lane effects are simulated as follows. Batch effects are accounted for by adding Gaussian noise to each noisy count  $y'_{i,j}$ ,

$$\begin{aligned} y''_{i,j} &= y'_{i,j} + [\eta'_{i,j}] \\ \eta'_{i,j} &\sim N\left(0, \frac{y'_{i,j}}{10}\right) \end{aligned}$$

Lane effects are simulated by Poisson sampling from  $y''_{1,j}$  and  $y''_{2,j}$  at different rates varying between lanes,

$$\begin{aligned} x_{i,j} &\sim Poisson(\delta_j \cdot y''_{i,j}) \\ \delta_j &\sim \text{Uniform}\{0.65, 0.8, 0.95\} \end{aligned}$$

## References

1. Varuzza, L.; Gruber, A.; de B. Pereira, C. Significance tests for comparing digital gene expression profiles. *Nat. Preced.* **2008**, hdl:10101/npre.2008.2002.3.
2. Audic, S.; Claverie, J. The significance of digital expression profiles. *Genome Res.* **1997**, *7*, 986–995.
3. Medina, C.; Rotter, B.; Horres, R.; Udupa, S.; Besser, B.; Bellarmino, L.; Baum, M.; Matsumura, H.; Terauchi, R.; Kahl, G.; *et al.* SuperSAGE: The drought stress-responsive transcriptome of chickpea roots. *BMC Genomics* **2008**, *9*, e553.
4. Kim, H.; Baek, K.; Lee, S.; Kim, J.; Lee, B.; Cho, H.; Kim, W.; Choi, D.; Hur, C. Pepper EST database: Comprehensive *in silico* tool for analyzing the chili pepper (*Capsicum annuum*) transcriptome. *BMC Plant Biol.* **2008**, *8*, e101.
5. Cervigni, G.; Paniego, N.; Pessino, S.; Selva, J.; Diaz, M.; Spangenberg, G.; Echenique, V. Gene expression in diplosporous and sexual *Eragrostis curvula* genotypes with differing ploidy levels. *BMC Plant Biol.* **2008**, *67*, e11.
6. Miles, J.; Blomberg, A.; Krisher, R.; Everts, R.; Sonstegard, T.; Tassell, C.V.; Zeulke, K. Comparative transcriptome analysis of *in vivo* and *in vitro*-produced porcine blastocysts by small amplified RNA-serial analysis of gene expression (SAR-SAGE). *Mol. Reprod. Dev.* **2008**, *75*, 976–988.
7. Cuevas-Tello, J.C.; Tiño, P.; Raychaudhury, S. How accurate are the time delay estimates in gravitational lensing? *Astron. Astrophys.* **2006**, *454*, 695–706.
8. Cuevas-Tello, J.C.; Tiño, P.; Raychaudhury, S.; Yao, X.; Harva, M. Uncovering delayed patterns in noisy and irregularly sampled time series: An astronomy application. *Pattern Recognit.* **2010**, *43*, 1165–1179.
9. Pelt, J.; Hjorth, J.; Refsdal, S.; Schild, R.; Stabell, R. Estimation of multiple time delays in complex gravitational lens systems. *Astron. Astrophys.* **1998**, *337*, 681–684.
10. Press, W.; Rybicki, G.; Hewitt, J. The time delay of gravitational lens 0957+561, I. Methodology and analysis of optical photometric Data. *Astrophys. J.* **1992**, *385*, 404–415.
11. Tiño, P. Basic properties and information theory of audic-claverie statistic for analyzing cDNA arrays. *BMC Bioinform.* **2009**, *10*, e310.
12. Tiño, P. One-shot Learning of Poisson Distributions in cDNA Array Analysis. In *Advances in Neural Networks*, Proceedings of the 8th International Symposium on Neural Networks (ISNN 2011), Guilin, China, 29 May – 1 June, 2011; Liu, D., Zhang, H., Polycarpou, M., Alippi, C., He, H., Eds.; Lecture Notes in Computer Science (LNCS 6676), Springer-Verlag: Berlin, Heidelberg, Germany, 2011; pp. 37–46.
13. Auer, P.; Doerge, R. Statistical design and analysis of RNA sequencing data. *Genetics* **2010**, *185*, 405–416.