BAYESIAN ORACLE FOR BOUNDING INFORMATION GAIN IN NEURAL ENCODING MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

In recent years, deep learning models have set new standards in predicting neural population responses. Most of these models currently focus on predicting the mean response of each neuron for a given input. However, neural variability around this mean is not just noise and plays a central role in several theories on neural computation. To capture this variability, we need models that predict full response distributions for a given stimulus. However, to measure the quality of such models, commonly used correlation-based metrics are not sufficient as they mainly care about the mean of the response distribution. An interpretable alternative evaluation metric for likelihood-based models is *Information Gain* (IG) which evaluates the likelihood of a model relative to a lower and upper bound. However, while a lower bound is usually easy to obtain, constructing an upper bound turns out to be challenging for neural recordings with relatively low numbers of repeated trials, high (shared) variability, and sparse responses. In this work, we generalize the jackknife oracle estimator for the mean—commonly used for correlation metrics—to a flexible Bayesian oracle estimator for IG based on posterior predictive distributions. We describe and address the challenges that arise when estimating the lower and upper bounds from small datasets. We then show that our upper bound estimate is data-efficient and robust even in the case of sparse responses and low signal-to-noise ratio. We further provide the derivation of the upper bound estimator for a variety of common distributions including the state-of-the-art zero-inflated mixture models, and relate IG to common mean-based metrics. Finally, we use our approach to evaluate such a mixture model resulting in 90% IG performance.

1 INTRODUCTION

In recent years, systems neuroscience has seen great advancements in building neural encoding models of population activity [23; 1; 3; 10; 20; 15; 6; 22]. Most of these models focus on estimating the conditional mean of the response distribution given a stimulus and are consequently evaluated on mean-based measures such as correlation or fraction of explainable variance explained (FEVE). However, neural responses exhibit a great deal of variability even when the animal is presented with the same stimulus. This variability is not just noise, but might be a symptom of underlying neural computations. In fact, many normative theories that link first principles to neural response properties, like the Bayesian brain hypothesis [17], neural sampling [11; 4] or probabilistic population codes [16], make predictions or rely on the variability of neural activity around the mean [14; 12; 5]. If we want to use neural encoding models as a quantitative underpinning for these theories, models are needed which accurately predict and are evaluated on complete response distributions. While progress has been made at building such models [21; 2], it is not clear what upper bound on the performance we can expect. However, this question is important as it gives us an indication how close our models are to the true system.

In the case of mean-predicting models, correlation-based metrics are often used for evaluation [15; 8]. Correlation is an interpretable measure since it is naturally bounded between -1 and 1. However, for vanilla correlation, it is impossible for any model to achieve a correlation of 1 in the presence of trial-to-trial fluctuations. Therefore, model correlation is often normalized by an upper bound oracle estimator [18; 15], which is commonly obtained by computing point estimates of the conditional mean using the responses to repeated presentations of the same stimulus. For a likelihood-based metric, a similar normalization to a bounded and interpretable scale would be desirable. To this end,

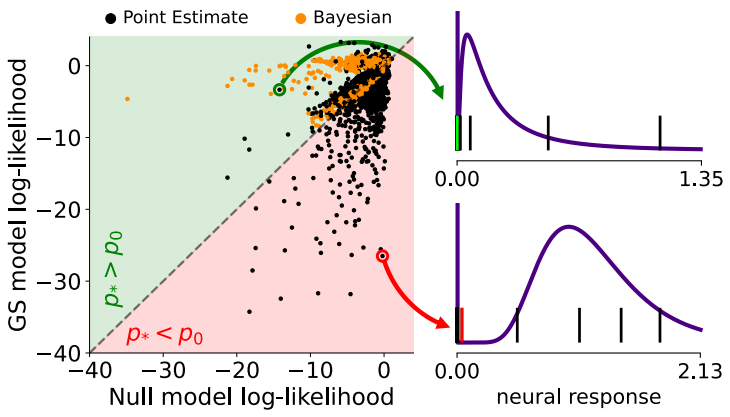

Figure 1: Comparison of lower and upper bound likelihood estimates (Null vs GS) per neuron. **Left:** For many neurons, the PE approach yields worse GS than the Null score. The Bayesian method results in the expected outcome of upper bound scores being higher than lower bound scores. **Right:** Two example neurons demonstrating where the PE method fails (red) or succeeds (green).

one can use Information Gain (IG) [13], which uses an estimate of both upper and lower bound, to put the likelihood between two meaningful values. However, the challenge lies in how these bounds can be obtained for noisy neural responses.

In this work, we develop a robust way to estimate such lower and upper bounds for IG on neuronal responses. We show that a point estimate approach for obtaining the upper bound fails and demonstrate that this is caused by the lack of robustness for the estimate of moments beyond the mean. This is especially pronounced when dealing with data that have few samples, sparse responses, and low signal-to-noise ratios which are common characteristics of neural responses. To mitigate this problem, we propose a generalization of the point estimate approach to a full Bayesian treatment using posterior predictive distributions. Our approach yields lower and upper bounds which are prove to be robust to all the above-mentioned complexities in neural data. We derive a general expression for the Bayesian estimator for zero-inflated distributions that can be efficiently estimated under very general conditions by solving only a single one-dimensional integral on a bounded interval. These distributions capture the sparse nature of neural responses, in particular for 2-photon recordings, and include state-of-the-art zero-inflated mixture models [21; 2]. Using this full-likelihood-based metric, we then evaluate a zero-inflated mixture model and find that it performs remarkably well at 90% IG. Finally we experimentally and mathematically relate IG to other common metrics for the performance of neural prediction models which are based on the mean and derive general conditions under which likelihood and correlation as a metric identify the same predictive function.

