Maximally informative models and diffeomorphic modes in the analysis of large data sets

Justin B. Kinney
Simons Center for Quantitative Biology
Cold Spring Harbor Laboratory
Cold Spring Harbor, NY 11724
jkinney@cshl.edu

Gurinder S. Atwal
Simons Center for Quantitative Biology
Cold Spring Harbor Laboratory
Cold Spring Harbor, NY 11724
atwal@cshl.edu

Abstract

Motivated by data-rich experiments in transcriptional regulation and sensory neuroscience, we consider the following general problem in statistical inference. A system of interest, when exposed to a stimulus S, adopts a deterministic response R of which a noisy measurement M is made. Given a large number of measurements and corresponding stimuli, we wish to identify the correct "response function" relating R to S. However the "noise function" relating M to R is unknown a priori. Here we show that maximizing likelihood over *both* response functions and noise functions is equivalent to simply identifying *maximally informative* response functions – ones that maximize the mutual information $I[R;M]$ between predicted responses and corresponding measurements. Moreover, if the correct response function is in the class of models being explored, maximizing mutual information becomes equivalent to simultaneously maximizing *every* dependence measure that satisfies the Data Processing Inequality. We note that experiments of the type considered are unable to distinguish between parametrized response functions lying along certain "diffeomorphic modes" in parameter space. We show how to derive these diffeomorphic modes and observe, fortunately, that such modes typically span a very low-dimensional subspace. Therefore, given sufficient data, maximizing mutual information can pinpoint nearly all response function parameters without requiring any model of experimental noise.

1 Introduction

This paper discusses a familiar problem in statistical inference, but focuses on an under-studied limit which is becoming increasingly relevant in both neuroscience and molecular biology. Consider an experiment having the following form:

$$\underset{\text{stimulus}}{S} \xrightarrow[\theta(S)]{\text{response function}} \underset{\text{response}}{R} \xrightarrow[\pi(M|R)]{\text{noise function}} \underset{\text{measurement}}{M}. \tag{1}$$

When presented with a stimulus S, a system of interest adopts a deterministic response R, of which a noisy measurement M is made. Specifically, stimuli are drawn from a probability distribution $p(S)$, the response R to each stimulus is determined by a "response function" θ, and a measurement M is thus generated with probability given by the "noise function" $\pi(M|R)$. We refer to this as an "SRM-type" experiment. From a large number of independent stimulus-response pairs, (S_n, M_n), $n = 1, 2, \ldots, N$, we wish to reconstruct θ.

This is a standard regression problem and is typically solved [3] by first assuming a specific noise function π, then searching a space Θ of model response functions for the one $\theta \in \Theta$ which maxi-

mizes the likelihood

$$p(\{M_n\} \,|\, \{S_n\}, \theta, \pi) = e^{N\mathcal{L}(\theta,\pi)} \quad \text{where} \quad \mathcal{L}(\theta,\pi) = \frac{1}{N}\sum_{n=1}^{N} \log \pi(M_n|\theta(S_n)). \tag{2}$$

For instance, the method of least squares regression corresponds to assuming a Gaussian noise function π. Often the assumed noise model is only approximate, or is adopted primarily for analytical convenience. Nevertheless, an incorrect π can work reasonably well, allowing one to infer a tolerably accurate response model when the data are limiting.

However, certain experiments in sensory neuroscience and transcriptional regulation operate in the data-saturated limit. In such cases, systematic error in θ caused by an incorrect noise function can dominate over the uncertainty due to finite sampling. Such bias has been documented when the receptive fields of sensory neurons are characterized using natural stimuli. In one study [24], anesthetized cats were shown a series of woodland scenes, providing neurons in V1 cortex with stimuli S of $\sim 10^2$ - 10^3 pixels each. Measurements $M \in \{\text{spike}, \text{no spike}\}$ were taken for individual V1 neurons until as many spikes as relevant pixels had been recorded, yielding $N \gg N_{spike} \gtrsim \dim(\theta)$. From these data the authors inferred a receptive field for each neuron, defined as a stimulus vector $\hat{e}$ such that the projection $R = S \cdot \hat{e}$ determined spiking probability. Inference using the standard reverse-correlation spike-triggered average [21], corresponding to maximum likelihood with $\pi(\text{spike}|R) \sim \exp[R]$ and assuming Gaussian stimuli, was shown to strongly bias the inferred receptive field.

Analogous experiments probing the fine structure of the transcriptional regulatory code are now possible [9, 16, 10, 14, 17, 22, 13], thanks to the development of ultra-high-throughput DNA sequencing technologies. To characterize how a specific transcriptional regulatory sequence (TRS) functions, a large number ($\sim 10^4$ - 10^6) of variants S of the TRS are used to control the expression of a gene, and a measurement M of the transcription rate R resulting from each variant is made. Modeling the quantitative dependence R has on S can then be used to characterize the sequence-dependent energy with which each regulatory protein binds the TRS, as well as measure the interaction energies between bound factors [10]. In this case, the possibility of systematic error from the inference procedure distorting biochemical measurements presents a serious concern.

