

Parametric inference in the large data limit using maximally informative models

Justin B. Kinney^{1, 2} and **Gurinder S. Atwal**¹

¹Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory,
Cold Spring Harbor, NY 11724

² Please direct correspondence to jkinney@cshl.edu.

Keywords: mutual information; diffeomorphic modes; machine learning; statistical inference; sensory neuroscience; transcriptional regulation

Abstract

Motivated by data-rich experiments in transcriptional regulation and sensory neuroscience, we consider the following general problem in statistical inference. When exposed to a high-dimensional signal S , a system of interest computes a representation R of that signal which is then observed through a noisy measurement M . From a large number of signals and measurements, we wish to infer the “filter” that maps S to R .

However, the standard method for solving such problems, likelihood-based inference, requires perfect *a priori* knowledge of the “noise function” mapping R to M . In practice such noise functions are usually known only approximately, if at all, and using an incorrect noise function will typically bias the inferred filter. Here we show that, in the large data limit, this need for a pre-characterized noise function can be circumvented by searching for filters that instead maximize the mutual information $I[M; R]$ between observed measurements and predicted representations. Moreover, if the correct filter lies within the space of filters being explored, maximizing mutual information becomes equivalent to simultaneously maximizing every dependence measure that satisfies the Data Processing Inequality. It is important to note that maximizing mutual information will typically leave a small number of directions in parameter space unconstrained. We term these directions “diffeomorphic modes” and present an equation that allows these modes to be derived systematically. The presence of diffeomorphic modes reflects a fundamental and nontrivial substructure within parameter space, one that is obscured by standard likelihood-based inference.

1 Introduction

This paper discusses a familiar problem in statistical inference, but focuses on an understudied limit that is becoming increasingly relevant in the era of large data sets. Consider an experiment having the following form:

$$\begin{array}{ccccc} S & \xrightarrow{\text{filter}} & R & \xrightarrow{\text{noise function}} & M \\ \text{signal} & & \text{representation} & & \text{measurement} \end{array} \quad (1)$$

When presented with a signal S , a system of interest applies a deterministic filter θ thereby producing an internal representation R of that signal. For each representation R , a noisy measurement M is then generated. The conditional probability distribution $\pi(M|R)$ from which M is drawn is called the “noise function” of the system. From data consisting of N signal-measurement pairs, $\{S_n, M_n\}_{n=1}^N$, we wish to reconstruct the filter θ . This paper focuses on how to infer θ properly in the $N \rightarrow \infty$ limit when the noise function π is unknown *a priori*.

All statistical regression problems have this “SRM” form (Bishop, 2006), but we will focus on two biological applications for which this problem is particularly relevant. In neuroscience, SRM experiments are commonly used to characterize the response of neurons to stimuli (Schwartz et al., 2006). For instance, S may be an image to which a retina is exposed, while M is a binary variable (‘spike’ or ‘no spike’) indicating the response of a single retinal ganglion cell. It is often assumed that the spiking probability depends on a linear projection R of S . The specific probability of a spike given R is determined by the noise function π .

More recently, analogous experiments have been used to characterize the biophysical mechanisms of transcriptional regulation. In the context of work by Kinney et al. (2010), S is the DNA sequence of a transcriptional regulatory region, R is the rate of mRNA transcription produced by this sequence, and M is a (noisy) measurement of the resulting level of gene expression. The filter θ is a function of DNA sequence that reflects the underlying molecular mechanisms of transcript initiation. The noise function π accounts for both biological noise¹ and instrument noise.

¹Such as stochastic gene expression (Elowitz et al., 2002).

The standard approach for solving inference problems like these is to adopt a specific noise function π , then search a space Θ of possible filters for the one filter θ that maximizes the likelihood $p(\{M_n\}|\{S_n\}, \theta, \pi) = e^{NL(\theta, \pi)}$, where,

$$L(\theta, \pi) = \frac{1}{N} \sum_{n=1}^N \log \pi(M_n|\theta(S_n)), \quad (2)$$

is the per-datum log likelihood. For instance, the method of least squares regression corresponds to maximum likelihood inference assuming a homogenous Gaussian noise function π (Bishop, 2006).

Although the correct filter θ does indeed maximize $L(\theta, \pi)$ when the correct noise function π is used, full *a priori* knowledge of this noise function is rare in practice. Often π is chosen primarily for computational convenience, as is standard with least-squares regression. This can be problematic because using an incorrect π will *typically* produce bias in the inferred filter θ , bias that does not disappear in the $N \rightarrow \infty$ limit. The reason for this is illustrated in Fig. 1.

Sometimes this problem can be partially alleviated by performing a separate “calibration experiment” in which the noise function $\pi(M|R)$ is measured directly. For instance, one might be able to make repeated measurements M for a select number of known representations R . However, there will always be residual measurement error in π that will propagate to θ in a manner that is not properly accounted for by simply plugging π into likelihood calculations via Eq. 2.

An alternative inference procedure (Sharpee et al., 2004; Paninski, 2003; Kinney et al., 2007) that circumvents the need for an assumed noise function is to maximize the

mutual information (Cover & Thomas, 1991),

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

between predictions R and measurements M .² Here, $p(R, M)$ is the empirical joint distribution between predictions and measurements, and thus depends implicitly on θ . This method has been proposed, studied, and applied in the specific contexts of receptive field inference (Sharpee et al., 2004; Paninski, 2003; Sharpee et al., 2006; Pillow & Simoncelli, 2006) and transcriptional regulation (Kinney et al., 2007; Elemento et al., 2007; Kinney, 2008; Kinney et al., 2010; Melnikov et al., 2012). However, this alternative approach can be applied to a much wider range of statistical regression problems, and a general discussion of how maximizing mutual information relates to maximizing likelihood for arbitrary SRM systems has yet to be presented.