2 Methods

2.1 Information Gain and problems with point estimates

Information Gain Let $p(y|x)$ denote the distribution of a neuron's response y to a stimulus x. In order to evaluate and interpret the modeled distribution $\hat{p}(y|x)$ we use Information Gain (IG) [13; 19] which sets the model likelihood on an interpretable scale between an estimated lower and upper bound:

$$IG = \frac{\langle \log \hat{p}(y \mid x) \rangle_{y,x} - \langle \log p_0(y) \rangle_{y,x}}{\langle \log p_*(y \mid x) \rangle_{y,x} - \langle \log p_0(y) \rangle_{y,x}} \tag{1}$$

using a Null distribution $p_0(y)$ and a Gold Standard distribution $p_*(y|x)$. This method of computing IG can be interpreted as a normalized comparison of lower bounds of mutual information [19, and Appendix 4]. The *Null model* should reflect basic aspects of the response. Here, we choose a Null model that does not account for any stimulus-related information, resulting in the marginal distribution of responses: $p_0(y|x) := \hat{p}(y)$. The *Gold Standard (GS) model* $p_*(y|x)$, on the other hand, should be the best possible approximation of the true conditional distribution $p(y|x)$. It can also be an oracle model that has more information than the model under evaluation, such as access to repeated presentations of the same stimulus. Here, we use such an oracle and estimate its parameters from responses to repeated stimulus presentations. Importantly, we do this in a leave-one-out fashion:

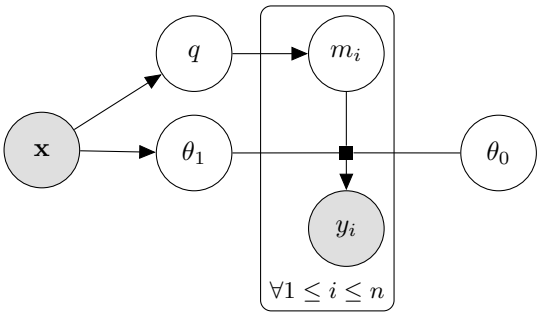

Figure 2: Graphical model for a zero-inflated distribution. A stimulus $\mathbf{x}$ determines the probability q whether a neurons fires or not and the parameters θ_1 of the response distribution of the non-zero response distribution. A Bernoulli random variable m_i determines whether a neuron fires on a particular trial i. If $m_i = 1$ a response is drawn from $p(y_i|\theta_1)$, otherwise from $p(y_i|\theta_0)$.

given a set of n repeats, the GS parameters of a target repeat i are estimated from $n-1$ left-out repeats $\backslash i$. However, as we demonstrate below, estimating a robust GS model can be challenging.

Point Estimate (PE) GS model The parameters θ of the upper bound estimator can be obtained as point estimates (PE) from $n-1$ left-out repeats:

$$p_*(y_i|\mathbf{y}_{\backslash i}) = p(y_i|\theta_i) \quad \text{with } \theta_i = f(\mathbf{y}_{\backslash i}).$$

Often, moment matching is used to obtain the point estimate of θ. For correlation-based performance metrics of neural prediction models, a jack-knifed mean estimator over repeated presentations of the same stimulus $E[y_i|x] = \frac{1}{n-1}\sum_{y_j \in y_{\backslash i}} y_j$ is commonly used as an oracle predictor for the conditional mean to obtain an upper bound on the achievable performance in the presence of noise [18; 15].

Problems with the PE approach To demonstrate the problems with point estimate GS models, we modeled neural responses with a zero-inflated Log-Normal likelihood and estimated the upper bound using the PE approach (see Appendix A.1 for details on data, [21; 2] for details on zero-inflated distributions, and Appendix B for the moment matching derivations). Since the GS model estimates parameters per stimulus, it should yield higher likelihood values than the Null model whose parameters are not stimulus-specific. However, applying the PE approach to neural data, we observed that the Null model outperforms the GS model for the majority of neurons (Fig. 1, black points). The reason for this effect is that the PE approach is sensitive to the sparse distribution of the data, which combined with few responses per stimulus results in an overconfident estimation of the GS parameters (see Fig. 1 on the right).

2.2 Bayesian Gold Standard Model

Gold Standard Model based on posterior predictive distributions To avoid an overconfident GS model, we add uncertainty to the parameter estimation and estimate the GS model in a fully Bayesian fashion via the full posterior predictive distribution:

$$p_*(y_i|\mathbf{y}_{\backslash i}) = \int_{-\infty}^{\infty} \underbrace{p(y_i|\theta)}_{\text{likelihood}} \underbrace{p(\theta|\mathbf{y}_{\backslash i})}_{\text{posterior}} \, \mathrm{d}\theta \tag{2}$$

Note that the PE approach is a special case within this framework for which $p(\theta|\mathbf{y}_{\backslash i}) = \delta(\theta - f(\mathbf{y}_{\backslash i}))$ is collapsed onto a delta distribution.

Efficient estimation for zero-inflated distributions In general, the integral in equation 2 is intractable and the posterior predictive distribution can only be evaluated using numerical approximations. Only for certain choices of likelihood the integral can be solved analytically if the right conjugate prior was chosen.