An alternative inference procedure that is free of systematic bias is to maximize the mutual information [5] between predictions R and measurements M,[1]

$$I(\theta) = I[R;M] = \int dR\, dM\; p(R,M) \log \frac{p(R,M)}{p(R)p(M)}. \tag{3}$$

Here, $p(R,M)$ is the empirical joint distribution of predictions and measurements, and thus depends implicitly on θ.[2] This method has been proposed and applied in the specific contexts of both receptive field inference [23, 24, 15, 18] and transcriptional regulation [8, 7, 9, 10, 14]. However, a general discussion of how maximizing mutual information relates to maximizing likelihood has yet to be presented.

Here we study the general problem of identifying optimal responses models in the $N \to \infty$ limit when the noise function π is unknown *a priori*. We show that maximizing $I(\theta)$ is equivalent to maximizing $\mathcal{L}(\theta,\pi)$ over both θ and π, and further becomes equivalent to simultaneously maximizing every dependence measure which satisfies the Data Processing Inequality (DPI) [5] when some candidate θ fully explains the data. Tests for whether or not an inferred θ fully explains the data are also described. We then address the issue that SRM-type experiments cannot distinguish between θ within certain equivalence classes. This leads to "diffeomorphic modes" in parameter space which cannot be pinned down by data. An equation for diffeomorphic modes is presented, and is used to derive all the diffeomorphic modes of general linear models and a specific linear-nonlinear model that has been studied previously [10].

Throughout this manuscript, R is specifically used to represent predictions of the model θ, i.e. $R = \theta(S)$ for an implicit stimulus S. Similarly $R^* = \theta^*(S)$, $R_1 = \theta_1(S)$, etc.. Responses R

[1] The notation $I(\theta)$ and $I[R;M]$ will be used interchangeably.

[2] For $I(\theta)$ to work as an objective function, one typically expects $N \gg \dim(\theta)$ will be required for reliable estimation of $p(R,M)$ under all choices of θ. A rapid and accurate method for estimating the density $p(R,M)$ from finite data is also needed.

are assumed to be multidimensional with components $\{R^\mu\}$, and $\partial_\mu \equiv \partial/\partial R^\mu$. θ denotes both a response model and the parameters of that model. Θ is used to represent both an abstract space of response models, as well as the space of parameters for models θ assumed to have a specific functional form. In the latter case, $\{\theta^i\}$ denotes coordinates in parameter space, and $\partial_i \equiv \partial/\partial\theta^i$. Implicit summation notation over repeated indices i or μ is assumed.

2 Mutual information and likelihood

In the $N \to \infty$ limit, the per-datum log likelihood of the pair (θ, π) can be decomposed as follows,

$$\mathcal{L}(\theta, \pi) = \int dR\, dM\; p(R, M) \log \pi(M|R) = I(\theta) - D(\theta, \pi) - H[M]. \tag{4}$$

The first term on the right is the mutual information (Eq. 3), which is independent of π. The second term,

$$D(\theta, \pi) = \int dR\, dM\; p(R, M) \log \frac{p(M|R)}{\pi(M|R)}, \tag{5}$$

is the Kullback-Leibler (KL) divergence between the empirical distribution $p(M|R)$ observed for the response model θ and the assumed noise function $\pi(M|R)$, and thus depends on both θ and π. The last term, $H[M] = -\int dM\; p(M) \log p(M)$, is the entropy of the measurements M, is independent of both θ and π, and can thus be ignored in the optimization problem.

In the $N \to \infty$ limit, the problem of finding pairs (θ, π) which maximize $\mathcal{L}(\theta, \pi)$ is identical to the problem of only finding response functions θ which maximize $I(\theta)$. This follows from the fact that, for a given choice of θ, choosing π to match the empirical noise model, $\pi(M|R) = p(M|R)$, globally minimizes $D(\theta, \pi)$ and causes it to vanish. The maximum likelihood problem therefore reduces to the problem of finding θ which maximize $\max_\pi \mathcal{L}(\theta, \pi) = I(\theta) - H[M]$; this is identical to the problem of maximizing $I(\theta)$.

It has been noted that when N is large and one's prior knowledge about π can be formalized with a prior $p(\pi)$, then the per-datum log of the marginal likelihood, $\mathcal{L}(\theta)$, is essentially equal to the mutual information $I(\theta)$ [8, 19]. This can be seen by computing the marginal likelihood,

$$p(\{M_n\} \,|\, \{S_n\}, \theta) = \int d\pi\; p(\pi) p(\{M_n\} \,|\, \{S_n\}, \theta, \pi) = e^{N[I(\theta) - \Delta(\theta) - H[M]]} \tag{6}$$

where

$$\Delta(\theta) = -\frac{1}{N} \log \left[\int d\pi\; p(\pi) e^{-N D(\theta, \pi)} \right] \tag{7}$$

is the only term affected by the prior $p(\pi)$. Under weak assumptions about $p(\pi)$, $\Delta \to 0$ as $N \to \infty$.[3] Therefore,

$$\mathcal{L}(\theta) \equiv \frac{1}{N} \log p(\{M_n\} \,|\, \{S_n\}, \theta) = I(\theta) - \Delta(\theta) - H[M], \tag{8}$$

is equal to $I(\theta)$ up to a constant and a θ-dependent correction which vanishes as $N \to \infty$.

