High Dimensional Estimation and Data Analysis: Entropy and Regularized Regression

by

Vincent Quang Vu

B.A. (University of California, Berkeley) 2002
M.A. (University of California, Berkeley) 2005

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy

in

Statistics

in the

GRADUATE DIVISION
of the
UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:
Professor Bin Yu, Chair
Professor John Rice
Professor Jack L. Gallant

Spring 2009

The dissertation of Vincent Quang Vu is approved:

Chair Date

Date

Date

University of California, Berkeley

Spring 2009

High Dimensional Estimation and Data Analysis:

Entropy and Regularized Regression

by

Vincent Quang Vu

Abstract

High Dimensional Estimation and Data Analysis:

Entropy and Regularized Regression

by

Vincent Quang Vu

Doctor of Philosophy in Statistics

University of California, Berkeley

Professor Bin Yu, Chair

High-dimensional data are a frequent occurrence in many areas of application of statistics. For example, the analysis of data from neuroscience often involves fundamentally high-dimensional variables such as natural images or patterns of spiking in neural spike trains. These applications are often concerned with the relationship between these variables and another variate. What is the strength of the relationship? What is the nature of the relationship? This work is concerned with some of the statistical challenges in high-dimensional data analysis that arise when answering these questions; it is grounded in applications to data problems in neuroscience, and examines some challenges in entropy estimation and regularized regression that arise there.

Professor Bin Yu
Dissertation Committee Chair

To my family

Contents

Acknowledgments

Much of the work presented in this dissertation comes from collaborations that I have had over the course of my Ph.D with: Bin Yu, Rob Kass, Frederic Theunissen, Jack Gallant, Pradeep Ravikumar, Thomas Naselaris, and Kendrick Kay. I thank everyone. This would not be possible without your help.

David Aldous provided me with a wonderful research opportunity while I was an undergraduate in the Statistics Department. I am grateful for having been given that opportunity. It opened my eyes and was the catalyst that led me to the graduate program.

When I entered the graduate program, I was assigned (by chance?) to an office on the fourth floor of Evans Hall where Bin Yu also had an office. (She was in a northwest corner office at the time.) We frequently crossed paths in the hallway near the rear elevators. Her persistence and gregariousness made it difficult for me not to agree to work together on a project. One project led to many. I am very grateful for her patience and teaching. Thank you.

Chapter 1

Introduction

High-dimensional data problems are a frequent occurrence in many areas of application of statistics. For example, the analysis of data from neuroscience often involves fundamentally high-dimensional variables such as natural images or patterns of spiking in neural spike trains. The problem is high-dimensional when a variable lives in a high-dimensional space, say $\mathbb{R}^p$, but the number of observations, n, is of the same or a smaller order of magnitude, i.e. $p \gg n$. This poses many challenges for statistical estimation. Many applied problems are concerned with the relationship between variables. There are two natural questions in the investigation:

- What is the strength of the relationship?
- What is the nature of the relationship?

This work is concerned with some of the methodological and theoretical challenges in high-dimensional data analysis that arise when answering these questions; it is grounded in applications to data problems in neuroscience.

The first question can be answered within the framework of measuring statistical dependence. Entropy and mutual information are general measures of statistical variability and dependence. They originated in the work of Shannon (1948), where he proposed their use in his mathematical theory of communication systems. There entropy and mutual information have concrete meaning in the engineering domain in terms of data compression and transmission. However, they have also found application in a variety of areas outside of engineering such as ecology and neuroscience. In such applications, these quantities are usually calculated from data. The application of these measures to data analysis problems involves the fundamental problem of statistical estimation. Part I (Chapters 2 and 3) of this dissertation deals with this problem at a general theoretical level and also in the context of analyzing neuronal data.

Chapter 2 is concerned with the general problem of non-parametric entropy estimation in a high-dimensional setting motivated by entropy calculations for neural spike trains. The idealized setup in that chapter is useful for understanding the general difficulty of entropy estimation and also understanding when and why certain methods should work. In practice, the situation is never ideal. In particular when the data is collected from an experiment where the phenomenon of interest is fundamentally of dynamic, time-varying nature, the issue of stationary versus non-stationary becomes very important. Since mutual information is a difference of entropies, the results in that chapter are also applicable to mutual information estimation. Chapter 3 specifically examines mutual information estimates in the context of time dependent experiments that are common in neuroscience. There the meaning of the estimate changes depending on whether or not there is stationarity.

Regression is a natural framework for answering the second question, "What is the nature of the relationship?" In the most basic case, the problem is to estimate the conditional mean function $\mathbb{E}(Y|X = x)$. The natural interpretation is that it provides the best prediction of the response Y given the predictor $X = x$, in the least squared error sense. At the coarsest level, the different methodology differ in the basic assumptions about the nature of the conditional mean function. Even with the strictest assumption that $\mathbb{E}(Y|X = x)$ is a linear function of x, regression in the high-dimensional setting can be very difficult. The problem is ill-posed when the dimension of X is comparable to or exceeds the sample size—regression must be regularized for any hope of success.

Part II (Chapters 4 and 5) addresses some specific aspects of regularized regression in high dimensions. Chapter 4 describes state-of-the-art results in a long investigation into neural coding in area V1 of the visual cortex of the human brain. There we describe a progression of technique that begins with regularized linear models and culminates with non-linear sparse models in predicting V1 fMRI response to novel natural image stimuli. The goal of it all is to answer the question: what is the nature of the relationship between natural image stimuli and V1 functional MRI (magnetic resonance imaging) response?

Chapter 5 contains some mathematical theory on the use of generalized ridge regression for prediction. The chapter attempts to address the problem of specification of the penalty/constraint (or prior for Bayesians) in generalized ridge regression. This problem was motivated by preliminary work on the data analysis problem in Chapter 4, where linear models were fit using a variant of ridge regression known as power ridge regression (Hoerl and Kennard, 1970; Goldstein and Smith, 1974). In preliminary investigations it

was found that the choice of the power parameter q in power ridge had a drastic effect on prediction performance. Chapter 5 abstracts the problem and presents a general framework for understanding the effect of penalty/constraint misspecification. The results are very mathematical, and further investigation of the results with concrete examples is planned.

Part I

Entropy

Chapter 2

Coverage Adjusted Entropy Estimation

2.1 Introduction

The problem of "neural coding" is to elucidate the representation and transformation of information in the nervous system (Perkel and Bullock, 1968). An appealing way to attack neural coding is to take the otherwise vague notion of "information" to be defined in Shannon's sense, in terms of entropy (Shannon, 1948). This project began in the early days of cybernetics (Wiener, 1948; MacKay and McCulloch, 1952), received considerable impetus from work summarized in the book *Spikes: Exploring the Neural Code* (Rieke et al., 1997), and continues to be advanced by many investigators. In most of this research, the findings concern the mutual information between a stimulus and a neuronal spike train response. For a succinct overview see (Borst and Theunissen, 1999). The mutual information, how-

ever, is the difference of marginal and expected conditional entropies; to compute it from data one is faced with the basic statistical problem of estimating the entropy[1]

$$H := -\sum_{x \in \mathcal{X}} P(x) \log P(x) \tag{2.1.1}$$

of an unknown discrete probability distribution P over a possibly infinite space $\mathcal{X}$, the data being conceived as random variables $X_1, \ldots, X_n$ with X_i distributed according to P. An apparent method of estimating the entropy is to apply the formula after estimating $P(x)$ for all $x \in \mathcal{X}$, but estimating a discrete probability distribution is, in general, a difficult nonparametric problem.

2.2 Background

In linguistic applications, $\mathcal{X}$ could be the set of words in a language, with P specifying their frequency of occurrence. For neuronal data, X_i often represents the number of spikes (action potentials) occurring during the ith time bin. Alternatively, when a fine resolution of time is used (such as $dt = 1$ millisecond), the occurrence of spikes is indicated by a binary sequence, and X_i becomes the pattern, or "word," made up of 0-1 words or "letters," for the ith word. This is described in Figure 2.1, and it is the basis for the widely-used "direct method" proposed by Strong et al. (1998). The number of possible words $m := |\{x \in \mathcal{X} : P(x) > 0\}|$ is usually unknown and possibly infinite. In the example in Figure 2.1, the maximum number of words is the total number of 0-1 strings of length L. For $L = 10$ this number is 1024; for $L = 20$ it is well over one million, and in general there is an exponential explosion with increasing L. Furthermore, the phenomenon under

[1] Unless otherwise stated, we take all logarithms to be base 2 and define $0 \log 0 = 0$.