We begin by pointing out that, in the $N \rightarrow \infty$ limit, maximizing mutual information over θ alone is equivalent to maximizing likelihood over *both* θ and π . We then prove that when the correct filter θ lies within the class of filters being considered, maximizing mutual information is also equivalent to simultaneously maximizing every dependence measure that satisfies the Data Processing Inequality (DPI). However, in the absence of a known noise function π , SRM experiments are fundamentally incapable of constraining certain directions in the parameter space of θ ; we call these directions “diffeomorphic modes.” An equation for diffeomorphic modes is described and then applied to filters having various functional forms. In particular, our analysis of a linear-nonlinear filter used by Kinney et al. (2010) to model transcriptional regulation demonstrates how model nonlinearities can eliminate diffeomorphic modes in

²The notation $I(\theta)$ and $I[R; M]$ will be used interchangeably.

useful and non-obvious ways. This has important consequences for biophysical studies of transcriptional regulation that use recently developed DNA-sequencing-based assays (Kinney et al., 2010; Melnikov et al., 2012).

Throughout this manuscript, R is used to implicitly denote the representation predicted by the filter θ for signal S , i.e. $R \equiv \theta(S)$. $\mathcal{D}(\theta) \equiv \mathcal{D}[R; M]$ is used to denote any DPI-satisfying dependence measure. Representations R are assumed to be multidimensional with components R^μ , and $\partial_\mu \equiv \partial/\partial R^\mu$. θ is used to denote both a filter and the parameters governing that filter. Θ represents both an abstract space of filters, as well as the space of parameters for filters assumed to have a specific functional form. In the latter case, θ^i denotes coordinates in parameter space, and $\partial_i \equiv \partial/\partial \theta^i$.

2 Mutual information and likelihood

We begin by discussing the connection between likelihood and mutual information in the $N \rightarrow \infty$ limit. In this limit, the per-datum log likelihood (Eq. 2) can be rewritten as,

$$L(\theta, \pi) = \int dR dM p(R, M) \log \pi(M|R) \quad (4)$$

$$= I(\theta) - D(\theta, \pi) - H[M]. \quad (5)$$

The first term, $I(\theta)$, is the mutual information between R and M (Eq. 3) and is independent of the noise function π . The second term,

$$D(\theta, \pi) = \int dR dM p(R, M) \log \frac{p(M|R)}{\pi(M|R)}, \quad (6)$$

is the Kullback-Leibler (KL) divergence between the empirical distribution $p(M|R)$, which results from the choice of θ , and the assumed noise function $\pi(M|R)$. The last

term, $H[M] = - \int dM p(M) \log p(M)$, is the entropy of the measurements M . $H[M]$ is independent of both θ and π and can thus be ignored in the optimization problem.

The key point is that finding maximally informative filters θ is equivalent to solving the maximum likelihood problem over *both* filters θ and noise functions π . This is because if θ maximizes $I(\theta)$, simply choosing a noise function that matches the empirical noise function, i.e. setting $\pi(M|R) = p(M|R)$, will minimize $D(\theta, \pi)$ and thus maximize $L(\theta, \pi)$.

If one can formalize prior assumptions about the noise function π using a Bayesian prior $p(\pi)$, the relevant objective function becomes the per-datum marginal likelihood,

$$L_m(\theta) \equiv \frac{1}{N} \log \int d\pi p(\pi) p(\{M_n\} | \{S_n\}, \theta, \pi). \quad (7)$$

This is analogous to Eq. 4 computed after all possible noise functions have been integrated out. As has been shown in previous work (Kinney et al., 2007; Rajan et al., 2013), maximizing marginal likelihood and maximizing mutual information are essentially equivalent in the $N \rightarrow \infty$ limit. This can be seen by decomposing $L_m(\theta)$ as,

$$L_m(\theta) = I(\theta) - \Delta(\theta) - H[M], \quad (8)$$

where,

$$\Delta(\theta) \equiv -\frac{1}{N} \log \left[\int d\pi p(\pi) e^{-ND(\theta, \pi)} \right]. \quad (9)$$

Under weak assumptions about the prior $p(\pi)$,³ $\Delta \rightarrow 0$ as $N \rightarrow \infty$ (see Appendix A).

³E.g. $p(\pi)$ does not vanish at the true noise function π^* .

3 DPI-optimal filters

Mutual information is just one measure among many that satisfy DPI (see Appendix B). In this section, we discuss the importance of DPI for the SRM inference problem and introduce the notion of “DPI-optimal” filters.

Paninski (2003) has argued as follows for using DPI-satisfying dependence measures as objective functions for inferring filters. If θ^* is the correct filter in an SRM experiment, then for every filter θ ,

$$R \xleftarrow{\theta} S \xrightarrow{\theta^*} R^* \xrightarrow{\pi^*} M, \quad (10)$$

is a Markov chain. This implies $\mathcal{D}(\theta) \leq \mathcal{D}(\theta^*)$ for every DPI-satisfying measure \mathcal{D} . If θ^* resides within the space Θ of filters being explored, it must therefore fall within the subset of $\Theta_{\mathcal{D}} \subseteq \Theta$ on which \mathcal{D} is maximized. As a simple extension of this argument, we point out that, because θ^* maximizes *all* DPI-satisfying measures, θ^* must actually lie within the intersection of all such sets, i.e.,

$$\theta^* \in \Theta_{\text{DPI}} \equiv \bigcap_{\mathcal{D} \text{ satisfying DPI}} \Theta_{\mathcal{D}} \quad (11)$$

Filters in Θ_{DPI} can properly be said to be “DPI-optimal.”