3 DPI-optimal response models

In practice, one typically searches for optimal θ within a limited class Θ of possible response models. In this section, we present results which obtain when any model $\theta \in \Theta$ fully explains the data, i.e. $I(\theta) = I(\theta^*)$ where θ^* is the true response function.

[3] In certain cases $\Delta(\theta)$ can be computed explicitly and thus be shown to vanish [8]. More generally, when π is taken to be finite-dimensional, a saddle-point computation (valid for large N) gives $\Delta(\theta) \approx \frac{1}{2N} \mathrm{Tr}[\log \partial\partial \tilde{D}] + \mathrm{const}$. Here, $\partial\partial\tilde{D}$ is the π-space Hessian of $\tilde{D}(\theta, \pi) \equiv D(\theta, \pi) - \frac{1}{N} \log p(\pi)$ computed at $\pi(M|R) = p(M|R)$. If $\log p(\pi)$ and its derivatives are bounded, then the θ-dependent part of $\Delta(\theta)$ decays as N^{-1}. If π is infinite dimensional, this saddle-point computation becomes a problem in field theory akin to the inference problem studied by [1]. If this field theory is properly formulated through an appropriate choice of $p(\pi)$, $\Delta(\theta)$ can be expected to exhibit different decay behavior, but still vanish as $N \to \infty$. See also [19].

First we observe that $\theta = \theta^*$ globally maximizes $I(\theta)$ over all possible response functions. Given any hypothesized θ together with θ^*, and letting π^* denote the true noise function, the chain of stochastic variables

$$R \xleftarrow{\theta(S)} S \xrightarrow{\theta^*(S)} R^* \xrightarrow{\pi^*(M|R^*)} M \tag{9}$$

forms a Markov chain [5], i.e. $p(R, S, R^*, M) = p(R|S)p(S)p(R^*|S)p(M|R^*)$. The fact that mutual information satisfies DPI allows us to read off the inequality

$$I[R; M] \leq I[S; M] = I[R^*; M], \tag{10}$$

proving that $\theta = \theta^*$ globally maximizes $I[R; M] = I(\theta)$.

As has been noted [15], the same argument can be made not just for mutual information, but for any dependence measure $\mathcal{D}[M; R]$ which satisfies DPI: the simple fact that Eq. 9 is a Markov chain implies $\mathcal{D}[R; M] \leq \mathcal{D}[R^*; M]$, proving $\theta = \theta^*$ globally maximizes $\mathcal{D}(\theta) \equiv \mathcal{D}[R; M]$.[4] Letting $\Theta_{\mathcal{D}} \subseteq \Theta$ denote the set of all $\theta \in \Theta$ which maximize a dependence measure $\mathcal{D}(\theta)$, we see that if $\theta^* \in \Theta$, then $\theta^* \in \Theta_{\mathcal{D}}$ for every $\mathcal{D}$ satisfying DPI. So in fact θ^* must be contained within the infinite intersection of all such $\Theta_{\mathcal{D}}$, which we shall denote by Θ_{DPI}:

$$\theta^* \in \Theta_{DPI} \equiv \bigcap_{\mathcal{D}\ satisfying\ DPI} \Theta_{\mathcal{D}}. \tag{12}$$

We now prove that when any $\theta \in \Theta_I$ achieves $I(\theta) = I(\theta^*)$, the set Θ_{DPI} of such "DPI-optimal" response models is in fact identical to the set Θ_I of maximally informative models, i.e.

$$\Theta_I = \Theta_{DPI}. \tag{13}$$

First, since mutual information satisfies DPI, $\Theta_{DPI} \subseteq \Theta_I$. Next, the fact (from Eq. 9) that $R \leftrightarrow R^* \leftrightarrow M$ is a Markov Chain means R contains no information about M which is not conveyed by R^*, and so $I[R^*; M] = I[R^*, R; M]$. This gives,

$$I[R^*; M] - I[R; M] = I[R^*, R; M] - I[R; M] = I[R^*; M|R] \tag{14}$$

This conditional mutual information $I[R^*; M|R]$ must be the same for all $\theta \in \Theta_I$. If we further assume $I(\theta) = I(\theta^*)$ for some (and thus all) $\theta \in \Theta_I$, then $I[R^*; M|R] = 0$. This implies $p(R^*, M|R) = p(R^*|R)p(M|R)$, or equivalently, $p(M|R^*, R) = p(M|R)$. Therefore, $R^* \leftrightarrow R \leftrightarrow M$ is also a Markov chain. Reconciling this with the fact that $R \leftrightarrow R^* \leftrightarrow M$ is a Markov chain as well, we get $\mathcal{D}[R; M] = \mathcal{D}[R^*; M]$ for any DPI-satisfying $\mathcal{D}$. All such $\mathcal{D}$ are therefore maximized by θ, meaning $\Theta_I \subseteq \Theta_{DPI}$. This completes the proof.