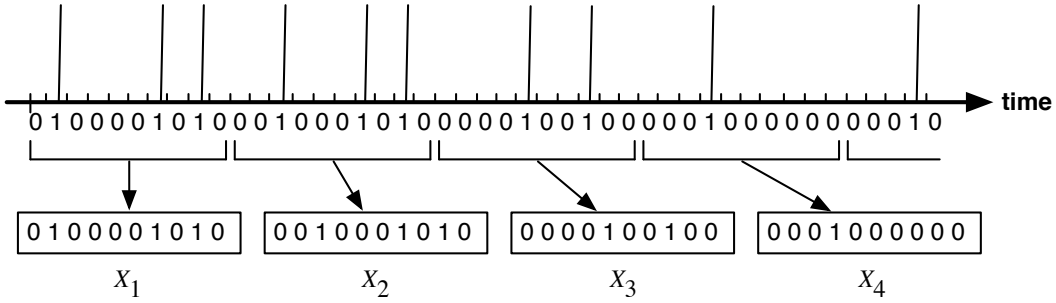


Figure 2.1: The top row depicts 45 milliseconds of a hypothetical spike train. The ticks on the time axis demarcate $dt = 1$ millisecond bins (intervals). The spike train is discretized into a sequence of counts. Each count is the number of spikes that fall within a single time bin. Subdividing this sequence into words of length $L = 10$ leads to the words shown at the bottom. The words $X_1, X_2, \ldots$ take values in the space $\mathcal{X} = \{0, 1\}^{10}$ consisting of all 0-1 strings of length 10.

investigation will often involve fine time resolution, necessitating a small bin size dt and thus a large L. For large L, the estimation of $P(x)$ is likely to be challenging.

We note that Strong et al. calculated the entropy rate. Let $\{W_t : t = 1, 2, \ldots\}$ be a discretized (according to dt) spike train as in the example in Figure 2.1. If $\{W_t\}$ is a stationary process, the entropy of a word, say $X_1 = (W_1, \ldots, W_L)$, divided by its length L is non-increasing in L and has a limit as $L \to \infty$, i.e.

$$\lim_{L\to\infty} \frac{1}{L} H(X_1) = \lim_{L\to\infty} \frac{1}{L} H(W_1, \ldots, W_L) =: H' \tag{2.2.1}$$

exists (Cover and Thomas, 1991). This is the entropy rate of $\{W_t\}$. The word entropy is used to estimate the entropy rate. If $\{W_t\}$ has finite range dependence, then the above entropy factors into a sum of conditional entropies and a single marginal entropy. Generally, the word length is chosen to be large enough so that $H(W_1, \ldots, W_L)/L$ is a close approximation to H', but not so large that there are not enough words to estimate $H(W_1, \ldots, W_L)$. Strong

et al. proposed that the entropy rate estimate be extrapolated from estimates of the word entropy over a range of word lengths. We do not address this extrapolation, but rather focus on the problem of estimating the entropy of a word.

In the most basic case the observations $X_1, \ldots, X_n$ are assumed to be independent and identically distributed (i.i.d.). Without loss of generality, we assume that $\mathcal{X} \subseteq \mathbb{N}$ and that the *words* [2] are labeled $1, 2, \ldots$. The seemingly most natural estimate is the *empirical plug-in* estimator

$$\hat{H} := -\sum_x \hat{P}(x) \log \hat{P}(x), \tag{2.2.2}$$

which replaces the unknown probabilities in Equation (2.1.1) with the empirical probabilities $\hat{P}(x) := n_x/n$, that is the observed proportion n_x/n of occurrences of the word x in $X_1, \ldots, X_n$. The empirical plug-in estimator is often called the "naive" estimate or the "MLE"—after the fact that $\hat{P}$ is the maximum likelihood estimate of P. We will use "MLE" and "empirical plug-in" interchangeably. From Jensen's Inequality it is readily seen that the MLE is negatively biased unless P is trivial. In fact no unbiased estimate of entropy exists (see Paninski (2003) for an easy proof).

In the finite m case, Basharin (1959) showed that the MLE is biased, consistent, and asymptotically normal with variance equal to the *entropy variance* $\mathrm{Var}[\log P(X_1)]$. Miller (1955) previously studied the bias independently and provided the formula

$$\mathbb{E}\hat{H} - H = -\frac{m-1}{2n} + \mathcal{O}\left(1/n^2\right). \tag{2.2.3}$$

The bias dominates the mean squared error of the estimator (Antos and Kontoyiannis,

[2] The information theory literature traditionally refers to $\mathcal{X}$ as an *alphabet* and its elements as *symbols*. It is natural to call a tuple of symbols a word, but the problem of estimating the entropy of the L-tuple word reduces to that of estimating the entropy in an enlarged space (of L-tuples).

2001), and has been the focus of recent studies (Victor, 2000; Paninski, 2003).

The original "direct method" advocated an ad-hoc strategy of bias reduction based on a subsampling extrapolation (Strong et al., 1998). A more principled correction based on the jackknife technique was proposed earlier by Zahl (1977). The formula Equation (2.2.3) suggests a bias correction of adding $(m-1)/(2n)$ to the MLE. This is known as the Miller–Maddow correction. Unfortunately, it is an asymptotic correction that depends on the unknown parameter m. Paninski (2003) observed that both the MLE and Miller–Maddow estimates fall into a class of estimators that are linear in the frequencies of observed word counts $f_j = |\{n_x : n_x = j\}|$. He proposed an estimate, "Best Upper Bounds" (BUB), based on numerically minimizing an upper-bound on the bias and variance of such estimates when m is assumed finite and known. We note that in the case that m is unknown, it can be replaced by an upper-bound, but the performance of the estimator is degraded.

Bayesian estimators have also been proposed for the finite m case by Wolpert and Wolf (1995). Their approach is to compute the posterior distribution of entropy based on a symmetric Dirichlet prior on P. Nemenman et al. (2004) found that the Dirichlet prior on P induces a highly concentrated prior on entropy. They argued that this property is undesirable and proposed an estimator based on a Dirichlet mixture prior with the goal of flattening the induced prior distribution on entropy. Their estimate requires a numerical integration and also the unknown parameter m, or at least an upper-bound. The estimation of m is even more difficult than the estimation of entropy (Antos and Kontoyiannis, 2001), because it corresponds to estimating $\lim_{a \downarrow 0} \sum_x [P(x)]^a$.

In the infinite m case, Antos and Kontoyiannis (2001) proved consistency of the

empirical plug-in estimator and showed that there is no universal rate of convergence for any estimator. However, Wyner and Foster (2003) have shown that the best rate (to first order) for the class of distributions with with finite *entropy variance* or equivalently finite log-likelihood second moment

$$\sum_x P(x)(\log P(x))^2 < \infty$$

is $\mathcal{O}_{\mathrm{P}}(1/\log n)$. This rate is achieved by the empirical plug-in estimate as well as an estimator based on match lengths. Despite the fact that the empirical plug-in estimator is asymptotically optimal, its finite sample performance leaves much to be desired.

Chao and Shen (2003) proposed a coverage adjusted entropy estimator intended for the case when there are potentially unseen words in the sample. This is always the case when m is relatively large or infinite. Intuitively, low probability words are typically absent from most sequences, i.e. the *expected sample coverage* is < 1, but in total, the missing words can have a large contribution to H. The estimator is based on plug-in of a coverage adjusted version of the empirical probability into the Horvitz–Thompson (Horvitz and Thompson, 1952) estimator of a population total. They presented simulation results showing that the estimator seemed to perform quite well, especially in the small sample size regime, when compared to the usual empirical plug-in and several bias corrected variants. The estimator does not require knowledge of m, but they assumed a finite m. We prove here (Theorem 2.2) that the coverage adjusted estimator also works in the infinite m case. Chao and Shen also provided approximate confidence intervals for the coverage adjusted estimate, however they are asymptotic and depend on the assumption of finite m.