This raises an important question: would optimizing a variety of different measures \mathcal{D} , not just mutual information, narrow the search for θ^* ? Here we show the answer is ‘no’; when $\theta^* \in \Theta$, maximizing mutual information is equivalent to simultaneously maximizing every DPI-satisfying measure, i.e.,

$$\Theta_I = \Theta_{\text{DPI}}. \quad (12)$$

To prove this, we first define on the space of all possible filters a weak and strong partial ordering, as well as an equivalence relation. These mathematical structures are

a natural consequence of DPI. For any two filters θ_1 and θ_2 ,⁴ we write,

$$\text{weak ordering : } \theta_1 \leq \theta_2 \iff \mathcal{D}(\theta_1) \leq \mathcal{D}(\theta_2) \text{ for all } \mathcal{D}, \quad (13)$$

$$\text{strong ordering : } \theta_1 < \theta_2 \iff \theta_1 \leq \theta_2 \text{ but not } \theta_2 \leq \theta_1, \quad (14)$$

$$\text{equivalence : } \theta_1 \simeq \theta_2 \iff \theta_1 \leq \theta_2 \text{ and } \theta_2 \leq \theta_1. \quad (15)$$

Note that $\theta_1 \leq \theta_2$ if $R_1 \leftrightarrow R_2 \leftrightarrow M$ is a Markov chain. The set Θ_{DPI} of DPI-optimal filters is the supremum of Θ under this partial ordering. The equivalence $\Theta_I = \Theta_{\text{DPI}}$, which occurs when $\theta^* \in \Theta$, follows directly from the fact, proven in Appendix C, that $\theta < \theta^*$ implies $I(\theta) < I(\theta^*)$. We note that this is not true for all DPI-satisfying measures. For instance, the trivial measure $\mathcal{D} = 0$ satisfies DPI but reveals no information about whether a given θ resides in Θ_{DPI} . These results are illustrated in Fig. 2.

4 Diffeomorphic modes

Whether or not two filters θ_1 and θ_2 satisfy the above equivalence relation (Eq. 15) can depend on the true filter θ^* and on the specific noise function π^* of the SRM experiment. However, certain pairs of filters will satisfy $\theta_1 \simeq \theta_2$ under *all* SRM experiments. We will refer to such pairs of filters as being “information equivalent.” In Appendix D we prove that two filters are information equivalent if and only if their predicted representations are related by an invertible transformation.

As an objective function, mutual information is inherently incapable of distinguishing between information equivalent filters. In practice this means that selecting maximally informative filters from a parametrized set of filters can leave some directions in

⁴The subscripts 1 and 2 label two different filters, not two parameters of a single filter.

parameter space unconstrained. Here we term these directions “diffeomorphic modes.”

The diffeomorphic modes of linear filters have an important and well-recognized consequence in neuroscience: the technique of maximally informative dimensions can identify only the relevant subspace of signal space, not a specific basis within that subspace (Sharpee et al., 2004; Paninski, 2003; Pillow & Simoncelli, 2006). However, an interesting twist occurs in applications to transcriptional regulation. Here, linear filters are often used to model the sequence-dependent binding energies of proteins to DNA (Stormo, 2013). Any mechanistic hypothesis about how DNA-bound proteins interact with one another predicts that the transcription rate will depend on these binding energies in a *specific* nonlinear manner (Bintu et al., 2005; Stormo, 2013). Such up-front knowledge about the nonlinearities of linear-nonlinear filters can eliminate diffeomorphic modes of the underlying linear filters in useful and non-obvious ways (Kinney, 2008; Kinney et al., 2010).

4.1 An equation for diffeomorphic modes

Consider a filter θ , representing a point in Θ , whose parameters θ^i are infinitesimally transported along a vector field having components $g^i(\theta)$. This yields a new filter θ' with components $\theta'^i = \theta^i + \epsilon g^i(\theta)$. If the representation R predicted by θ for a specified signal S has components R^μ in representation space, these will be transformed to $R'^\mu = R^\mu + \epsilon \sum_i g^i(\theta) \partial_i R^\mu$.

If the vector field $g^i(\theta)$ represents a diffeomorphic mode of Θ , this transformation must be invertible, meaning the values $\sum_i g^i(\theta) \partial_i R^\mu$ cannot depend on S except through the value of R . This is a nontrivial condition because $\partial_i R$ can depend on the

underlying signal S in an arbitrary manner. However, if $\sum_i g^i(\theta) \partial_i R^\mu$ does indeed depend only on the value of R then,

$$\sum_i g^i(\theta) \partial_i R^\mu = h^\mu(R, \theta). \quad (16)$$

for some vector function $h^\mu(R, \theta)$. This is the equation that any diffeomorphic mode $g^i(\theta)$ must satisfy.

4.2 General linear filters

We now use Eq. 16 to derive the diffeomorphic modes of general linear filters. By definition, a linear filter θ yields a representation R that is a linear combination of signal “features” F_i^μ , i.e.,

$$R^\mu = \sum_i \theta^i F_i^\mu(S). \quad (17)$$

As is standard with regression problems (Bishop, 2006), the term “linear” describes how R depends on the parameters θ^i ; the features F_i^μ need not be linear functions of S .