Here, we show that the posterior predictive distribution can be efficiently computed for the class of zero-inflated distributions [21, Fig. 2]. In principle, such a distribution is a mixture of a delta-distribution at zero and a density of a positive part. In practice, we replace the delta-distribution with a uniform distribution in a small interval $[0, \tau)$ and shift the positive part to the interval $[\tau, \infty)$ so that the two mixture components do not overlap

The probability density for a single stimulus x is then defined as:

$$p(y) = (1-q) \cdot \underbrace{p_u(y)}_{\text{uniform}} + q \cdot \underbrace{p(y|\theta_1)}_{\text{positive distribution on } [\tau,\infty)}, \tag{3}$$

where the mixing proportion q and the parameters θ_1 are dependent on x. When using a predictive model, q and θ_1 are typically predicted from the stimulus x using an encoding model. Zero-inflated likelihoods are the basis of current state-of-the-art likelihood-based neural encoding models [21; 2].

Here we show that the posterior predictive distribution for a zero-inflated likelihood boils down to a one dimensional integral over q if the posterior predictive distribution of the positive part is known.

Lemma 1. *The posterior predictive distribution of a zero-inflated distribution as defined in equation 3 is given by*

$$p(y_i|\mathbf{y}_{\setminus i}) = \begin{cases} p(y_i|\mathbf{y}^0_{\setminus i}) \cdot \int_q (1-q) \cdot p(q|\mathbf{y}_{\setminus i})\, dq & \text{if } y_i < \tau \\ p(y_i|\mathbf{y}^1_{\setminus i}) \cdot \int_q q \cdot p(q|\mathbf{y}_{\setminus i})\, dq & \text{if } y_i \geq \tau \end{cases}$$

where $\mathbf{y}^0_{\setminus i}$ and $\mathbf{y}^1_{\setminus i}$ denote the set of zero and non-zero responses in $\mathbf{y}_{\setminus i}$, respectively.

Proof. See Appendix C. □

This means that the posterior predictive distribution can be computed efficiently for a large class of positive distributions, including the zero-inflated Log-Normal or even zero-inflated Flow models [2]. In the next section, we demonstrate that this Bayesian treatment yields a GS model that is more robust against outliers and yields higher likelihoods than the Null model, as expected (Fig. 1, orange points).

3 Experiments

3.1 Analysis of Gold Standard Models

In this section, we investigate why the Bayesian approach outperforms the PE and test its robustness under different numbers of left-out repeats $\mathbf{y}_{\setminus i}$ and different signal-to-noise ratios. For this analysis, we used responses from 7672 neurons to 360 stimuli, each repeated 20 times as well as simulated data. Details about the datasets are provided in Appendix A.1 and Appendix A.2, respectively. We base our posterior predictive distribution on a zero-inflated Log-Normal distribution with zero-threshold $\tau = \exp(-10)$ and parameters $\theta_1 = (\mu, \sigma^2)$, referring to the first and second moment of the log-transformed positive responses (Appendix C). The conjugate prior for the Log-Normal part of the distribution is a Normal-Inverse-Gamma distribution $p(\theta_1) = \mathcal{N}G^{-1}(\mu, \lambda, \alpha, \beta)$ in log space whose parameters can be chosen freely. We discuss choices of prior parameters later in section 3.2.

Bayesian GS estimates higher order moments better. In order to determine which parameters of the likelihood profits most from the Bayesian estimation, we compared GS models where the individual parameters are either estimated via the PE or the Bayesian approach (Fig. 3**a**). For this we used the largest number of left-out repeats $n-1 = 19$ available in our real neuron dataset. First, we observe that the likelihood improves with the Bayesian estimation of μ (orange vs. yellow bar) as well as σ^2 (light blue vs. yellow bar) individually. Consequently, the highest performance is achieved when both parameters are estimated via the Bayesian approach (dark blue vs. yellow bar). Interestingly, the relative gain in performance is much higher for σ^2 than for the μ, reflecting a lower robustness of the higher moments compared to the first moment in log space.

Bayesian GS is data-efficient Datasets can vary in how many repeats per stimulus they contain. Since a metric should be comparable across datasets, IG ought to yield consistent estimates for different numbers of left-out repeats $\mathbf{y}_{\setminus i}$, in particular in the regime of low $n-1$. The right panel of Fig. 3**b** shows this on neural data where we first observe that the results of Fig. 3**a** are qualitatively consistent for different numbers of repeats. The effect of the Bayesian parameter estimation on the likelihood performance, however, is much more pronounced in the low $n-1$ regime: As the number of left-out repeats decreases the PE approach suffers much more than the Bayesian and it completely fails at $n = 3$ (vanishing yellow bar). To test the higher $n-1$ regime, we simulated