The problems of entropy estimation and estimation of the distribution P are dis-

tinct. Entropy estimation should be no harder than estimation of P, since H is a functional of P. However, several of the entropy estimators considered here depend either implicitly or explicitly on estimating P. BUB is linear in the frequency of observed word counts f_j, and those are 1-to-1 with the empirical distribution $\hat{P}$ up to labeling. In general, any symmetric estimator is a function of $\hat{P}$. The only estimators mentioned above that does not depend on $\hat{P}$ is the match length estimator. For the coverage adjusted estimator, the dependence on estimating P is only through estimating $P(k)$ for observed words k.

2.3 Theory

Unobserved words—those that do not appear in the sample, but have non-zero probability–can have a great impact on entropy estimation. However, these effects can be mitigated with two types of corrections: Horvitz–Thompson adjustment and coverage adjustment of the probability estimate. Section 2.3.1 contains an exposition of some of these effects. The adjustments are described in Section 2.3.2 along with the definition of the resulting coverage adjusted entropy estimator. A key ingredient of the estimator is a coverage adjusted probability estimate. We provide a novel derivation from the viewpoint of regularization in Section 2.3.3. Lastly, Section 2.3.4 concludes the theoretical study with our rate of convergence results.

Throughout this section we assume that $X_1, \ldots, X_n$ is an i.i.d. sequence from the distribution P on the countable set $\mathcal{X}$. Without loss of generality, we assume that the $P(k) > 0$ for all $k \in \mathcal{X}$ and write p_k for $P(k) = \mathbb{P}(X_i = k)$. As before, $m := |\mathcal{X}|$ and

possibly $m = \infty$. Let

$$n_k := \sum_{i=1}^{n} 1_{\{X_i = k\}}$$

be the number of times that the word k appears in the sequence $X_1, \ldots, X_n$.

2.3.1 The Unobserved Word Problem

The set of observed words S is the set of words that appear at least once in the sequence $X_1, \ldots, X_n$, i.e.

$$S := \{k : n_k > 0\}.$$

The complement of S, i.e. $\mathcal{X} \backslash S$, is the set of unobserved words. There is always a non-zero probability of unobserved words, and if $m > n$ or $m = \infty$ then there are always unobserved words. In this section we describe two effects of the unobserved words pertaining to entropy estimation.

Given the set of observed words S, the entropy of P can be written as the sum of two parts:

$$H = -\sum_{k \in S} p_k \log p_k - \sum_{k \notin S} p_k \log p_k. \tag{2.3.1}$$

One part is the contribution of observed words; the other is the contribution of unobserved words. Suppose for a moment that p_k is known exactly for $k \in S$, but unknown for $k \notin S$. Then we could try to estimate the entropy by

$$-\sum_{k \in S} p_k \log p_k, \tag{2.3.2}$$

but there would be an error in the estimate unless the *sample coverage*

$$C := \sum_{k \in S} p_k$$

is identically 1. The error is due to the contribution of unobserved words and thus the unobserved summands:

$$-\sum_{k \notin S} p_k \log p_k.$$

This error could be far from negligible, and its size depends on the p_k for $k \notin S$. However, there is an adjustment that can be made so that the adjusted version of Equation (2.3.2) is an unbiased estimate of H. This adjustment comes from the Horvitz–Thompson estimate of a population total, and we will review it in Section 2.3.2.

Unfortunately, p_k is unknown for both $k \in S$ and $k \notin S$. A common estimate for p_k is the MLE/empirical $\hat{p}_k := n_k/n$. Plugging this estimate into Equation (2.3.2) gives the MLE/empirical plug-in estimate of entropy:

$$\hat{H} := -\sum_{k} \hat{p}_k \log \hat{p}_k = -\sum_{k \in S} \hat{p}_k \log \hat{p}_k,$$

because $\hat{p}_k = 0$ for all $k \notin S$. If the sample coverage C is < 1, then this is a degenerate estimate because $\sum_{k \in S} \hat{p}_k = 1$ and so $\hat{p}_k = 0$ for all $k \notin S$. Thus, we could shrink the estimate of p_k on S toward zero so that its sum over S is < 1. This is the main idea behind the coverage adjusted probability estimate, however we will derive it from the viewpoint of regularization in Section 2.3.3.

We have just seen that unobserved words can have two negative effects on entropy estimation: unobserved summands and error-contaminated summands. The "size," or non-coverage, of the set of unobserved words can be measured by 1 minus the sample coverage:

$$1 - C = \sum_{k \notin S} p_k = \mathbb{P}(X_{n+1} \notin S | S).$$

Thus, it is also the conditional probability that a future observation X_{n+1} is not a previously

observed word. So the average non-coverage is

$$\mathbb{E}(1-C) = \mathbb{P}(X_{n+1} \notin S) = \sum_k p_k(1-p_k)^n.$$

and in general $\mathbb{E}(1-C) > 0$. Its rate of convergence to 0, as $n \to \infty$, depends on P and can be very slow. (See the corollary to Theorem 2.3 below). It is necessary to understand how to mitigate the effects of unobserved words on entropy estimation.

2.3.2 Coverage Adjusted Entropy Estimator

Chao and Shen (2003) observed that entropy can be thought of as the total $\sum_k y_k$ of an unknown population consisting of elements $y_k = -p_k \log p_k$. For the general problem of estimating a population total, the Horvitz–Thompson estimator,

$$\sum_{k \in S} \frac{y_k}{\mathbb{P}(k \in S)} = \sum_k \frac{y_k}{\mathbb{P}(k \in S)} 1_{\{k \in S\}}, \tag{2.3.3}$$

provides an unbiased estimate of $\sum_k y_k$, under the assumption that the inclusion probabilities $\mathbb{P}(k \in S)$ and y_k are known for $k \in S$. For the i.i.d. sequence $X_1, \ldots, X_n$ the probability that word k is unobserved in the sample is $(1-p_k)^n$. So the inclusion probability is $1-(1-p_k)^n$. Then the Horvitz–Thompson adjusted version of Equation (2.3.2) is

$$\sum_{k \in S} \frac{-p_k \log p_k}{1-(1-p_k)^n}.$$

All that remains is to estimate p_k for $k \in S$. The empirical $\hat{p}_k$ can be plugged into the above formula, however, as we stated in the previous section, it is a degenerate estimate when $C < 1$ because it assigns 0 probability to $k \notin S$ and, thus, tends to overestimates the inclusion probability. We will discuss this further in Section 2.3.3.

In a related problem, Ashbridge and Goudie (2000) considered finite populations with elements $y_k = 1$, so that Equation (2.3.3) becomes an estimate of the population size. They found that $\hat{P}$ did not work well and suggested using instead a coverage adjusted estimate $\tilde{P} := \hat{C}\hat{P}$, where $\hat{C}$ is an estimate of C. Chao and Shen recognized this and proposed using the Good–Turing coverage estimator (Good, 1953; Robbins, 1968):

$$\hat{C} := 1 - \frac{f_1}{n},$$

where $f_1 := \sum_k 1_{\{n_k=1\}}$ is the number of singletons in the sequence $X_1, \ldots, X_n$. This leads to the coverage adjusted entropy estimator:

$$\tilde{H} := -\sum_k \frac{\tilde{p}_k \log \tilde{p}_k}{1 - (1 - \tilde{p}_k)^n},$$

where $\tilde{p}_k := \hat{C}\hat{p}_k$. Chao and Shen gave an argument for $C\hat{P}$ based on a conditioning property of the multinomial distribution. In the next section we give a different derivation from the perspective of regularization of an empirical risk, and give upper-bounds for the bias and variance of $\hat{C}$.