To find the diffeomorphic modes of these filters, we apply the operator $\sum_i g^i(\theta) \partial_i$ to both sides of Eq. 17. Using Eq. 16 we then find $\sum_i g^i(\theta) \partial_i F_i^\mu(S) = h^\mu(\{\sum_j \theta^j F_j^\nu(S)\}, \theta)$. The left-hand side is linear in signal features, so unless something unusual happens,⁵ $h^\mu(R, \theta)$ must also be a linear function of R , i.e. have the form,

$$h^\mu(R, \theta) = a^\mu(\theta) + \sum_\nu b_\nu^\mu(\theta) R^\nu. \quad (18)$$

⁵E.g. if the various features $F_i^\mu(S)$ exhibit complicated interdependencies, either because of their functional form or because signals S are restricted to a particular subspace. We ignore such possibilities here.

The number of diffeomorphic modes is bounded above by the number of independent parameters on which h^μ depends (at each θ).⁶ For a general linear filter we see that there can be no more than $\dim(R)[\dim(R) + 1]$ diffeomorphic modes, which is the number of parameters a^μ and b_ν^μ in Eq. 18. This bound is independent of the number of signal features, i.e. the dimensionality of S . In particular, if R is a scalar, then $h = a + bR$. In this case we observe two diffeomorphic modes, corresponding to additive and multiplicative transformations of R .

4.3 A linear-nonlinear filter

Kinney et al. (2010) performed experiments probing the biophysical mechanism of transcriptional regulation at the *Escherichia coli lac* promoter (Fig. 3A). These experiments are of the SRM form where S is the DNA sequence of a mutated *lac* promoter, M is a measurement of the resulting gene expression, and the mRNA transcription rate T is the internal representation the system. Linear filters were used to model the binding energies Q and P of the two proteins CRP and RNAP. The specific parametric form used for these filters was,

$$Q = \sum_{bl} \theta_Q^{bl} S_{bl} + \theta_Q^0, \quad P = \sum_{bl} \theta_P^{bl} S_{bl} + \theta_P^0, \quad (19)$$

where b indexes the four possible bases (A,C,G,T), l indexes nucleotide positions within the 75 bp promoter DNA region, $S_{bl} = 1$ if base b occurs at position l and $S_{bl} = 0$

⁶Technically the number of diffeomorphic modes is the number of independent vector fields g^i that correspond to such transformations. However, here we consider only *proper* diffeomorphic modes, not gauge transformations; as in physics, we define gauge transformations to be vector fields g^i along which transformation of θ leaves all predicted representations invariant.

otherwise.⁷

Measurements M were taken for $\sim 5 \times 10^4$ mutant *lac* promoters S . These data were then used to fit a model for the sequence-dependent binding energy of CRP. This was done by maximizing $I[Q; M]$. Because of the diffeomorphic modes of Q , the parameters θ_Q^{bl} were inferred up to an unknown scale and the additive constant θ_Q^0 was left undetermined. This is shown in Fig. 3B. Analogous results were obtained for RNAP (Fig. 3C).

Next, a full thermodynamic model of transcriptional regulation was proposed and fit to the data. Based on the hypothesized biophysical mechanism, the transcription rate T was assumed to depend on S via,

$$T = \frac{1}{1 + R^{-1}} \quad \text{where} \quad R = e^{-P} \frac{1 + e^{-Q-\gamma}}{1 + e^{-Q}}. \quad (20)$$

This quantity R is called the “regulation factor” of the promoter (Bintu et al., 2005). Because R is an invertible function of T , it serves equally well as the representation of the SRM system. In the following analysis we work with R instead of T due to its simpler functional form.

When the parameters of the linear filters P and Q were *simultaneously* fit to data by maximizing $I[T; M]$ (or equivalently, maximizing $I[R; M]$), three of the four diffeomorphic modes described above were eliminated (Fig. 3D). Specifically, the overall scale of the parameters θ_Q^{bl} and θ_P^{bl} were fixed, allowing binding energy predictions for CRP and RNAP in physically meaningful units of $k_B T$. The parameter θ_Q^0 , corresponding to the intracellular concentration of CRP, was also fixed by the data. The only

⁷To fix the gauge freedoms of these filters, Kinney et al. (2010) adopted the convention that $\min_b \theta_Q^{bl} = \min_b \theta_P^{bl} = 0$ for all positions l .

diffeomorphic mode left unbroken was θ_R^0 , corresponding to the intracellular concentration of RNAP.

We now show how the nonlinearity in R was able to break three of the four diffeomorphic modes of P and Q . First observe that any diffeomorphic mode of a linear-nonlinear filter must also be a diffeomorphic mode of *each individual* linear filter if, as here, the linear filters are independent functions of S . This means any diffeomorphic mode g^i of the full thermodynamic model for R must satisfy,

$$\sum_i g^i \partial_i R = h = (a_P + b_P P) \partial_P R + (a_Q + b_Q Q) \partial_Q R + a_\gamma \partial_\gamma R, \quad (21)$$

for coefficients $a_P, b_P, a_Q, b_Q, a_\gamma$ which do not depend on S . Evaluating the right-hand side derivatives and substituting for P in terms of Q and R we find,

$$h = -R \left[a_P - b_P \log \left\{ \frac{R(1 + e^{-Q})}{1 + e^{-Q-\gamma}} \right\} - \frac{(a_Q + b_Q Q) e^{-Q} (1 - e^{-\gamma})}{(1 + e^{-Q-\gamma})(1 + e^{-Q})} + \frac{a_\gamma e^{-Q-\gamma}}{1 + e^{-Q-\gamma}} \right]. \quad (22)$$

For g^i to be a diffeomorphic mode, the right-hand side must be independent of S for fixed R . The terms dependent on Q must therefore vanish, rendering $b_P = a_Q = b_Q = a_\gamma = 0$.⁸ Any diffeomorphic modes g^i must therefore satisfy $\sum_i g^i \partial_i R = -a_P R$. Thus only one mode remains, corresponding to an additive shift in the binding energy P .