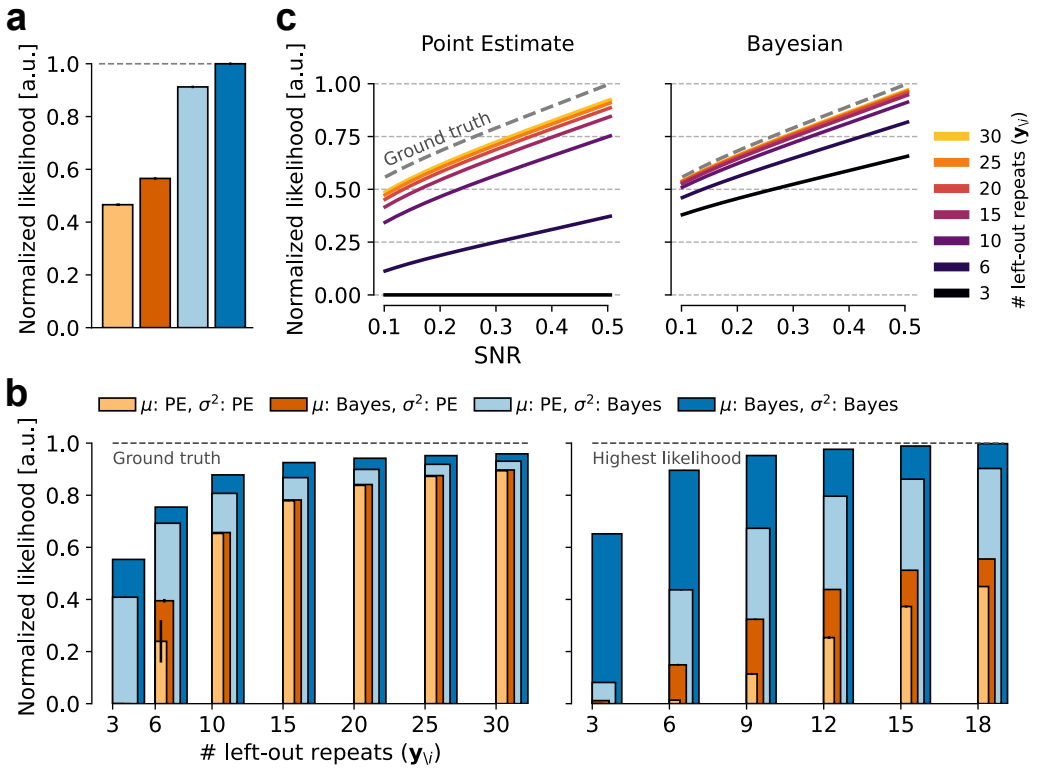

Figure 3: Comparison of the point estimate and Bayesian GS model. **a:** Comparison of different GS models where the individual parameters are either estimated via the PE or the Bayesian approach. The number of left-out repeats $\mathbf{y}_{\setminus i}$ is 19. Colors are the same as in **b**. Normalized wrt. the max. likelihood value, i.e. the dark blue bar. **b:** Similar to **a** but for different numbers of left-out repeats $\mathbf{y}_{\setminus i}$. **Left:** Simulated data. The likelihood is normalized w.r.t. the ground truth likelihood. **Right:** Neural responses. Likelihood is normalized w.r.t. the maximum likelihood value, i.e. the dark blue bar at 18 left-out repeats. **c:** Upper bound likelihood scores for different signal-to-noise ratios and different number of left-out repeats $\mathbf{y}_{\setminus i}$. **Left:** Point Estimate. **Right:** Bayesian. Likelihood is normalized w.r.t. the ground truth likelihood. In all panels the likelihood values are averaged over stimuli and neurons, and the error bars and shaded areas show SEM over 5 random selections of the left-out repeats.

neural responses since the real neural dataset contained maximally 20 repeats. In the left panel of Fig. 3**b** we explored the differences between the two approaches for up to 30 repeats and observed that the Bayesian treatment consistently yields a better likelihood than the PE estimate (yellow bar does not completely converge to dark blue bar). The probabilistic treatment of μ, however, seems to become less important in the high $n-1$ regime than that of σ^2, reflecting the higher robustness of the first moment compared to the second moment in log space (compare the difference between orange and yellow for high vs. low $n-1$).

Bayesian GS is robust to different SNRs Apart from different numbers of repeats, datasets can also vary in terms of signal-to-noise ratio. We therefore simulated neural data with different underlying means and variances per stimulus, resulting in different SNR values (see Appendix A.2 for details). We then tested Bayesian and point estimate GS models on this data (Fig. 3 **c**) and observed that the Bayesian approach consistently outperforms the PE approach across all SNRs (data for Fig. 3**b** left panel had an SNR of 0.42).

3.2 Evaluating neural encoding models via IG

In Section 3.1 we demonstrated that the Bayesian approach for obtaining an upper bound estimator greatly outperforms the naive point estimate (PE) approach in several aspects. From now on, we

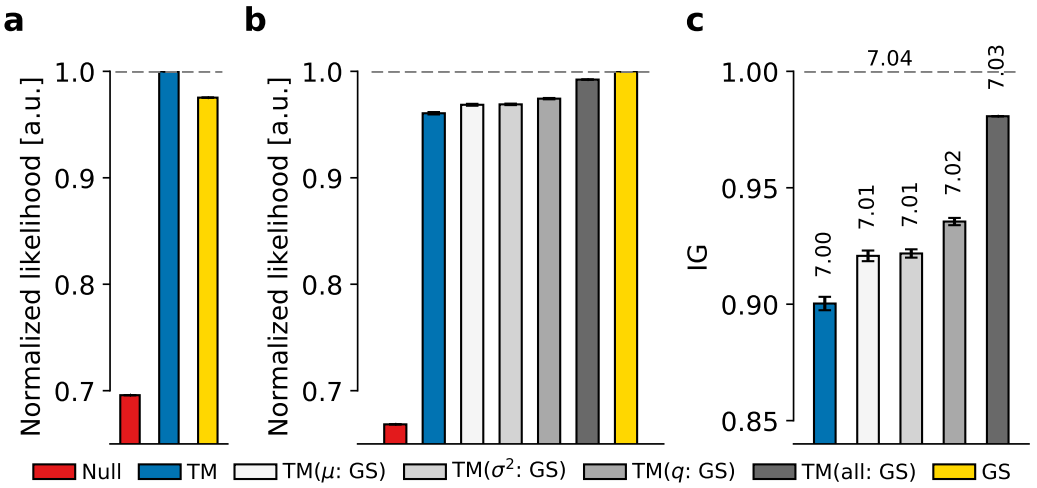

Figure 4: Real neural data evaluation of a neural encoding model trained on a zero-inflated Log-Normal likelihood (ZIL) using the Bayesian upper bound estimate (GS). **a:** A sub-optimal prior for the GS model results in a lower performance than the trained model (TM). Normalized wrt. the max. likelihood value, i.e. the blue bar. **b:** The GS model outperforms the TM when its prior was optimized. The TM estimates all parameters similarly well. Grey scale colors indicate models from the same distribution as the TM (ZIL) for which all parameters are estimated by the trained model except for the parameters in parentheses. Normalized wrt. the max. likelihood value, i.e. the yellow bar. **c:** Information Gain (IG) for the TM and the models grey scale models from **b**. Values on top of bars indicate the likelihood per image and per neuron in bits. In all panels the likelihood values are averaged over stimuli and neurons, and the error bars and shaded areas show SEM over 5 random initializations of the TM weights. There are no errorbars on the red, yellow and dark grey bars since they do not involve a TM.

therefore only use the Bayesian estimator when referring to the Gold Standard Model. In this section, we use the lower and upper bound estimates to evaluate neural encoding models via Information Gain.