2.3.3 Regularized Probability Estimation

Consider the problem of estimating P under the entropy loss $L(q, x) = -\log Q(x)$, for Q satisfying $Q(k) = q_k \geq 0$ and $\sum q_k = 1$. This loss function is closely aligned with the problem of entropy estimation because the risk, i.e. the expected loss on a future observation,

$$R(Q) := -\mathbb{E} \log Q(X_{n+1}) \tag{2.3.4}$$

is uniquely minimized by $Q = P$ and its optimal value is the entropy of P. The MLE $\hat{P}$ minimizes the empirical version of the risk

$$\hat{R}(Q) := -\frac{1}{n}\sum_{i=1}^{n} \log Q(X_i). \tag{2.3.5}$$

As stated previously in Section 2.3.1, this is a degenerate estimate when there are unobserved words. More precisely, if the expected coverage $\mathbb{E}C < 1$ (which is true in general), then $R(\hat{P}) = \infty$.

Analogously to Equation (2.3.1), the expectation in Equation (2.3.4) can be split into two parts by conditioning on whether X_{n+1} is a previously observed word or not:

$$\begin{aligned} R(Q) = &- \mathbb{E}[\log Q(X_{n+1}) | X_{n+1} \in S] \, \mathbb{P}(X_{n+1} \in S) \\ &- \mathbb{E}[\log Q(X_{n+1}) | X_{n+1} \notin S] \, \mathbb{P}(X_{n+1} \notin S). \end{aligned} \tag{2.3.6}$$

Since $\mathbb{P}(X_{n+1} \in S)$ does not depend on Q, minimizing Equation (2.3.6) with respect to Q is equivalent to minimizing

$$- \mathbb{E}[\log Q(X_{n+1}) | X_{n+1} \in S] - \lambda^* \mathbb{E}[\log Q(X_{n+1}) | X_{n+1} \notin S], \tag{2.3.7}$$

where $\lambda^* = \mathbb{P}(X_{n+1} \notin S) / \mathbb{P}(X_{n+1} \in S)$. We cannot distinguish the probabilities of the unobserved words on the basis of the sample. So consider estimates Q which place constant probability on $x \notin S$. Equivalently, these estimates treat the unobserved words as a single class and so the risk reduces to the equivalent form:

$$-\mathbb{E}[\log Q(X_{n+1}) | X_{n+1} \in S] - \lambda^* \mathbb{E} \log \left[1 - \sum_{k \in S} Q(k) \right].$$

The above expectations only involve evaluating Q at observed words. Thus, Equation (2.3.5) is more natural as an estimate of $-\mathbb{E}[\log Q(X_{n+1}) | X_{n+1} \in S]$, than as an estimate of $R(Q)$.

If we let λ be any estimate of the odds ratio $\lambda^* = \mathbb{P}(X_{n+1} \notin S)/\,\mathbb{P}(X_{n+1} \in S)$, then we arrive at the *regularized empirical risk*,

$$\tilde{R}(q;\lambda) := -\frac{1}{n}\sum_i \log Q(X_i) - \lambda \log\left[1 - \sum_i Q(X_i)\right]. \tag{2.3.8}$$

This is the usual empirical risk with an additional penalty on the total mass assigned to observed words. It can be verified that the minimizer, up to an equivalence, is $(1+\lambda)^{-1}\hat{P}$. This estimate shrinks the MLE towards 0 by the amount $(1+\lambda)^{-1}$. Any Q which agrees with $(1+\lambda)^{-1}\hat{P}$ on S is a minimizer of Equation (2.3.8). Note that $(1+\lambda^*)^{-1} = \mathbb{P}(X_{n+1} \in S) = \mathbb{E}C$ is the expected coverage, rather than the sample coverage C. $\hat{C}$ can be used to estimate both $\mathbb{E}C$ and C, however it is actually better as an estimate of $\mathbb{E}C$ because McAllester and Schapire (2000) have shown that $\hat{C} = C + \mathcal{O}_{\mathrm{P}}(\log n/\sqrt{n})$, whereas we prove in the appendix the following proposition.

Proposition 2.1.

$$0 \geq \mathbb{E}(\hat{C} - C) = -\sum_k p_k^2(1-p_k)^{n-1} \geq (1-1/n)^{n-1}/n \sim -e^{-1}/n$$

and $\operatorname{Var}\hat{C} \leq 4/n$.

So $\hat{C}$ is a $1/\sqrt{n}$ consistent estimate of $\mathbb{E}C$. Using $\hat{C}$ to estimate $\mathbb{E}C = (1+\lambda^*)^{-1}$, we obtain the coverage adjusted probability estimate $\tilde{P} = \hat{C}\hat{P}$.

2.3.4 Convergence Rates

In the infinite m case, Antos and Kontoyiannis (2001) proved that the MLE is universally consistent almost surely and in L^2, provided that the entropy exists. However, they also showed that there can be no universal rate of convergence for entropy estimation.

Some additional restriction must be made beyond the existence of entropy in order to obtain a rate of convergence. Wyner and Foster (2003) found that for the weakest natural restriction, $\sum_k p_k(\log p_k)^2 < \infty$, the best rate of convergence, to first order, is $\mathcal{O}_{\mathrm{P}}\left(1/\log n\right)$. They proved that the MLE and an estimator based on match lengths achieves this rate. Our main theoretical result is that the coverage adjusted estimator also achieves this rate.

Theorem 2.2. *Suppose that* $\sum_k p_k(\log p_k)^2 < \infty$. *Then as* $n \to \infty$,

$$\tilde{H} = H + \mathcal{O}_{\mathrm{P}}\left(1/\log n\right).$$

In the previous section we employed $\hat{C} = 1 - f_1/n$, in the regularized empirical risk Equation (2.3.8). As for the observed sample coverage, $C = \mathbb{P}(X_{n+1} \in S|S)$, McAllester and Schapire (2000) proved that $\hat{C} = \mathbb{P}(X_{n+1} \in S|S) + \mathcal{O}_{\mathrm{P}}\left(\log n/\sqrt{n}\right)$, regardless of the underlying distribution. Our theorem below together with that of McAllester and Schapire implies a rate of convergence on the total probability of unobserved words.

Theorem 2.3. *Suppose that* $\sum_k p_k|\log p_k|^q < \infty$. *Then as* $n \to \infty$, *almost surely,*

$$\hat{C} = 1 - \mathcal{O}\left(1/(\log n)^q\right).$$

Corollary 2.4. *Suppose that* $\sum_k p_k|\log p_k|^q < \infty$. *Then as* $n \to \infty$,

$$1 - C = \mathbb{P}(X_{n+1} \notin S|S) = \mathcal{O}_{\mathrm{P}}\left(1/(\log n)^q\right). \tag{2.3.9}$$

Proof. This follows from the above theorem and McAllester and Schapire (2000, Theorem 3) which implies $|\hat{C} - \mathbb{P}(X_{n+1} \in S|S)| \leq \mathrm{o}_{\mathrm{P}}\left(1/(\log n)^q\right)$ because

$$0 \leq \mathbb{P}(X_{n+1} \notin S|S) \leq |1 - \hat{C}| + |\hat{C} - \mathbb{P}(X_{n+1} \in S|S)|$$

and $\mathcal{O}_{\mathrm{P}}\left(1/(\log n)^q\right) + \mathrm{o}_{\mathrm{P}}\left(1/(\log n)^q\right) = \mathcal{O}_{\mathrm{P}}\left(1/(\log n)^q\right)$. ∎

The proofs of Theorem 2.2 and Theorem 2.3 are contained in Section 3.5. At the time of writing, the only other entropy estimators proved to be consistent and asymptotically first-order optimal in the finite entropy variance case that we are aware of are the MLE and Wyner and Foster's modified match length estimator. However, the $\mathcal{O}_{\mathrm{P}}\left(1/\log n\right)$ rate, despite being optimal, is somewhat discouraging. It says that in the worst case we will need an exponential number of samples to estimate the entropy. Furthermore, the asymptotics are unable to distinguish the coverage adjusted estimator from the MLE, which has been observed to be severely biased. In the next section we use simulations to study the small-sample performance of the coverage adjusted estimator and the MLE, along with other estimators. The results suggest that in this regime their performances are quite different.

2.4 Simulation Study

We conducted a large number of simulations under varying conditions to investigate the performance of the coverage adjusted estimator (CAE) and compare with four other estimators.