5 Discussion

Likelihood-based inference masks the fundamentally different ways in which data constrain the parameters that lie along diffeomorphic modes versus those that lie along nondiffeomorphic modes. Standard likelihood inference constrains all model parameters, including both diffeomorphic and nondiffeomorphic modes, with error bars that

⁸This assumes $\gamma \neq 0$, i.e. that CRP actually interacts with RNAP. Which is true.

scale as $N^{-1/2}$.⁹ These constraints will be consistent with the correct underlying filter θ^* when the correct noise function is used (Fig. 4A). However, use of an incorrect noise function will typically cause θ^* to fall outside the error bars inferred along *both* diffeomorphic and nondiffeomorphic modes (Fig. 4B).

This problem is rectified if we use a prior $p(\pi)$ that reflects our uncertainty about what the true noise function is. From Eq. 8 it can be seen that using the resulting marginal likelihood to compute a posterior distribution on θ will constrain diffeomorphic and nondiffeomorphic modes in fundamentally different ways (Fig. 4C). Nondiffeomorphic modes will be constrained by $I(\theta)$, which remains finite in the large N limit. This produces error bars on nondiffeomorphic modes comparable to those produced by likelihood when the correct noise function π^* is used. However, constraints along diffeomorphic modes will come only from Δ . Because Δ vanishes as N^{-1} ,¹⁰ diffeomorphic constraints become independent of N once N is sufficiently large.

Fortunately, one does not need to posit a specific prior probability over all possible noise functions in order to confidently infer filters from SRM data. Using mutual information as an objective function instead of likelihood, i.e. sampling filters according to $p(\theta|\text{data}) \sim e^{NI(\theta)}$, will constrain nondiffeomorphic modes the same way that marginal likelihood does while putting no constraints along diffeomorphic modes (Fig. 4D).

One might worry that a large fraction of filter parameters will be diffeomorphic, and that the analysis of SRM experiments will require an assumed noise function in order to obtain useful results even if doing so yields unreliable error bars. Such situations are

⁹In this discussion we ignore gauge parameters, which do not alter model predictions and are therefore non-identifiable.

¹⁰More precisely, given any direction i in filter space, $\partial_i^2 \Delta|_{\theta^*} \sim N^{-1}$ for N large enough.

conceivable, but in practice this is often not the case. We have shown that for linear filters, the number of diffeomorphic modes will typically not exceed $\dim(R)[\dim(R) + 1]$ regardless of how large $\dim(S)$ is. Some of these diffeomorphic modes may also be eliminated if these linear filters are combined using a nonlinearity of known functional form. Indeed, of the 204 independent parameters comprising the biophysical model of transcriptional regulation inferred by Kinney et al. (2010), only one was diffeomorphic.

A bigger concern, perhaps, is the practical difficulty of using mutual information as an objective function. Specifically, it remains unclear how to compute $I(\theta)$ rapidly and reliably enough to confidently sample from $p(\theta|\text{data}) \sim e^{NI(\theta)}$. Still, various methods for estimating mutual information are available (Khan et al., 2007; Panzeri et al., 2007), and the information optimization problem has been successfully implemented using a variety of techniques (Sharpee et al., 2004; Sharpee et al., 2006; Kinney et al., 2007, 2010; Melnikov et al., 2012). We believe the exciting applications of mutual-information-based inference provide compelling motivation for making progress on these practical issues.

6 Appendix A: marginal likelihood

In certain cases $\Delta(\theta)$ can be computed explicitly and thereby be shown to vanish (Kinney et al., 2007). More generally, when π is taken to be finite-dimensional, a saddle-point computation (valid for large N) gives $\Delta(\theta) \approx \frac{1}{2N} \text{Tr}[\log \partial\partial\tilde{D}] + \text{const.}$ 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 using $\pi(M|R) = p(M|R)$. 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 semiclassical computation in field theory akin to the density estimation problem studied by Bialek et al. (1996). If this field theory is properly formulated through an appropriate choice of $p(\pi)$, then $\Delta(\theta)$ may exhibit different decay behavior, but will still vanish as $N \rightarrow \infty$. See also Rajan et al. (2013).

7 Appendix B: DPI-satisfying measures

DPI is satisfied by all measures of the F -information form (Csiszár & Shields, 2004; Kinney & Atwal, 2013),

$$I_F[M; R] \equiv \int dR dM p(R)p(M)F\left(\frac{p(R, M)}{p(R)p(M)}\right), \quad (23)$$

where $F(x)$ is a convex function for $x \geq 0$. Mutual information corresponds to $F(x) = x \log x$ whereas $F(x) = (x^\alpha - 1)/(\alpha - 1)$ yields a more general ‘‘Rényi information’’ measure (Rényi, 1961) that reduces to mutual information when $\alpha = 1$. DPI-satisfying measures other than mutual information have been used for filter inference in a number of works, including Paninski (2003) and Kouh & Sharpee (2009). A discussion of the differences between DPI-satisfying measures and some non-DPI-satisfying measures can be found in (Kinney & Atwal, 2013).