To this end, we trained an encoding model on a dataset which is publicly available [15] containing the responses of 5335 neurons in mouse primary visual cortex to 5094 unique natural images. The dataset was split into 4472 stimulus-response pairs for training, 522 for validation, and 100 for testing. The stimuli in the test set were each repeated 10 times resulting in $n - 1 = 9$ repeats for the GS model to be computed on. The neural encoding model and the training are the same as in the model provided by Lurz et al. [15] with a few minor differences: While Lurz et al. [15] uses Poisson loss to train the models, we chose the negative loglikelihood of a zero-inflated Log-Normal (ZIL) distribution as a loss function. This means that the model needs to predict three parameters (q, μ, σ^2) instead of a single mean firing rate λ as in Lurz et al. [15]. The readout thus learns three feature vectors which all read out from the same learned position of the core network. Since the metric that the model will be evaluated on is IG, the early-stopping criterion was changed from correlation to likelihood.

The results of these experiments with neural encoding models are summarized in Fig. 4 and will be explained in detail in the following paragraphs.

Choice of prior is crucial for the GS model Up to this point, we chose the prior hyper-parameters neuron-independently: We chose the parameters of the Normal-Inverse-Gamma prior based on the average conditional mean $\mathrm{E}_x[\mathrm{E}_y[y|x]]$ and average conditional variance $\mathrm{E}_x[\mathrm{Var}_y[y|x]]$ of our dataset. This left us with number of neurons samples which we fit the prior on, resulting in one identical prior for every neuron. However, we observed that the resulting GS model (yellow bar) was outperformed by the trained encoding model (TM model, blue bar) and did not provide a good estimate of a performance upper bound (Fig. 4 **a**). To obtain a true upper bound oracle model, we therefore optimized the prior hyper-parameters directly on the sum of leave-one-out GS likelihoods via gradient descent for each single neuron. This is analogous to other oracle models that have access

to more information than the predictive model under evaluation, thus providing an upper bound. Note that this approach results in a more conservative estimate of the model performance (Fig. 4 **b**, compare the blue and yellow bars).

Encoding model captures all parameters similarly well In order to investigate which parameters of the response distribution the encoding model predicts well, we conducted an experiment similar to the one shown in Fig. 3 **a**: We compared the likelihood of the trained model (TM, blue) to cases where we matched one or all of its three parameters (q, μ, σ^2) to the GS model (see Fig. 4 **b** blue bar vs. grey bars). While we could match the parameter q directly, we used moment matching to obtain the parameters of the non-zero part of the trained model (Log-Normal distribution) from the non-zero part of the GS model (Log-Student-t). We observed that the likelihood of the trained model improved when each of the three parameters were matched with the GS model individually (three lighter grey bars vs. blue bar), where q yielded the highest increase. Matching all three parameters, however, did not result in the same performance as the GS model itself (darkest grey vs. yellow). Since all parameters of the ZIL model were fitted on the GS model, the reason must be the difference in distributional shape of the positive part: Log-Normal for ZIL vs. Log-Student-t for the GS model.

Encoding model performance is at 90% IG The final IG value of the trained model using the Null model and the GS model with optimized prior hyper-parameters can be seen in Fig. 4 **c**. It performs remarkably well at 90% IG (blue bar), which corresponds to a likelihood of 7.00 bits per image and neuron (printed value above the bar). The effect of the parameter matching (grey bars) is more emphasized and suggests that the largest performance gain can be achieved in future models by improving the prediction of the parameter q.

4 Relation between IG and other metrics

Neural predicting models have so far been mostly evaluated using metrics such as fraction of explainable variance explained (FEVE) [7], correlation, and fraction oracle [18; 15]. While we cannot expect that there is a one-to-one relationship between IG and these metrics, as IG is sensitive to the entire response distribution whereas the commonly used metrics mostly focus on the mean response of a neuron for a given stimulus, we can nevertheless relate IG to these metric under certain assumptions. In this section we provide a summary of the relationships to these metrics. Details and proofs can be found in Appendices E and F.

Generally, we will demonstrate the relations in two steps: First we show that IG linearly depends on the expected Kullback-Leibler (KL) divergence between the true distribution $p(y|x)$ and the model distribution $\hat{p}(y|x)$. Then we derive how this KL divergence relates to other metrics.

IG is a linear function of $\langle D_{KL}[p(y|x), \hat{p}(y|x)]\rangle_x$ The IG described in Eq. 1 can be re-written in terms of KL divergences between estimated/fitted distributions and the true conditional distribution $p(y|x)$ (see Appendix D for the detailed derivation):

$$IG = \frac{\langle D_{KL}[p(y|x)||p_0(y)]\rangle_x - \langle D_{KL}[p(y|x)||\hat{p}(y|x)]\rangle_x}{\langle D_{KL}[p(y|x)||p_0(y)]\rangle_x - \langle D_{KL}[p(y|x)||p_*(y|x)]\rangle_x} \tag{4}$$

This shows that the Information Gain is a linear function of $\langle D_{KL}[p(y|x)||\hat{p}(y|x)]\rangle_x$ with a negative slope. It is maximized when average Kullback-Leibler divergence between the true distribution and the model distribution is minimized.