We pause to offer some intuition for these results. For any hypothesized θ, the resulting joint distribution $p(R, M)$ will be a convolution of the true joint distribution $p(R^*, M)$ with the conditional distribution $p(R|R^*)$,

$$p(R, M) = \int dR^* p(R|R^*)p(R^*, M). \tag{15}$$

In general the reverse is *not* true, i.e. $p(R^*, M) \neq \int dR\ p(R^*|R)p(R, M)$, because $p(R^*|R) \neq p(R^*|R, M)$. This reflects a basic asymmetry among joint distributions $p(R, M)$: sometimes one can be derived from another by convolution, sometimes not.

All DPI-satisfying measures $\mathcal{D}[R, M]$ are either decreased or left unchanged by such convolutions. Every such $\mathcal{D}$ therefore imposes a weak ordering on the space of joint distributions. When neither of two distributions $p(R, M)$ and $p(R', M)$ can be expressed as a convolution of the other, then different DPI satisfying measures $\mathcal{D}$ can potentially rank these distributions differently. However, when

[4] We note that there are an infinite number of dependence measures other than mutual information which satisfy DPI. For instance, information measures of the f-divergence form [6, 15],

$$I_f[M; R] \equiv \int dR\, dM p(R)p(M) f\left(\frac{p(R, M)}{p(R)p(M)}\right) \tag{11}$$

satisfy DPI when the function $f(x)$ is convex for $x \geq 0$. Mutual information corresponds to $f(x) = x \log x$, while $f(x) = (\alpha - 1)^{-1} x^\alpha$ for $\alpha \geq 0$ is the more general Rényi divergence [20, 11].

$p(R, M)$ can be gotten from $p(R', M)$ by convolution, then $\mathcal{D}[R; M] \leq \mathcal{D}[R'; M]$ is guaranteed. Thus, because all $p(R, M)$ under consideration derive from a single $p(R^*, M)$ by a convolution of the form in Eq. 15, *every* DPI-satisfying measure $\mathcal{D}$ ranks $p(R^*, M)$ no lower than any other $p(R, M)$. So if $\theta^* \in \Theta$, one gets $\theta^* \in \Theta_{DPI}$.

The equivalence $\Theta_I = \Theta_{DPI}$, realized when $I(\theta) = I(\theta^*)$ for $\theta \in \Theta_I$, stems from the fact that mutual information is maximally sensitive to such convolutions: if $\mathcal{D}[R; M] < \mathcal{D}[R^*; M]$ for *any* measure $\mathcal{D}$ satisfying DPI, then $I[R; M] < I[R^*; M]$. Mutual information is not unusual in this regard. For example, every information measure $I_f[M; R]$ for which $f(x)$ is *strictly convex* satisfies $\Theta_{I_f} = \Theta_{DPI}$. There are, however, some dependence measures which are less sensitive than mutual information: the trivial dependence measure $\mathcal{D} = 0$ satisfies DPI, but reveals nothing about $p(R, M)$.

4 Are the data fully explained?

We now discuss how to check whether a given $\theta \in \Theta_I$ fully explains the data, i.e. whether $I(\theta) = I(\theta^*)$. Verifying that a maximally informative model is also fully informative is an important part of the modeling process. Showing $I(\theta) \neq I(\theta^*)$ for any $\theta \in \Theta_I$ will prove that the available data require a different (or enlarged) space Θ of response models. On the other hand, showing $I(\theta) = I(\theta^*)$ means no further information about θ can be gotten from the data in hand.

One method [4] is to directly measure the total stimulus-dependent information $I[S; M]$ in the measurements. From Eq. 10 this is seen to equal $I(\theta^*)$. To do this, we rewrite the formula for $I[S; M]$ as

$$I[S; M] = \int dS\, dM\, p(S, M) \log \frac{p(S, M)}{p(M)p(S)} = H[M] - \langle H_S[M] \rangle_S \tag{16}$$

where the expectation value $\langle \cdot \rangle_S$ is taken over stimuli S drawn from $p(S)$, and

$$H_S[M] = -\int dM\, p(M|S) \log p(M|S) \tag{17}$$

is the measurement entropy for a particular stimulus S. If one has many measurements for a given stimulus S, the entropy $H_S[M]$ can be estimated. If such measurements are available for a *representative* sample of stimuli S, the expectation value $\langle H_S[M] \rangle_S$ in Eq. 16 can also be estimated. This approach has been applied to experiments in both sensory neuroscience [24] and transcriptional regulation [9]. In practice, however, experiments must be appropriately designed in order to provide the measurements needed to estimate $H_S[M]$ for a large, representative sample of stimuli.