- Empirical Plug-in (MLE): defined in Equation (2.2.2).
- Miller–Maddow corrected MLE (MM): based on the asymptotic bias formula provided by Miller (1955) and Basharin (1959). It is derived from Equation (2.2.3) by estimating m by the number of distinct words observed $\hat{m} = \sum_k 1_{\{n_k \geq 1\}}$ and adding $(\hat{m} - 1)/(2n)$ to the MLE.
- Jackknife (JK): proposed by Zahl (1977). It is a bias-corrected version of the MLE

obtained by averaging all n leave-one-out estimates.

- Best Upper Bounds (BUB): proposed by Paninski (2003). It is obtained by numerically minimizing a worst case error bound for a certain class of linear estimators for a distribution with known support size m.

The NSB estimator proposed by Nemenman et al. (2004) was not included in our simulation comparison because of problems with the software and its computational cost. We also tried their asymptotic formula for their estimator in the "infinite (or unknown)" m case:

$$\psi(1)/\ln(2) - 1 + 2\log n - \psi(n - \hat{m}), \tag{2.4.1}$$

where $\psi(z) = \Gamma'(z)/\Gamma(z)$ is the digamma function. However, we were also unable to get it to work because it seemed to increase unboundedly with the sample size, even for $m = \infty$ cases.

There are two sets of experiments consisting of multiple trials. The first set of experiments concern some simple, but popular model distributions. The second set of experiments deal with neuronal data recorded from primate visual and avian auditory systems. It departs from the theoretical assumptions of Section 2.3 in that the observations are dependent.

Chao and Shen (2003) also conducted a simulation study of the coverage adjusted estimator for distributions with small m and showed that it performs reasonably well even when there is a relatively large fraction of unobserved words. Their article also contains examples from real data sets concerning diversity of species. The experiments presented here are intended to complement their results and expand the scope.

2.4.1 Practical Considerations

There were a few practical hurdles when performing these experiments. The first is that the coverage adjusted estimator is undefined when the sample consists entirely of singletons. In this case $\hat{C} = 0$ and $\tilde{p} = 0$. The probability of this event decays exponentially fast with the sample size, so it is only an issue for relatively small samples. To deal with this matter we replaced the denominator n in the definition of $\hat{C}$ with $n + 1$. This minor modification does not affect the asymptotic behavior of the estimator, and allows it to be defined for all cases.[3]

The BUB estimator assumes that the number of words m is finite and requires that it be specified. m is usually unknown, but sometimes an upper-bound on m may be assumed. To understand the effect of this choice we tried three different variants on the BUB estimator's m parameter:

- Understimate (BUB-): The naive $\hat{m}$ as defined above for the Miller–Maddow corrected MLE.
- Oracle value (BUB.o): The true m in the finite case and $\lceil 2^H \rceil$ in the infinite case.
- Overestimate (BUB+): Twice the oracle value for the first set of experiments and the maximum number of words $|\mathcal{X}|$ for the second set of neuronal data experiments.

Although the BUB estimator is undefined for the m infinite case, we still tried using it, defining the m parameter of the oracle estimator to be $\lceil 2^H \rceil$. This is motivated by the Asymptotic Equipartition Property (AEP) (Cover and Thomas, 1991), which roughly says

[3] Another variation is to add .5 to the numerator and 1 to the denominator.

	support ($k=$)	p_k	H	$\mathrm{Var}[\log p(X)]$
Uniform	$1, \ldots, 1024$	$1/1024$	10	0
Zipf	$1, \ldots, 1024$	$k^{-1}/\sum_k k^{-1}$	7.51	9.59
Poisson	$1, \ldots, \infty$	$1024^k/(k!e^{1024})$	7.05	1.04
Geometric	$1, \ldots, \infty$	$(1023/1024)^{k-1}/1024$	11.4	2.08

Table 2.1: Standard distributions considered in the first set of experiments.

that, asymptotically, 2^H is the effective support size of the distribution. There are no theoretical guarantees for this heuristic use of the BUB estimator, but it did seem to work in the simulation cases below. Again, this is an oracle value and not actually known in practice. The implementation of the estimator was adapted from software provided by Paninski (2003) and its numerical tuning parameters were left as default.

2.4.2 Experimental Setup

In each trial we sample from a single distribution and compute each estimator's estimate of the entropy. Trials are repeated, with 1,000 independent realizations.

Standard Models We consider the four discrete distributions shown in Table 2.1. The uniform and truncated Zipf distributions have finite support ($m = 1,024$), while the Poisson and geometric have infinite support. The Zipf distribution is very popular and often used to model linguistic data. It is sometimes referred to as a "power law." We generated i.i.d. samples of varying sample size (n) from each distribution and computed the respective estimates. We also considered the distribution of distinct words in James Joyce's novel Ulysses. We found that results were very similar to that of the Zipf distribution and did not include them here.

Neuronal Data Here we consider two real neuronal data sets first presented in Theunissen et al. (2001). A subset of the data are available from the *Neural Prediction Challenge*[4]. We fit a variable length Markov chain (VLMC) to subsets of each data set and treated the fitted models as the truth. Our goal was not to model the neuronal data exactly, but to construct an example which reflects real neuronal data, including any inherent dependence. This experiment departs from the assumption of independence for the theoretical results. See Mächler and Bühlmann (2002) for a general overview of the VLMC methodology.

From the first data set, we extracted 10 repeated trials, recorded from a single neuron in the Field L area of avian auditory system during natural song stimulation. The recordings were discretized into $dt = 1$ millisecond bins and consist of sequences of 0's and 1's indicating the absence or presence of a spike. We concatenated the ten recordings before fitting the VLMC (with state space $\{0, 1\}$). A complete description of the physiology and other information theoretic calculations from the data can be found in Hsu et al. (2004).

The other data set contained several separate single neuron recording sequences from the V1 area of primate visual system, during a dynamic natural image stimulation. We used the longest contiguous sequence from one particular trial. This consisted of 3,449 spike counts, ranging from 0 to 5. The counts are number of spikes occurring during consecutive $dt = 16$ millisecond periods. (For the V1 data, the state space of the VLMC is $\{0, 1, 2, 3, 4, 5\}$). The resulting fits for both data sets are shown in Table 2.2. Note that for each VLMC, H/L is nearly the same for both choices of word length (cf. the remarks under equation Equation (2.3.3) in Section 2.2.

[4] `http://neuralprediction.berkeley.edu`

VLMC	depth (msec)	$\mathcal{X}$	word length L	$\lvert\mathcal{X}\rvert$	H	H/L
Field L	232 (232)	$\{0,1\}^{10}$	10	1,024	1.51	0.151
	232 (232)	$\{0,1\}^{15}$	15	32,768	2.26	0.150
V1	3 (48)	$\{0,1,\ldots,5\}^5$	5	7,776	8.32	1.66
	3 (48)	$\{0,1,\ldots,5\}^6$	6	46,656	9.95	1.66

Table 2.2: Fitted VLMC models. Entropy (H) was computed by Monte Carlo with 10^6 samples from the stationary distribution. H/L is the entropy of the word divided by its length.

The (maximum) depth of the VLMC is a measure of time dependence in the data. For the Field L data, the dependence is long, with the VLMC looking 232 time periods (232 msec) into the past. This may reflect the nature of the stimulus in the Field L case. For the V1 data, the dependence is short with the fitted VLMC looking only 3 time periods (48 msec) into the past.

Samples of length n were generated from the stationary distribution of the fitted VLMCs. We subdivided each sample into non-overlapping words of length L. Figure 2.1 shows this for the Field L model with $L = 10$. We tried two different word lengths for each model. The word lengths and entropies are shown in Table 2.2. We then computed each estimator's estimate of entropy on the words and divided by the word length to get an estimate of the *entropy rate* of the word.

We treat m as unknown in this example and did not include the oracle BUB.o in the experiment. We used the maximum possible value of m, i.e. $|\mathcal{X}|$ for BUB+. In the case of Field L with $L = 10$, this is 1,024. The other values are shown in Table 2.2.