IG vs. FEVE We will now make use of this fact to show that IG is linear in another commonly used metric, fraction of explainable variance explained (FEVE), for a Gaussian likelihood. For a fixed model noise variance $\hat{\sigma}_\epsilon^2$ the KL-divergence $\langle D_{KL}[p(y|x)||\hat{p}(y|x)]\rangle_x$ is a linear function of FEVE:

$$\langle D_{KL}[p(y|x)||\hat{p}(y|x)]\rangle_x = f(\hat{\sigma}_\epsilon) + \frac{1}{2}(1 - FEVE) \times \frac{\sigma_s^2}{\hat{\sigma}_\epsilon^2}, \tag{5}$$

where $f(\hat{\sigma}_\epsilon) = \log\left(\frac{\hat{\sigma}_\epsilon}{\sigma_\epsilon}\right) - \frac{1}{2} + \frac{\sigma_\epsilon^2}{2\hat{\sigma}_\epsilon^2}$ and σ_s^2 is the signal variance. Since $\langle D_{KL}[p(y|x)||\hat{p}(y|x)]\rangle_x$ and IG are also linearly related, then IG too is a linear function of FEVE. Note that when the estimated

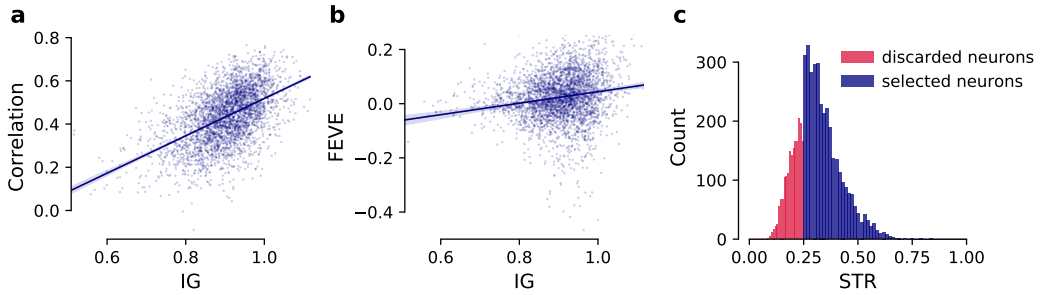

Figure 5: Per neuron comparison of IG with correlation and FEVE. The data comes from the TM in Fig. 4 (one of the 5 models from different initializations). **a:** Correlation vs. IG. **b:** FEVE vs. IG. **c:** The neurons displayed in panel **a** and **b** were selected based on whether their signal-to-total-variance ratio (STR) was above 0.25. The blue lines depict linear regressions with 95% confidence intervals.

noise variance is the same as the true noise variance, then Eq. 5 becomes:

$$\langle D_{KL}[p(y|x)||\hat{p}(y|x)]\rangle_x = \frac{1}{2}(1 - FEVE) \times SNR$$

See Appendix E for detailed derivations. If the estimated noise variance $\hat{\sigma}_\epsilon^2$ is not fixed, as is the general case, this trend will still be true but the linear relationship between the IG and FEVE is more noisy. In Fig. 5**b**, we show that by evaluating the trained model from Fig. 4 on IG and FEVE per neuron. Note that we only display neurons whose signal-to-total-variance ratio $STR = \sigma_s^2/\sigma_y^2 = \sigma_s^2/(\sigma_s^2 + \sigma_\epsilon^2)$ exceeds a threshold which we set to 0.25 (see Fig. 5**c**).

IG vs. correlation Another commonly used metric is the correlation $\rho(\hat{\mu}_x, \mu_x)$ between the model prediction $\hat{\mu}_x = \langle \hat{y} \rangle_{\hat{y}|x}$ and trial averaged responses $\mu_x = \langle y \rangle_{y|x}$. Assuming a Gaussian likelihood, like in the relation to FEVE, we show that for a fixed model noise and signal variance $\langle D_{KL}[p(y|x)||\hat{p}(y|x)]\rangle_x$ is a linear function of the correlation between μ_x and $\hat{\mu}_x$:

$$\langle D_{KL}[p(y|x)||\hat{p}(y|x)]\rangle_x = f(\hat{\sigma}_\epsilon) + \frac{1}{2}\left(1 + \frac{\hat{\sigma}_s^2}{\sigma_s^2} - \frac{2\hat{\sigma}_s}{\sigma_s}\rho(\hat{\mu}_x, \mu_x)\right) \times \frac{\sigma_s^2}{\hat{\sigma}_\epsilon^2} \tag{6}$$

As before, if the model noise variance matches the true noise variance, $\sigma_\epsilon^2 = \hat{\sigma}_\epsilon^2$, we have:

$$\langle D_{KL}[p(y|x)||\hat{p}(y|x)]\rangle_x = \frac{1}{2}\left(1 + \frac{\hat{\sigma}_s^2}{\sigma_s^2} - \frac{2\hat{\sigma}_s}{\sigma_s}\rho(\hat{\mu}_x, \mu_x)\right) \times SNR$$

If we further assume that the model's signal variance matches the true signal variance, $\hat{\sigma}_s^2 = \sigma_s^2$, we get:

$$\langle D_{KL}[p(y|x)||\hat{p}(y|x)]\rangle_x = (1 - \rho(\hat{\mu}_x, \mu_x)) \times SNR$$

It is worth noting that the assumptions for having a linear relationship with KL divergence are stronger in the case of correlation than that of the FEVE. Specifically, there is an additional dependence on the signal variance in the case of correlation. This makes sense as FEVE is inherently sensitive to signal variance while correlation is not (see Appendix F for the detailed derivation). As in the case of FEVE, we empirically show (Fig. 5 **a**) the relation between IG and correlation on our trained model from Fig. 4. While the general linear trend can be observed, this relation is very noisy, as expected. Note again, that we only show neurons that pass the threshold of $STR \geq 0.25$.