We therefore propose a second test which does not require modifications to the experiment. Repeating the argument of Eq. 14 with S in place of R^*, one sees that $I(\theta) = I(\theta^*)$ implies $S \leftrightarrow R \leftrightarrow M$ is a Markov chain, i.e. $p(S|R, M) = p(S|R)$. Because of this, any function $f(S)$ will satisfy

$$\langle f \rangle_{S|R,M} = \langle f \rangle_{S|R} \tag{18}$$

for all R and M.[5] The converse is true as well: if Eq. 18 is satisfied for all functions $f(S)$, then $I(\theta) = I(\theta^*)$. This can be seen by considering $f(S) = \delta(R^* - \theta^*(S))$, in which case Eq. 18 gives $p(R^*|R, M) = p(R^*|R)$. If this holds for all R^*, then $R^* \leftrightarrow R \leftrightarrow M$ is a Markov chain, and so $I[R; M] = I[R^*; M]$.

Therefore, if any function $f(S)$ can be found which violates Eq. 18 for any $\theta \in \Theta_I$, then $\theta^* \notin \Theta$. A down-side to this test is its open-ended nature. One must try different functions $f(S)$, of which there are an infinite number. We suggest that, as a practical matter, choosing $f(S) = \theta'(S)$ for other $\theta' \in \Theta$ encountered in the process of searching for Θ_I might make sense. Alternatively, setting f equal to the components of the gradient $\partial_i R^\mu$ seems sensible, since Eq. 18 applied to $f = \partial_i R^\mu$ causes $\partial_i I(\theta)$ to vanish at $\theta = \theta^*$ [23, 9].

5 Information equivalence and diffeomorphic modes

Certain response models cannot be distinguished from one another by *any* SRM-type experiment because their predictions are always equally informative about measurements. We say that two such

[5] $\langle \cdot \rangle_{S|R,M}$ denotes averaging with respect to $p(S|R, M)$; $\langle \cdot \rangle_{S|R}$ corresponds to averaging over $p(S|R)$.

models θ_1 and θ_2 are "information equivalent", and write $\theta_1 \simeq \theta_2$, since this insensitivity of SRM-type experiments leads to a natural equivalence relation among response models. While Θ_I can sometimes be influenced by a specific experiment's stimulus distribution $p(S)$ and noise function $\pi^*(M|R)$ (an issue we will not pursue here), information equivalence places hard constraints on the structure of Θ_I, implying that certain equivalence classes within Θ must either be fully contained within Θ_I or fully excluded.

We now prove that $\theta_1 \simeq \theta_2$ if and only if the predictions of θ_1 and θ_2 are isomorphic, i.e. there exists an invertible function f such that $\theta_1(S) = f(\theta_2(S))$ and $\theta_2(S) = f^{-1}(\theta_1(S))$ for all possible stimuli S. First, such an isomorphism implies $p(M|R_1) = p(M|f(R_2)) = p(M|R_2)$, which means $I[R_1; M] = I[R_2; M]$, and thus $\theta_1 \simeq \theta_2$. Going the other direction, we can imagine an SRM-type experiment in which $\theta^* = \theta_1$ and $p(M|R^*) = \delta(M - R^*)$. Earlier we showed that, if $I[R_2; M] = I[R^*; M]$, then $R^* \leftrightarrow R_2 \leftrightarrow M$ is a Markov chain. With our choice of response function and noise function, $R_1 \leftrightarrow R_2 \leftrightarrow R_1$ is thus a Markov chain, implying R_1 must be a deterministic function of R_2. Imagining the same experiment with $\theta^* = \theta_2$ instead, we see that this function must be invertible.

If all $\theta \in \Theta$ of have a specific parametric form, information-equivalence implies that moving a response model θ along certain directions in parameter space may not change $I(\theta)$. Consider what happens when θ, having parameters $\{\theta^i\}$, is infinitesimally transported along a vector field $g^i(\theta)$, yielding a new model θ' with components $\theta'^i = \theta^i + \epsilon g^i(\theta)$. Each prediction R, having components $\{R^\mu\}$ in response space, will thus be transformed to $R'^\mu = R^\mu + \epsilon g^i(\theta)\partial_i R^\mu$. If $\theta' \simeq \theta$, the change $R'^\mu - R^\mu = \epsilon g^i(\theta)\partial_i R^\mu$ must, for all stimuli S, be fully specified by the value of predictions R and parameters θ and not otherwise depend on S. There must therefore be a vector field $h^\mu(R, \theta)$ in response space satisfying

$$g^i(\theta)\partial_i R^\mu = h^\mu(R, \theta). \tag{19}$$

We refer to vector fields $g^i(\theta)$ which satisfy this equation as "diffeomorphic modes". Movement of any model θ along its corresponding vector $g^i(\theta)$ induces a diffeomorphism of responses, predicted for all possible stimuli, defined by flows of the vector field $h^\mu(R, \theta)$.[6]

Importantly, diffeomorphic modes correspond to continuous changes of model parameters which cannot, in principle, be constrained by SRM-type data. The parametric form assumed for all $\theta \in \Theta$ determines which diffeomorphic modes exist, and identifying these modes analytically is critical when analyzing real data. For instance, if one is able to sample Θ_I, e.g. using Monte Carlo techniques, then the position of $\theta \in \Theta_I$ along diffeomorphic modes may have to be artificially fixed in order to arrive at values for individual model parameters. We therefore turn to the problem of computing diffeomorphic modes for models of different functional form.