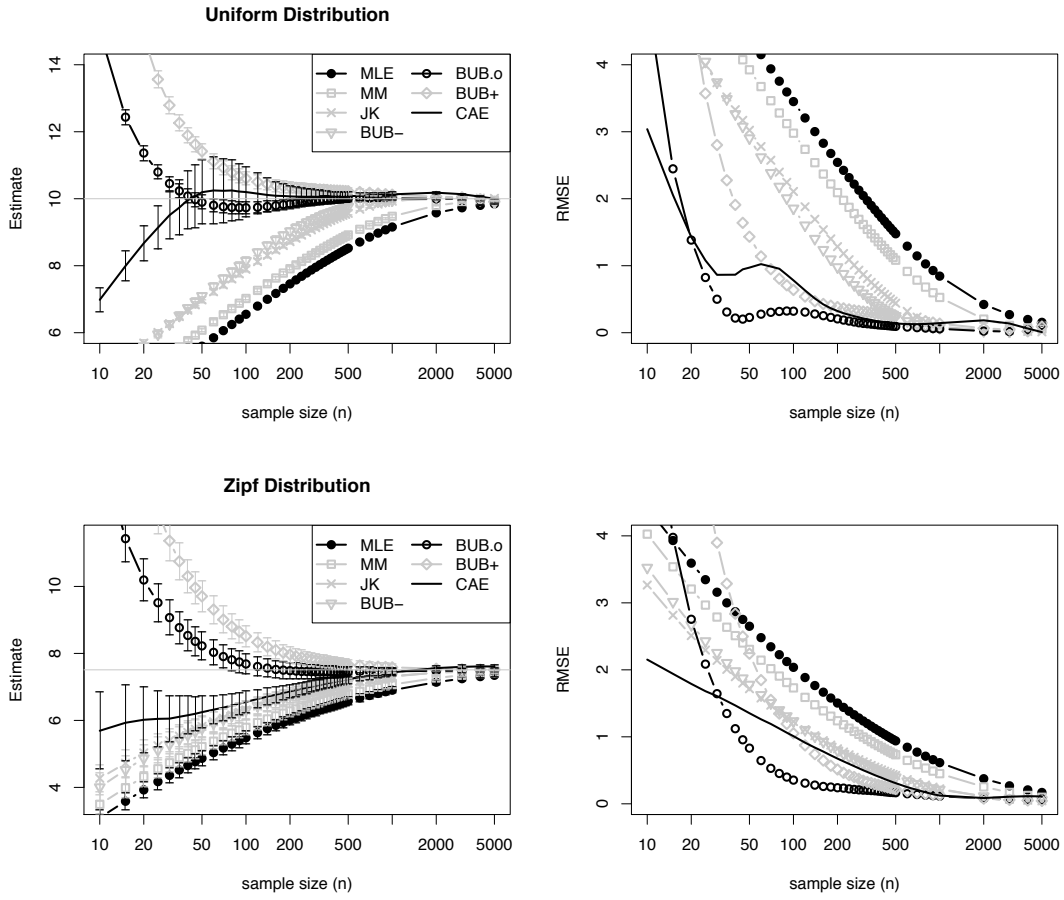


Figure 2.2: The two distributions considered here have finite support, with $m = 1,024$. (Left) The estimated entropy for several different estimators, over a range of sample sizes n. The lines are average estimates taken over 1,000 independent realizations, and the vertical bars indicate $\pm$ one standard deviation of the estimate. The actual value of H is indicated by a solid gray horizontal line. MM and JK are the Miller–Maddow and Jackknife corrected MLEs. BUB-, BUB.o, and BUB+ are the BUB estimator with its m parameter set to a naive $\hat{m}$, oracle $m = 1024$, and twice the oracle m. CAE is the coverage adjusted estimator. (Right) The corresponding root mean squared error (RMSE). Bias dominates most estimates. For the uniform distribution, CAE and BUB.o have relatively small biases and perform very well for sample sizes as small as several hundred. For the Zipf case, the CAE estimator performs nearly as well as the oracle BUB.o for sample sizes larger than 500.

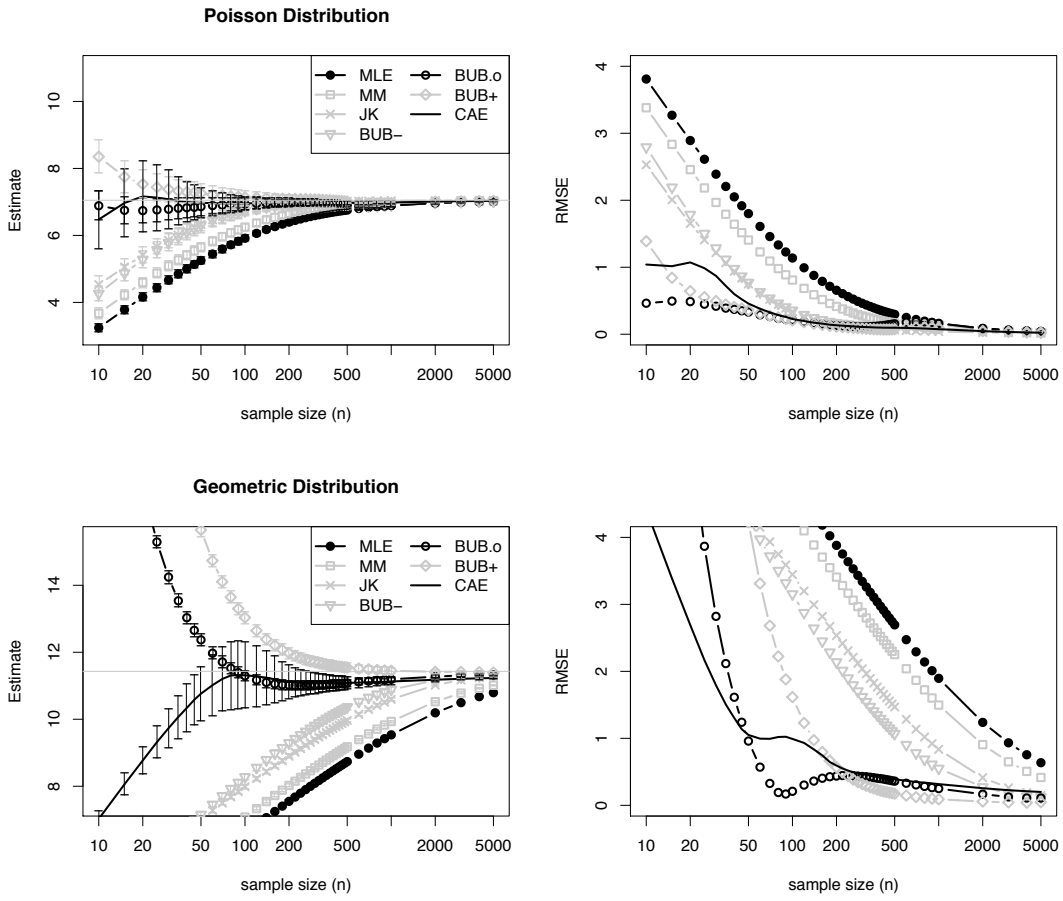


Figure 2.3: Simulation results when sampling from two distributions considered with infinite support ($m = \infty$). (Methodology is identical to that shown in Figure 2.2.) Results are very similar to those in Figure 2.2: the CAE estimator performs nearly as well as the oracle BUB.o.