Does a high likelihood always correspond to a high correlation? Correlation only reflects good estimation of the mean of a distribution (see Appendix G). IG, on the other hand is a likelihood-based metric and as such depends on the correct estimation of all parameters of the likelihood. So does a high likelihood correspond to a high correlation? The answer is: it depends.

Correlation is high if the conditional mean is accurately estimated. However, the mean of the distribution is in general not directly a parameter of the likelihood and is therefore not necessarily

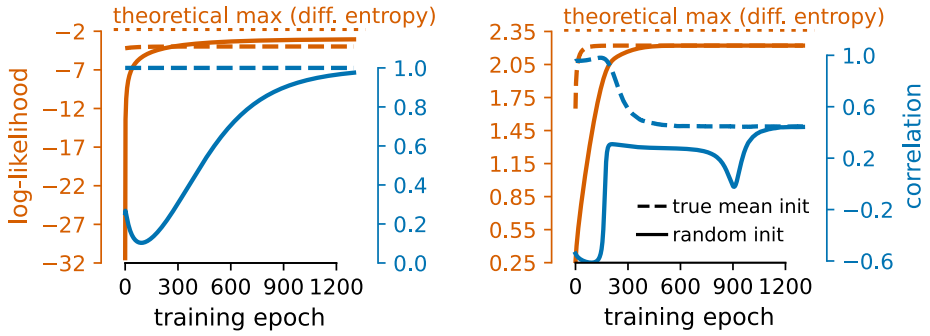

Figure 6: Comparison of log-likelihood and correlation of a model which is optimized via maximum likelihood estimation. **Left:** The model distribution is from the exponential family and has the mean as a sufficient statistic. The optima of the two metrics coincide but an improvement in likelihood does not necessarily mean an improvement in correlation. Data distribution: Normal. Model distribution: Gamma. **Right:** The model distribution does *not* fulfill these criteria. No similar behavior of likelihood and correlation can be guaranteed. If initialized with optimal correlation, the correlation after training has decreased (dotted line). Data distribution: Gamma. Model distribution: Chi-squared.

accurately estimated even if optimal likelihood is achieved. Therefore there is in general no guarantee that a model that outperforms another model on likelihood also outperforms this other model on correlation. However, if the distribution is 1) a member of the exponential family and 2) has the mean as a sufficient statistic [9], an optimal likelihood also implies an optimal correlation. A distribution that fulfills these two criteria guarantees that for optimal likelihood, the mean is optimally estimated and with it the correlation. We demonstrate this effect using simulated data (see Appendix A.2) in Fig. 6 in the left panel: A model of such a distribution is being optimized on toy data via maximum likelihood estimation and at the end of training the optima of likelihood and correlation coincide. Note however that during training, i.e. for non-optimal likelihood, a training step which improves likelihood does not necessarily yield an improvement of correlation (compare decreasing blue line vs. increasing orange line). If the distribution does not fulfill the second criterion, i.e. if the mean is not a sufficient statistic, no relation between the correlation and likelihood can be guaranteed, not even at the optimum. This can be seen in Fig. 6 in the right panel: An improvement in likelihood (orange) does not guarantee an improvement in correlation (blue). If the model was initialized such that the correlation is optimal (dotted lines), the maximum likelihood estimation will even result in a model that has lower correlation than it had at the time of initialization (compare dotted blue line at epoch 0 vs 1200).

5 Conclusion

In this work, we discussed the challenges in obtaining the lower and upper bound estimates for likelihood-based measures like Information Gain (IG) in order to put the performance on an interpretable scale. We introduced a robust way of obtaining such an upper bound by using the Bayesian framework of posterior predictive distributions. Equipped with this metric, we showed that current neural encoding models are able to predict full response distributions to up to 90% IG and we examined which parameters of the distribution the model still fails at predicting. We also gave a detailed derivation for obtaining upper bound estimates for the state-of-the-art family of distributions, the zero-inflated distributions. Finally, we related likelihood-based metrics like IG to other metrics which are commonly used in Computational Neuroscience like correlation and FEVE.

There are, of course, also some limitations to the current work. While we were able to fix the catastrophic failure of naive PE estimates (Fig. 1), our estimates of IG or the GS model are not perfect as some neurons yield IG > 1 (Fig. 5) which is not mathematically impossible but would ideally not happen. Even better IG or GS estimators will mitigate this issue. Finally, our approach is for single neuron likelihoods. However, the neural variation is structured on a population level, for instance through brain states. Since the general approach of IG still works in that situation, our framework is flexible enough for a robust GS model of entire populations to be derived in future work.

ACKNOWLEDGMENTS

Hidden during review.

REFERENCES

[1] Ján Antolík, Sonja B Hofer, James A Bednar, and Thomas D Mrsic-Flogel. Model constrained by visual hierarchy improves prediction of neural responses to natural scenes. *PLoS computational biology*, 12(6):e1004927, 2016.