5.1 General linear response models

Linear response models are of particular interest. In neuroscience they are commonly used to represent neuron receptive fields, and the resulting challenge of identifying "maximally informative dimensions" in stimulus space has received focused attention [23, 24]. In transcriptional regulation, linear "energy matrix" models are often used to represent the sequence-dependent binding energies of transcription factors, and the problem of inferring these from microarray data [8, 7] and DNA sequence data [9, 10, 14] has also been studied.

Here we derive the diffeomorphic modes of arbitrary linear response models. Assume

$$R^\mu = \theta^i F_i^\mu(S) \tag{21}$$

for some set of stimulus features $F_i^\mu(S)$. Note that these models are linear in their parameters but not necessarily linear in the stimulus S. To find the diffeomorphic modes, we apply Eq. 19 to Eq.

[6] Alternatively, one can define diffeomorphic modes in terms of the generator equation [9],

$$g^i(\theta)\partial_i = h^\mu(R, \theta)\partial_\mu. \tag{20}$$

We have found that working with this formulation eases notation and aids interpretability when deriving the diffeomorphic modes of a specific parametric model, but we will use Eq. 19 in what follows for the sake of concreteness.

21, giving $g^i(\theta)F_i^\mu(S) = h^\mu(\theta^i F_i^\nu(S), \theta)$. The left hand side is linear in stimulus features, and so $h^\mu(R,\theta)$ must[7] also be a linear function of R, i.e. have the highly restricted form

$$h^\mu(R,\theta) = a^\mu(\theta) + b^\mu_\nu(\theta)R^\nu. \tag{22}$$

Thus, the number of diffeomorphic modes of a general linear model, given by the number of parameters on which h^μ depends at each θ, is bounded above by $\dim(R)[\dim(R)+1]$. Importantly, this bound is independent of the number of stimulus features (i.e. $\dim(S)$); it depends only on the dimension of response space. In particular, if R is a scalar, then there are at most 2 diffeomorphic modes, corresponding to additive and multiplicative transformations of R.

5.2 A linear-nonlinear response model

We now show that combining multiple linear response models into a single linear-nonlinear model can eliminate diffeomorphic modes. This fact proved useful in a recent study by Kinney et al., [10]. In the context of their work, each stimulus S was a mutated version of a 75 base pair region of the *Escherichia coli lac* promoter DNA. A linear response function P was used to model the binding energy of RNA polymerase to its site on this promoter, while a separate linear function Q was used to model the interaction of the transcription factor CRP to its promoter binding site. The resulting rate of mRNA transcription was represented by the "regulation factor" R [2], which is related to the equilibrium occupancy of RNAP polymerase at its binding site ($occupancy = [1+R^{-1}]^{-1}$), and thus to the rate of mRNA transcription. In terms of P and Q, the regulation factor R was given by

$$R = e^{-P}\frac{1+e^{-Q-\gamma}}{1+e^{-Q}} \tag{23}$$

where γ is the interaction energy between CRP and RNA polymerase. Note that, in this equation, the energies P, Q, and γ are all in units of k_BT.

We now derive the diffeomorphic modes of R. Since P and Q depend on sequence features that can be varied independently, any diffeomorphic mode of R has to be a diffeomorphic mode of *both* P and Q. Since P and Q are linear in their parameters, and the only other parameter in the model is γ, any diffeomorphic mode of R must have the form

$$g^i(\theta)\partial_i R = h(R,\theta) = (a_P + b_P P)\partial_P R + (a_Q + b_Q Q)\partial_Q R + a_\gamma \partial_\gamma R. \tag{24}$$

Again, the coefficients $a_P, b_P, a_Q, b_Q, a_\gamma$ can be arbitrary functions of any of the model parameters, but cannot depend on S. Computing the derivatives and then substituting for P in terms of Q and R, we find

$$h(R,\theta) = -R\left[a_P - b_P \log\left\{\frac{R(1+e^{-Q})}{1+e^{-Q-\gamma}}\right\} - \frac{(a_Q+b_Q Q)e^{-Q}(1-e^{-\gamma})}{(1+e^{-Q-\gamma})(1+e^{-Q})} + \frac{a_\gamma e^{-Q-\gamma}}{1+e^{-Q-\gamma}}\right]. \tag{25}$$

For $g^i(\theta)$ to be a diffeomorphic mode, the right hand side must be independent of S for fixed R. But Q depends on S, so we must have $b_P = a_Q = b_Q = a_\gamma = 0$.[8] Diffeomorphic modes of R are thus defined by only one parameter, a_P, and satisfy

$$g^i(\theta)\partial_i R = -a_P R, \tag{26}$$

corresponding to an additive shift of P.