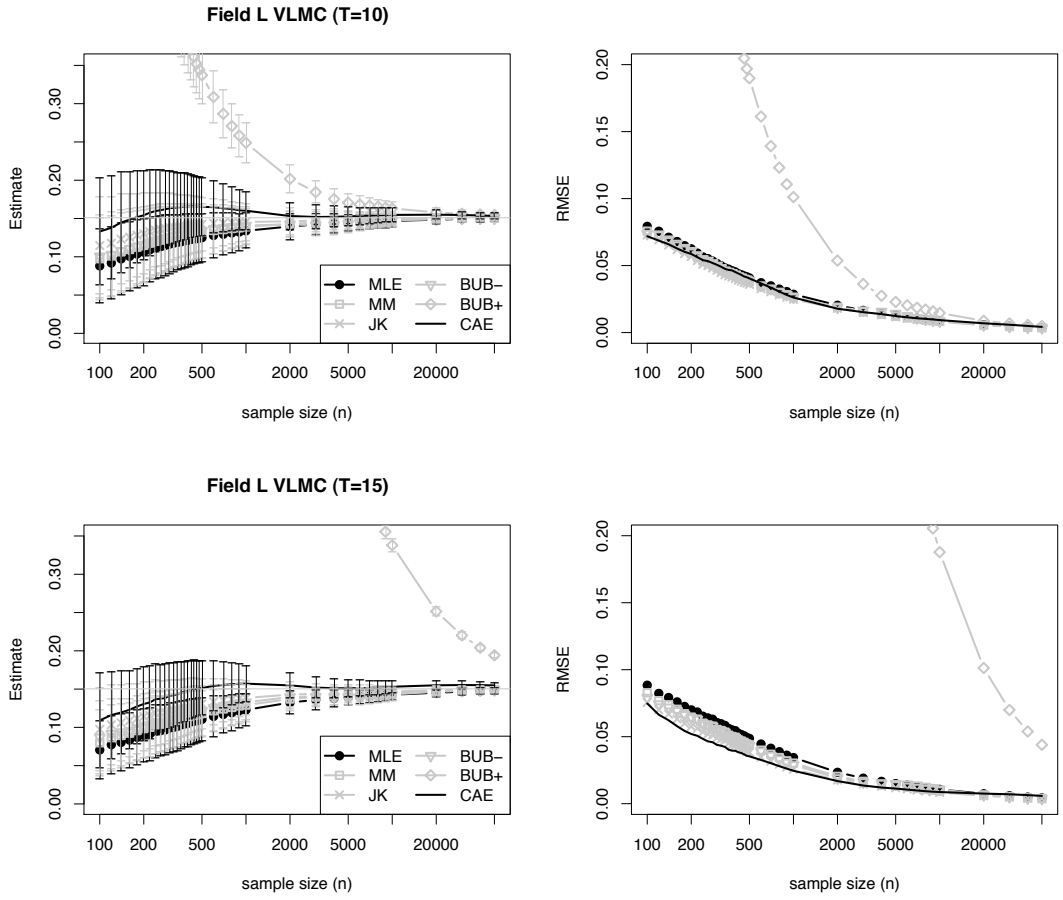


Figure 2.4: Simulation results when sampling from stationary a VLMC fit to Field L neuronal data. (Left) Estimates of entropy rate. Samples of size n, corresponding n 1 millisecond bins, are drawn from a stationary VLMC used to model neuronal data from Field L of avian auditory system. We applied the "direct method" with two different wordlengths L. (Top) $L = 10$ and the maximum number of words is $|\mathcal{X}| = 1,024$. (Bottom) $L = 15$ and $|\mathcal{X}| = 32,768$. The lines are averages taken over 1,000 independent realizations, and the vertical bars indicate $\pm$ one standard deviation of the estimate. The true H/L is indicated by a horizontal line. MM and JK are the Miller–Maddow and Jackknife corrected MLEs. BUB- and BUB+ are the BUB estimator with its m parameter set to a naive estimate $\hat{m}$ and the maximum possible number of words $|\mathcal{X}|$: 1,024 for the top and 32,768 for the bottom. CAE is the coverage adjusted estimator. (Right) Root mean squared error (RMSE). BUB+ has a considerably large bias in both cases. CAE has a moderate balance of bias and variance and shows a visible improvement over all other estimators in the larger ($L = 15$) word case.

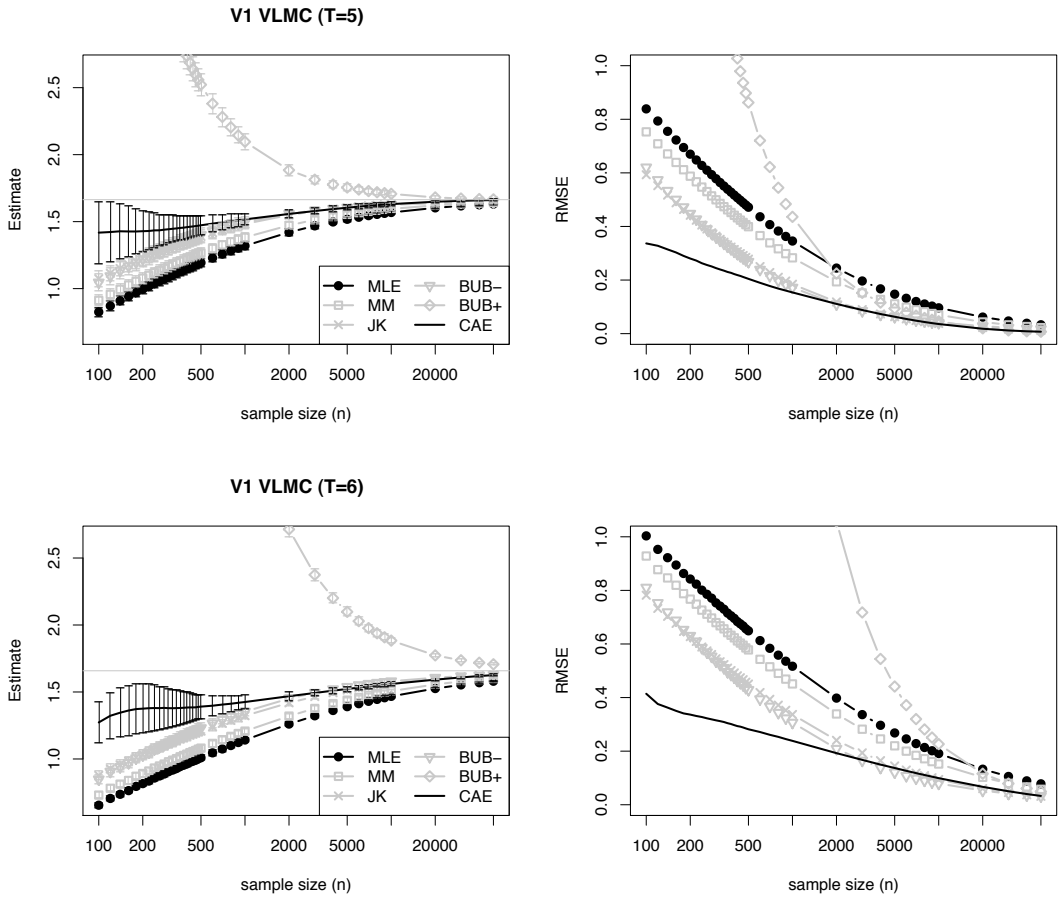


Figure 2.5: Simulation results when sampling from stationary a VLMC fit to primate V1 neuronal data. (Left) Estimates of entropy rate. Samples of size n are drawn from a stationary VLMC used to model neuronal data from V1 of primate visual system. A single sample corresponds to $dt = 16$ milliseconds of recording time. We applied the "direct method" with two different wordlengths L. (Top) $L = 5$ and the maximum number of words $|\mathcal{X}|$ is 7,776. (Bottom) $L = 6$ and $|\mathcal{X}| = 46,656$. The lines are averages taken over 1,000 independent realizations, and the vertical bars indicate $\pm$ one standard deviation of the estimate. The true H/L is indicated by a horizontal line. MM and JK are the Miller–Maddow and Jackknife corrected MLEs. BUB- and BUB+ are the BUB estimator with its m parameter set to a naive $\hat{m}$ and the maximum possible number of words: 7,776 for the top and 46,656 for the bottom. CAE is the coverage adjusted estimator. (Right) Root mean squared error (RMSE). CAE has the smallest bias and performs much better than the other estimators across all sample sizes.

2.4.3 Results

Standard Models The results are plotted in Figure 2.2 and Figure 2.3. It is surprising that good estimates can be obtained with just a few observations. Estimating m from its empirical value marginally improves MM over the MLE. The naive BUB-, which also uses the empirical value of m, performs about the same as JK.

Bias apparently dominates the error of most estimators. The CAE estimator trades away bias for a moderate amount of variance. The RMSE results for the four distributions are very similar. The CAE estimator performs consistently well, even for smaller sample sizes, and is competitive with the oracle BUB.o estimator. The Zipf distribution example seems to be the toughest case for the CAE estimator, but it still performs relatively well for sample sizes of at least 1,000.

Neuronal Data The results are presented in Figure 2.4 and Figure 2.5. The effect of the dependence in the sample sequences is not clear, but all the estimators seem to be converging to the truth. CAE consistently performs well for both V1 and Field L, and really shines in the V1 example. However, for Field L there is not much difference between the estimators, except for BUB+.