8 Appendix C: DPI-optimality

Assume $\theta < \theta^*$ by Eq. 14. Because $R \leftrightarrow R^* \leftrightarrow M$ is a Markov chain, the KL divergence between $p(R^*|R, M)$ and $p(R^*|R)$ can be decomposed as $D(p(R^*|R, M)||p(R^*|R)) =$

$I[R^*; M] - I[R; M]$. If this quantity is zero, then $R^* \leftrightarrow R \leftrightarrow M$ is also Markov chain, implying $\theta^* \leq \theta$, a contradiction. This KL divergence must therefore be positive, i.e. $I(\theta) < I(\theta^*)$. So if $\theta^* \in \Theta_{\text{DPI}}$, then for every $\theta \notin \Theta_{\text{DPI}}$, $\theta \notin \Theta_I$ as well. This proves $\Theta_I = \Theta_{\text{DPI}}$.

9 Appendix D: information equivalence

First we observe that if θ_1 and θ_2 make isomorphic predictions then they are information equivalent. This is readily shown from the fact that $\mathcal{D}[R; M]$ is invariant under arbitrary invertible transformations of R (Kinney & Atwal, 2013). Next we show the converse: if θ_1 and θ_2 are information equivalent, the predictions R_1 and R_2 must be isomorphic. Here is the proof. If $\theta_1 \simeq \theta_2$, then $\mathcal{D}[R_1; M] = \mathcal{D}[R_2; M]$ for all \mathcal{D} , and in particular $I[R_1; M] = I[R_2; M]$. In Appendix C we showed that $I[R; M] = I[R^*; M]$ implies $R^* \leftrightarrow R \leftrightarrow M$ is a Markov chain. Imagining an SRM experiment in which $\theta^* = \theta_1$ and $\pi(M|R) = \delta(M - R)$, we find that $R_1 \leftrightarrow R_2 \leftrightarrow R_1$ is a Markov chain. This implies the mapping $R_2 \rightarrow R_1$ is one-to-one. Similarly, $R_1 \rightarrow R_2$ is one-to-one. R_1 and R_2 are therefore bijective.

Acknowledgments

We thank William Bialek, Curtis Callan, Bud Mishra, Swagatam Mukhopadhyay, Anand Murugan, Michael 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

- Bialek, W., Callan, C., & Strong, S. (1996). Field theories for learning probability distributions. *Phys. Rev. Lett.*, *77*(23), 4693–4697.
- Bintu, L., Buchler, N., Garcia, H., Gerland, U., Hwa, T., Kondev, J., & Phillips, R. (2005). Transcriptional regulation by the numbers: models. *Curr. Opin. Genet. Dev.*, *15*(2), 116–124.
- Bishop, C. (2006). *Pattern recognition and machine learning*. Springer, New York.
- Brenner, N., Strong, S. P., Koberle, R., Bialek, W., & de Ruyter van Steveninck, R. R. (2000). Synergy in a neural code. *Neural Comput.*, *12*(7), 1531–1552.
- Cover, T. & Thomas, J. (1991). *Elements of information theory*. John Wiley & Sons, New York.
- Csiszár, I. & Shields, P.C. (2004). *Information theory and statistics: a tutorial* now Publishers, Hanover MA.
- Elemento, O., Slonim, N., & Tavazoie, S. (2007). A universal framework for regulatory element discovery across all genomes and data types. *Mol. Cell*, *28*(2), 337–350.
- Elowitz, M. B., Levine, A. J., Siggia, E. D., Swain, P. S. (2002). Stochastic gene expression in a single cell. *Science*, *297*(5584), 1183–1186
- Khan, S., Bandyopadhyay, S., Ganguly, A., Saigal, S., Erickson III, D., Protopopescu, V., & Ostouchov, G. (2007). Relative performance of mutual information estimation

methods for quantifying the dependence among short and noisy data. *Phys. Rev. E*, 76(2), 026209.

Kinney, J. B., Tkačik, G., & Callan, C. G. (2007). Precise physical models of protein-DNA interaction from high-throughput data. *Proc. Natl. Acad. Sci. USA*, 104(2), 501–506.

Kinney, J. B. (2008). *Biophysical models of transcriptional regulation from sequence data*. PhD thesis, Princeton University.

Kinney, J.B., Murugan, A., Callan, C.G., & Cox, E. (2010). Using deep sequencing to characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc. Natl. Acad. Sci USA*, 107(20),9158–9163.

Kinney, J. B. & Atwal, G. S. (2013). Equitability, mutual information, and the maximal information coefficient *arXiv:1301.7745* [q-bio.QM].

Kouh, M. & Sharpee, T. (2009). Estimating linear-nonlinear models using Rényi divergences. *Network*, 20(2), 49–68.

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. (2012). Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.*, 30(3), 271–277.

Paninski, L. (2003). Convergence properties of three spike-triggered analysis techniques. *Network*, 14(3), 437–464.

- Panzeri, S., Senatore, R., Montemurro, M. A., & Petersen, R. S. (2007). Correcting for the sampling bias problem in spike train information measures. *J. Neurophysiol.*, 98(3), 1064–1072.
- Pillow, J. W. & Simoncelli, E. P. (2006). Dimensionality reduction in neural models: an information-theoretic generalization of spike-triggered average and covariance analysis. *J. Vis.*, 6(4), 414–428.
- Rajan, K., Marre, O., & Tkačik, G. (2013). Learning quadratic receptive fields from neural responses to natural stimuli. *Neural Comput.*, 25(7), 1661–1692.
- Rényi, A. (1961). On measures of entropy and information. *Proc. 4th Berkeley Symp. Math. Statist. and Prob. (Univ. of Calif. Press)*, 1, 547–561.
- Schwartz, O., Pillow, J., Rust, N., & Simoncelli, E. (2006). Spike-triggered neural characterization. *J. Vis.*, 6(4), 484–507.
- Sharpee, T., Rust, N., & Bialek, W. (2004). Analyzing neural responses to natural signals: maximally informative dimensions. *Neural Comput.*, 16(2), 223–250.
- Sharpee, T., Sugihara, H., Kurgansky, A., Rebrik, S., Stryker, M., & Miller, K. (2006). Adaptive filtering enhances information transmission in visual cortex. *Nature*, 439(7079), 936–942.
- Stormo, G. D. (2013). *Introduction to protein-DNA interactions: structure, thermodynamics, and bioinformatics* Cold Spring Harbor Laboratory Press, New York.