In [10], measurements for $\sim 5\times 10^4$ mutant *lac* promoters were used to infer models for P and Q individually as well as in the context of R. When P and Q were inferred individually, each was determined only up to an unknown affine transformation. However, when P and Q were inferred simultaneously by fitting R, three of the four diffeomorphic modes of P and Q vanished, leaving only the additive mode shown in Eq. 26. Thus, inferring the nonlinear function R allowed the binding energies of RNA polymerase and CRP to be determined in meaningful physical units (k_BT), and the intracellular concentration of CRP, which manifests as an additive contribution to Q, to be pinned down [10]. In fact, of the 204 independent parameters which defined the model in Eq. 23, the only parameter which could not be pinned down by data was the single diffeomorphic mode in Eq. 26, corresponding to changes in RNA polymerase concentration.

[7] There are exceptions to this statement, e.g. if the various features $F_i^\mu(S)$ exhibit complicated interdependencies, either because of their functional form or because stimuli S are restricted to a particular subspace. We ignore such pathological cases here.

[8] This assumes $\gamma \neq 0$.

6 Discussion

Its inability to pin down diffeomorphic modes distinguishes mutual information from likelihood in an important and revealing way. When maximizing likelihood with an assumed noise model, all response model parameters are constrained by data.[9] However, the constraints likelihood places on diffeomorphic modes come entirely from the KL-divergence (Eq. 4), which enforces the assumption that the empirical noise function $p(M|R)$ should match the assumed noise function $\pi(M|R)$. The relative likelihood of response models θ along diffeomorphic modes is $\exp[-ND(\theta, \pi)]$, and so the weight given to one's assumed noise function π grows with N. If there is any uncertainty whatsoever about what the true noise function is, this term will become overly presumptuous when N is sufficiently large.

A more rigorous approach is to place an explicit prior on possible noise functions π, and then optimize the response model θ using the marginal likelihood in Eq. 6. This allows one's prior belief about the noise function to influence the choice of response model when N is small, but the relative influence of this prior diminishes as N becomes large. This "noise-function-averaged" or "error-model-averaged" likelihood can be computed explicitly in certain cases and has proven useful on real data [8]. However, in the large N limit the resulting inference procedure essentially amounts to first identifying maximally informative θ, then using the prior on π to fix the diffeomorphic modes.

Regardless of the specific implementation, one's inference procedure should reflect the fact that SRM-type experiments are fundamentally insensitive to diffeomorphic modes of the response model. Any constraints along diffeomorphic modes must come from a source of information other than the SRM data itself, e.g. a separate calibration experiment.

One might worry that a large number of response model parameters will be diffeomorphic, and that SRM-type experiments will effectively require an assumed noise function if they are to yield useful results. Such situations are conceivable, but in practice this is often not the case. When the stimulus S is high-dimensional and the response R is low dimensional, the vast majority of model parameters will typically be involved in reducing the dimensionality of S; very few will only parametrize diffeomorphisms of R. We showed that when θ is linear in its parameters, the number of diffeomorphic modes will not exceed $\dim(R)[\dim(R) + 1]$ (except in pathological cases) . This holds regardless of how large $\dim(\theta)$ is. In the specific linear-nonlinear model considered by [10] (Eq. 23), only one of the 204 independent parameters turned out to be diffeomorphic. So although diffeomorphic modes do appear in real-world applications, they are often very limited in number, and in such cases the vast majority of response model parameters can be inferred from SRM-type data without any systematic error stemming from an incorrect noise function.

Unfortunately, using mutual information as an objective function can present practical difficulties. One must be able to rapidly and reliably estimate $I(\theta)$ from finite data, and the resulting $I(\theta)$ may present a rugged optimization landscape. Still, various methods for estimating mutual information have been implemented (e.g. [12, 25]), and the information optimization problem has been successfully addressed in specific situations using stochastic gradient ascent [23, 24], standard Metropolis Monte Carlo [8], and parallel tempering Monte Carlo [10, 14]. How best to address these practical issues remains an open question, but we believe the exciting applications in neuroscience and molecular biology provide compelling reasons to make progress on these problems.

Acknowledgments

We thank Bill Bialek, Curtis Callan, Bud Mishra, Swagatam Mukhopadhyay, Anand Murugan, Mike Schatz, Bruce Stillman, and Gašper Tkačik for helpful conversations. Support for this project was provided by the Simons Center for Quantitative Biology at Cold Spring Harbor Laboratory.