[2] Mohammad Bashiri, Edgar Walker, Konstantin-Klemens Lurz, Akshay Jagadish, Taliah Muhammad, Zhiwei Ding, Zhuokun Ding, Andreas Tolias, and Fabian Sinz. A flow-based latent state generative model of neural population responses to natural images. *Advances in Neural Information Processing Systems*, 34:15801–15815, 2021.

[3] Eleanor Batty, Josh Merel, Nora Brackbill, Alexander Heitman, Alexander Sher, Alan Litke, EJ Chichilnisky, and Liam Paninski. Multilayer recurrent network models of primate retinal ganglion cell responses. 2016.

[4] Pietro Berkes, Gergő Orbán, Máté Lengyel, and József Fiser. Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science*, 331(6013):83–87, 2011.

[5] Adrian G Bondy, Ralf M Haefner, and Bruce G Cumming. Feedback determines the structure of correlated variability in primary visual cortex. *Nature neuroscience*, 21(4):598–606, 2018.

[6] Max F Burg, Santiago A Cadena, George H Denfield, Edgar Y Walker, Andreas S Tolias, Matthias Bethge, and Alexander S Ecker. Learning divisive normalization in primary visual cortex. *PLOS Computational Biology*, 17(6):e1009028, 2021.

[7] Santiago A Cadena, George H Denfield, Edgar Y Walker, Leon A Gatys, Andreas S Tolias, Matthias Bethge, and Alexander S Ecker. Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLoS computational biology*, 15(4):e1006897, 2019.

[8] Santiago A Cadena, Konstantin F Willeke, Kelli Restivo, George Denfield, Fabian H Sinz, Matthias Bethge, Andreas S Tolias, and Alexander S Ecker. Diverse task-driven modeling of macaque v4 reveals functional specialization towards semantic tasks. *bioRxiv*, 2022.

[9] Justin Domke. Moment-matching conditions for exponential families with conditioning or hidden data. *arXiv preprint arXiv:2001.09771*, 2020.

[10] Alexander S Ecker, Fabian H Sinz, Emmanouil Froudarakis, Paul G Fahey, Santiago A Cadena, Edgar Y Walker, Erick Cobos, Jacob Reimer, Andreas S Tolias, and Matthias Bethge. A rotation-equivariant convolutional neural network model of primary visual cortex. *arXiv preprint arXiv:1809.10504*, 2018.

[11] József Fiser, Pietro Berkes, Gergő Orbán, and Máté Lengyel. Statistically optimal perception and learning: from behavior to neural representations. *Trends in cognitive sciences*, 14(3): 119–130, 2010.

[12] Ralf M Haefner, Pietro Berkes, and József Fiser. Perceptual decision-making as probabilistic inference by neural sampling. *Neuron*, 90(3):649–660, 2016.

[13] Matthias Kümmerer, Thomas SA Wallis, and Matthias Bethge. Information-theoretic model comparison unifies saliency metrics. *Proceedings of the National Academy of Sciences*, 112 (52):16054–16059, 2015.

[14] Richard D Lange, Ankani Chattoraj, Jeffrey M Beck, Jacob L Yates, and Ralf M Haefner. A confirmation bias in perceptual decision-making due to hierarchical approximate inference. *PLoS Computational Biology*, 17(11):e1009517, 2021.

[15] Konstantin-Klemens Lurz, Mohammad Bashiri, Konstantin Friedrich Willeke, Akshay Kumar Jagadish, Eric Wang, Edgar Y Walker, Santiago Cadena, Taliah Muhammad, Eric Cobos, Andreas Tolias, et al. Generalization in data-driven models of primary visual cortex. *bioRxiv*, 2020.

[16] W J Ma, J M Beck, P E Latham, and A Pouget. Bayesian inference with probabilistic population codes. *Nat. Neurosci.*, 9:1432–1438, 2006.

[17] Wei Ji Ma, Jeffrey M Beck, Peter E Latham, and Alexandre Pouget. Bayesian inference with probabilistic population codes. *Nature neuroscience*, 9(11):1432–1438, 2006.

[18] Fabian Sinz, Alexander S Ecker, Paul Fahey, Edgar Walker, Erick Cobos, Emmanouil Froudarakis, Dimitri Yatsenko, Zachary Pitkow, Jacob Reimer, and Andreas Tolias. Stimulus domain transfer in recurrent models for large scale cortical population prediction on video. *Advances in neural information processing systems*, 31, 2018.

[19] Lucas Theis, Andrè Maia Chagas, Daniel Arnstein, Cornelius Schwarz, and Matthias Bethge. Beyond glms: a generative mixture modeling approach to neural system identification. *PLoS computational biology*, 9(11):e1003356, 2013.

[20] Edgar Y Walker, Fabian H Sinz, Erick Cobos, Taliah Muhammad, Emmanouil Froudarakis, Paul G Fahey, Alexander S Ecker, Jacob Reimer, Xaq Pitkow, and Andreas S Tolias. Inception loops discover what excites neurons most using deep predictive models. *Nature neuroscience*, 22(12):2060–2065, 2019.

[21] Xue-Xin Wei, Ding Zhou, Andres Grosmark, Zaki Ajabi, Fraser Sparks, Pengcheng Zhou, Mark Brandon, Attila Losonczy, and Liam Paninski. A zero-inflated gamma model for deconvolved calcium imaging traces. *arXiv preprint arXiv:2006.03737*, 2020.

[22] Konstantin F Willeke, Paul G Fahey, Mohammad Bashiri, Laura Pede, Max F Burg, Christoph Blessing, Santiago A Cadena, Zhiwei Ding, Konstantin-Klemens Lurz, Kayla Ponder, et al. The sensorium competition on predicting large-scale mouse primary visual cortex activity. *arXiv preprint arXiv:2206.08666*, 2022.

[23] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014.