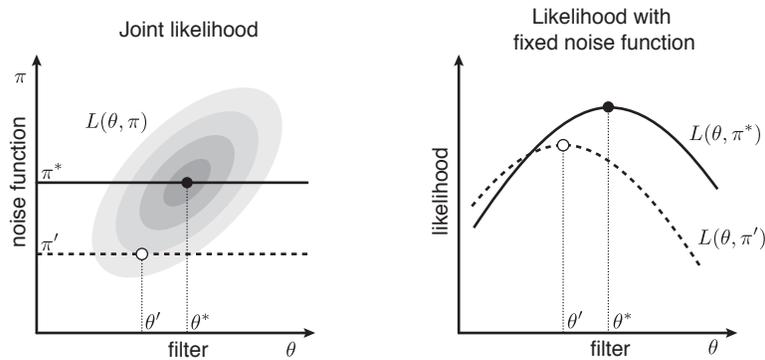


Figure 1.

Figure 1: Maximizing likelihood with an incorrect noise function will generally bias the inferred filter. The per-datum log likelihood $L(\theta, \pi)$ will typically depend on both the filter θ and the noise function π in a correlated manner (left panel). Values of a schematic $L(\theta, \pi)$ are illustrated in gray, with darker shades indicating larger likelihood. If the correct noise function π^* is assumed (solid line), maximizing $L(\theta, \pi^*)$ will yield the correct filter θ^* (filled dot). However, if an incorrect noise function π' is assumed (dashed line), maximizing $L(\theta, \pi')$ will typically lead to an incorrect filter θ' (open dot).

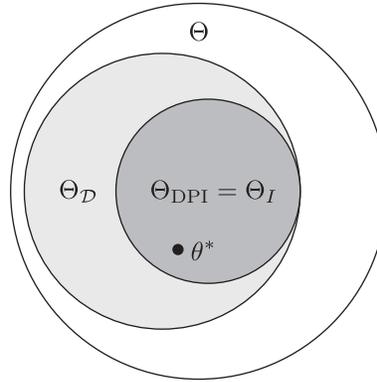


Figure 2.

Figure 2: Venn diagram illustrating filter sets maximizing different DPI-satisfying measures. In general, different DPI-satisfying dependence measures, e.g. mutual information I and some other measure \mathcal{D} , will be maximized by different sets of filters, respectively represented here by Θ_I and $\Theta_{\mathcal{D}}$. Θ_{DPI} is the intersection of the optimal sets of all such DPI-satisfying measures. Mutual information has the important property that $\Theta_I = \Theta_{\text{DPI}}$ whenever $\theta^* \in \Theta$; this is not true of all DPI-satisfying measures.