References

[1] Bialek ,W., Callan, C., Strong, S. Field Theories for Learning Probability Distributions. *Phys. Rev. Lett.* **77**(23):4693-7 (1996).

[2] Bintu, L., Buchler, N., Garcia, H., Gerland, U., Hwa, T., Kondev, J., Phillips, R. Transcriptional regulation by the numbers: models. *Curr. Opin. Genet. Dev.* **15**(2):116-24 (2005).

[9] This excludes parameters of θ which have no effect on R, or which affect R in ways that do not alter the M-distribution $\pi(M|R)$.

[3] Bishop C. *Pattern Recognition and Machine Learning*. Springer, New York (2006).

[4] Brenner, N., Strong, S.P., Koberle, R., Bialek, W., de Ruyter van Steveninck, R.R. Synergy in a neural code. *Neural Comput.* **12**(7):1531-52 (2000).

[5] Cover, T., Thomas, J. *Elements of information theory*. John Wiley & Sons, New York (1991).

[6] Csiszár, I., Shields, P.C. *Information Theory and Statistics: A Tutorial*. now Publishers Inc., Hanover (2004).

[7] Elemento, O., Slonim, N., Tavazoie, S. A universal framework for regulatory element discovery across all genomes and data types. *Mol. Cell* **28**(2):337-50 (2007).

[8] Kinney, J.B., Tkačik, G., Callan, C.G. Precise physical models of protein-DNA interaction from high-throughput data. *Proc. Natl. Acad. Sci. USA* **104**(2):501-6 (2007).

[9] Kinney, J.B. *Biophysical models of transcriptional regulation from sequence data*. Ph.D. Thesis, Princeton University (2008).

[10] Kinney, J.B., Murugan, A., Callan, C.G., Cox, E. Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc. Natl. Acad. Sci. USA* **107**(20):9158-63 (2010).

[11] Kouh, M., Sharpee, T. Estimating linear-nonlinear models using Rényi divergences. *Network* **20**(2):49-68 (2009).

[12] Kraskov, A., Stgbauer, H., Grassberger, P. Estimating mutual information. *Phys. Rev. E* **69**:066138 (2004).

[13] Kwasnieski, J.C., Mogno, I., Myers, C.A., Corbo, J.C., Cohen, B.A. Complex effects of nucleotide variants in a mammalian *cis*-regulatory element. *Proc. Natl. Acad. Sci. USA* **109**(47):19498503 (2012).

[14] Melnikov, A., Murugan, A., Zhang, X., Tesileanu, T., Wang, L., Rogov, P., Feizi, S., Gnirke, A., Callan, C.G., Kinney, J.B., Kellis, M., Lander, E.S., Mikkelsen, T.S. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* **30**(3):271-7 (2012).

[15] Paninski L. Convergence properties of three spike-triggered analysis techniques. *Network-Comp. Neural.* **14**:437-64 (2003).

[16] Patwardhan, R.P., Lee, C., Litvin, O., Young, D., Pe'er, D., Shendure, J. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat. Biotechnol.* **27**(12):1173-5 (2009).

[17] Patwardhan, R.P., Hiatt, J.B., Witten, D.M., Kim, M.J., Smith, R.P., May, D., Lee, C., Andrie, J.M., Lee, S., Cooper, G.M., Ahituv, N., Pennacchio, L.A., Shendure, J. Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.* **30**(3):265-70 (2012).

[18] Pillow, J.W., Simoncelli, E.P. Dimensionality reduction in neural models: an information-theoretic generalization of spike-triggered average and covariance analysis. *J. Vision* **6**(4):414-28 (2006).

[19] Rajan, K., Marre, O., Tkačik, G. Learning quadratic receptive fields from neural responses to natural stimuli. arXiv:1209.0121 [q-bio.NC] (2012).

[20] Rényi, A. On measures of entropy and information. *Proc. 4th Berkeley Symp. Math. Statist. and Prob.* **1**:547-61 (1961).

[21] Schwartz, O., Pillow, J., Rust, N., Simoncelli, E. Spike-triggered neural characterization. *J. Vision* **6**(4):484-507 (2006).

[22] Sharon, E., Kalma, Y., Sharp, A., Raveh-Sadka, T., Levo, M., Zeevi, D., Keren, L., Yakhini, Z., Weinberger, A., Segal, E. Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nature Biotechnol.* **30**:521-30 (2012).

[23] Sharpee, T., Rust, N., Bialek, W. Analyzing neural responses to natural signals: maximally informative dimensions. *Neural Comput.* **16**(2):223-50 (2004).

[24] Sharpee, T., Sugihara, H., Kurgansky, A., Rebrik, S., Stryker, M., Miller, K. Adaptive filtering enhances information transmission in visual cortex. *Nature* **439**(7079):936-42 (2006).

[25] Slonim, N., Atwal, G.S., Tkačik, G., Bialek, W. Estimating mutual information and multi-information in large networks. arXiv:cs/0502017 [cs.IT] (2005).