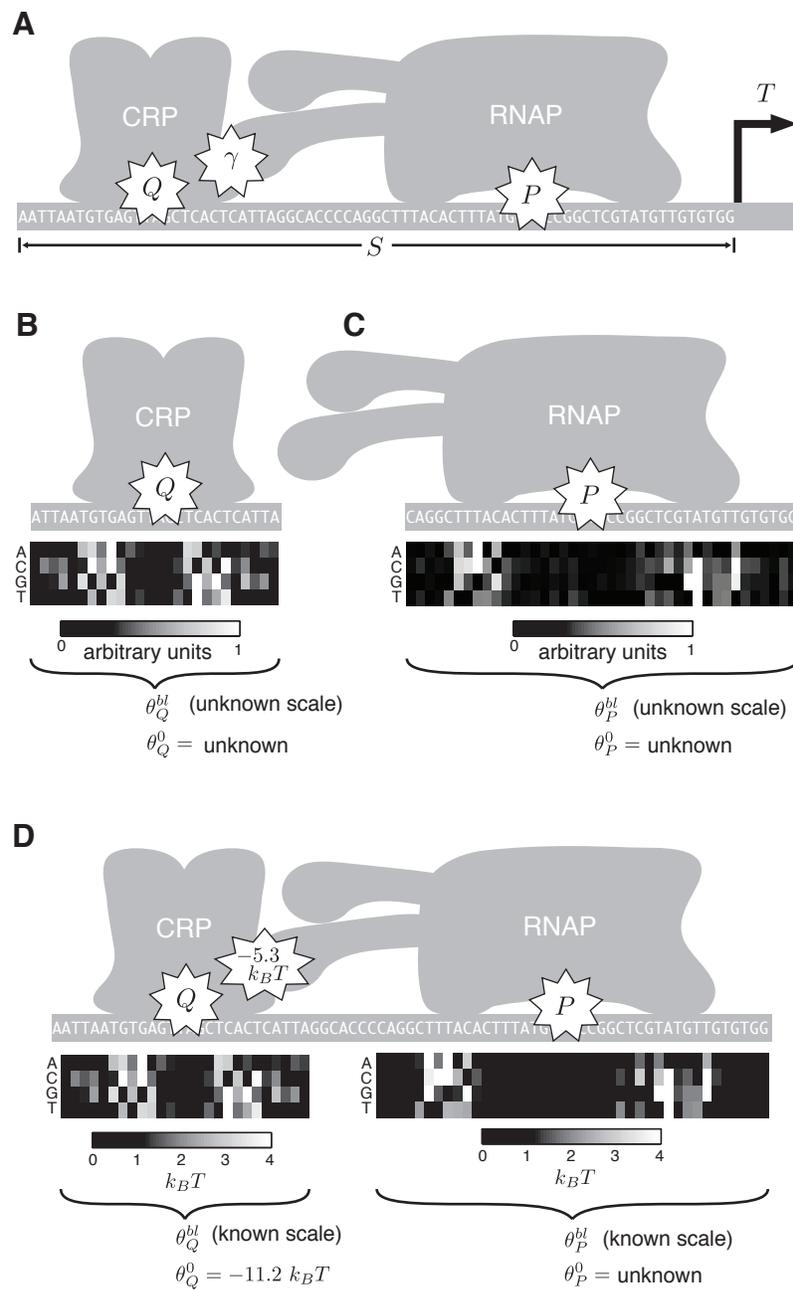


Figure 3.

Figure 3: A linear-nonlinear filter modeling the biophysics of transcriptional regulation at the *Escherichia coli lac* promoter. (A) The biophysical model inferred by Kinney et al. (2010) from Sort-Seq data. Each signal S is a 75 bp DNA sequence differing from the wildtype *lac* promoter by ~ 9 randomly scattered substitution mutations. Q and P denote the sequence-dependent binding energies of the proteins CRP and RNAP to their respective sites on this sequence S ; both Q and P were modeled as linear filters of S . γ is a sequence-independent interaction energy between CRP and RNAP. The resulting transcription rate T , of which the Sort-Seq assay produces noisy measurements M , is assumed to depend on Q , P , and γ in a specific nonlinear manner dictated by the hypothesized biophysical mechanism (Eq. 20; all energies are in units of $k_B T$). (B) The linear filter Q is defined by parameters θ_Q^{bl} and θ_Q^0 via Eq. 19. Inferring these parameters by maximizing the mutual information $I[Q; M]$ determines θ_Q^{bl} up to an unknown scale and leaves θ_Q^0 undetermined. (C) Analogous results are obtained for the parameters θ_P^{bl} and θ_P^0 when $I[P; M]$ is maximized. (D) Because of the inherent nonlinearity in Eq. 20 (right-hand side), maximizing $I[T; M]$ breaks diffeomorphic modes, fixing the values of θ_Q^{bl} , θ_P^{bl} , and θ_Q^0 in units of $k_B T$. The parameter θ_P^0 remains undetermined.

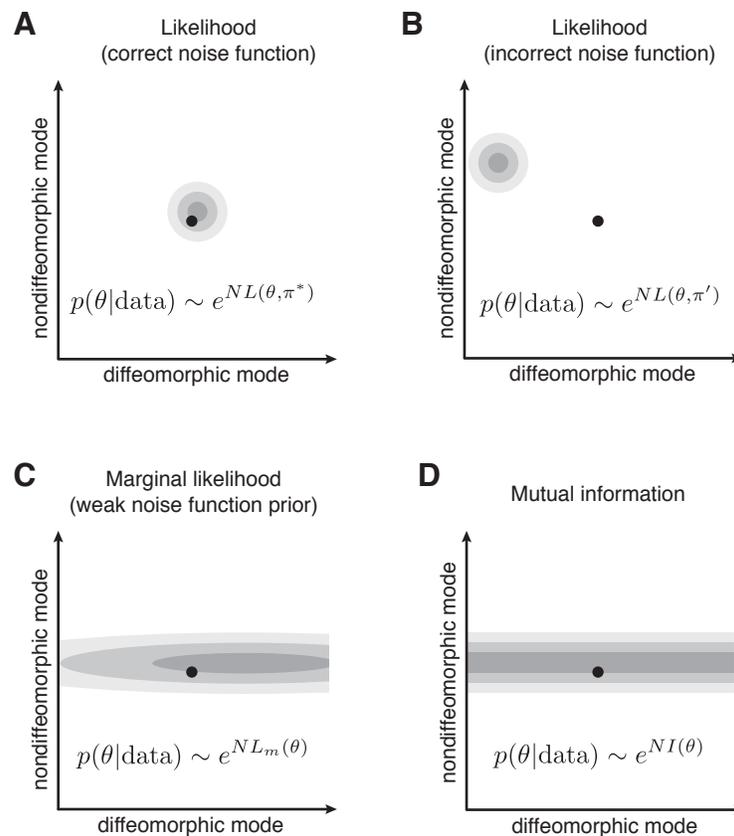


Figure 4.

Figure 4: Schematic illustration of constraints placed on diffeomorphic and non-diffeomorphic modes by different objective functions. The dot in each panel represents the correct filter θ^* ; shades of gray represent the posterior distribution $p(\theta|\text{data})$. (A,B) Likelihood (Eq. 2) places tight constraints (scaling as $N^{-1/2}$ as $N \rightarrow \infty$) along both diffeomorphic and nondiffeomorphic modes. (A) θ^* will typically lie within error bars if the correct noise function π^* is used. (B) However, if an incorrect noise function π' is used, θ^* will generally violate inferred constraints along both diffeomorphic and nondiffeomorphic modes. (C) Marginal likelihood (Eq. 7) computed using a sufficiently weak prior $p(\pi)$ will place tight constraints on nondiffeomorphic modes and weak constraints (scaling as N^0 as $N \rightarrow \infty$) along diffeomorphic modes. (D) Mutual information (Eq. 3)

places tight constraints on nondiffeomorphic modes but provides no constraints whatsoever on diffeomorphic modes.