BUB+ uses m equal to the maximum number of words $|\mathcal{X}|$ and performs terribly because the data are so sparse. The maximum entropy corresponding to $|\mathcal{X}|$ is much larger than the actual entropy. In the Field L case, the maximum entropies are 10 and 15, while the actual entropies are 1.51 and 2.26. In the V1 case, the maximum entropies are 12.9 and 15.5, while the actual entropies are 8.32 and 9.95. This may be the reason that the BUB+ estimator has such a large positive bias in both cases, because the estimator is designed to

approximately minimize a balance between upper-bounds on worst case bias and variance.

2.4.4 Summary

The coverage adjusted estimator is a good choice for situations where m is unknown and/or infinite. In these situations, the use of an estimator which requires specification of m is disadvantageous because a poor estimate (or upper-bound) of m, or the "effective" m in the infinite case, leads to further error in the estimate. BUB.o, which used the oracle m, performed well in most cases. However, it is typically not available in practice, because m is usually unknown.

The Miller–Maddow corrected MLE, which used the empirical value of m, improved on the MLE only marginally. BUB-, which is BUB with the empirical value of m, performed better than the MLE. It appeared to work in some cases, but not others. For BUB+, where we overestimated or upper-bounded m (by doubling the oracle m, or using the maximal $|\mathcal{X}|$), the bias and RMSE increased significantly over BUB.o for small sample sizes. It appeared to work in some cases, but not others–always alternating with BUB-. In the case of the neuronal data models, BUB+ performed very poorly. In situations like this, even though an upper-bound on m is known, it can be much larger than the "effective" m, and result in a gross error.

2.5 Conclusions

We have emphasized the value of viewing entropy estimation as a problem of sampling from a population, here a population of words made up of spike train sequence

patterns. The coverage adjusted estimator performed very well in our simulation study, and it is very easy to compute. When the word length m is known, the BUB estimator can perform better. In practice, however, m is usually unknown and, as seen in V1 and Field L examples, assuming an upper bound on it can result in a large error. The coverage-adjusted estimator therefore appears to us to be a safer choice.

Other estimates of the probabilities of observed words, such as the profile-based estimator proposed by Orlitsky et al. (2004), might be used in place of $\tilde{P}$ in the coverage adjusted entropy estimator.

As is clear from our simulation study, the dominant source of error in estimating entropy is often bias, rather than variance, which is typically not captured from computed standard errors. An important problem for future investigation would therefore involve data-driven estimation of bias in the case of unknown or infinite m.

The V1 and Field L examples have substantial dependence structure, yet methods derived under the i.i.d. assumption continue to perform well. It may be shown that both the direct method and the coverage-adjusted estimator remain consistent under the relatively weak assumption of stationarity and ergodicity, but the rate of convergence will depend on mixing conditions. On the other hand, in the non-stationary case these methods become inconsistent. Stationarity is, therefore, a very important assumption.

2.6 Proofs

We first prove Theorem 2.3. The proof builds on the following application of a standard concentration technique.

Lemma 2.5. *$\hat{C} \to 1$ almost surely.*

Proof. Consider the number of singletons f_1 as a function of $x_1^n = (x_1, \ldots, x_n)$. Modifying a single coordinate of x_1^n changes the number of singletons by at most 2 because the number of words affected by such a change is at most 2. Hence $\hat{C} = 1 - f_1/n$ changes by at most $2/n$. Using McDiarmid's method of bounded differences, i.e. the Hoeffding-Azuma Inequality, gives

$$\mathbb{P}(|\hat{C} - \mathbb{E}\hat{C}| > \epsilon) \leq 2e^{-\frac{1}{2}n\epsilon^2} \tag{2.6.1}$$

and by consequence of the Borel-Cantelli Lemma, $|C - \mathbb{E}\hat{C}| \to 0$ almost surely. To show that $\mathbb{E}\hat{C} \to 1$, we note that $1 \geq (1 - p_k)^{n-1} \to 0$ for all $p_k > 0$ and

$$\begin{aligned} |1 - \mathbb{E}\hat{C}| &= \mathbb{E}\frac{1}{n}\sum_k 1_{\{n_k=1\}} \\ &= \sum_k p_k(1-p_k)^{n-1} \to 0 \end{aligned} \tag{2.6.2}$$

as $n \to \infty$ by the Bounded Convergence Theorem. ∎

Proof of Proposition 2.1. The bias is

$$\begin{aligned} \mathbb{E}\hat{C} - \mathbb{P}(X_{n+1} \in S) &= \mathbb{P}(X_{n+1} \notin S) - \mathbb{E}(1 - \hat{C}) \\ &= \sum_k p_k(1-p_k)^n - \sum_k p_k(1-p_k)^{n-1} \\ &= -\sum_k p_k^2(1-p_k)^{n-1}. \end{aligned}$$

This quantity is trivially non-positive, and a little bit of calculus shows that the bias is maximized by the uniform distribution $p_k = 1/n$:

$$\begin{aligned}\sum_k p_k^2(1-p_k)^{n-1} &\le \sum_k p_k \max_{0\le x\le 1} x(1-x)^{n-1} \\ &= \max_{0\le x\le 1} x(1-x)^{n-1} \\ &= (1-1/n)^{n-1}/n\end{aligned}$$

The variance bound can be deduced from Equation (2.6.1), because $\operatorname{Var}\hat{C} = \int_0^\infty \mathbb{P}(|\hat{C} - \mathbb{E}\hat{C}|^2 > x)dx$ and Equation (2.6.1) implies

$$\int_0^\infty \mathbb{P}(|\hat{C} - \mathbb{E}\hat{C}|^2 > x)dx \le \int_0^\infty 2e^{-\frac{1}{2}nx}dx = 4/n. \qquad \blacksquare$$

Proof of Theorem 2.3. From Equation (2.6.1) we conclude that $\hat{C} = \mathbb{E}\hat{C} + \mathcal{O}_{\mathrm{P}}\left(n^{-1/2}\right)$. So it suffices to show that $\mathbb{E}\hat{C} = 1 + \mathcal{O}\left(1/(\log n)^q\right)$. Let $\epsilon_n = 1/\sqrt{n}$. We split the summation in Equation (2.6.2):

$$|1 - \mathbb{E}\hat{C}| = \sum_{k:p_k\le\epsilon_n} p_k(1-p_k)^{n-1} + \sum_{k:p_k>\epsilon_n} p_k(1-p_k)^{n-1}$$

Using Lemma 2.6 below, the first term on the right side is

$$\sum_{k:p_k\le\epsilon_n} p_k(1-p_k)^{n-1} \le \sum_{k:p_k\le\epsilon_n} p_k = \mathcal{O}\left(1/(\log n)^q\right)$$

The second term is

$$\begin{aligned}\sum_{k:p_k>\epsilon_n} p_k(1-p_k)^{n-1} &\le (1-\epsilon_n)^{n-1} \sum_{k:p_k>\epsilon_n} p_k \\ &\le (1-\epsilon_n)^{n-1} \\ &\le \exp(-(n-1)/\sqrt{n})\end{aligned}$$

by the well-known inequality $1 + x \le e^x$. $\blacksquare$

Lemma 2.6 (Wyner and Foster (2003))**.**

$$\sum_{k:p_k\le\epsilon} p_k \le \frac{\sum_k p_k|\log p_k|^q}{\log(1/\epsilon)^q}$$

Proof. Since $\log(1/x)$ is a decreasing function,

$$\sum_{k:p_k\le\epsilon} p_k \left|\log\frac{1}{p_k}\right|^q \ge \sum_{k:p_k\le\epsilon} p_k \left|\log\frac{1}{\epsilon}\right|^q$$

and then we collect the $\log(1/\epsilon)^q$ term to the left side to derive the claim. ∎

Proof of Theorem 2.2. Using the result of Wyner and Foster (2003) that under the above assumptions, $\hat{H} = H + \mathcal{O}_{\mathrm{P}}(1/\log n)$, it suffices to show $|\tilde{H} - \hat{H}| = \mathcal{O}_{\mathrm{P}}(1/\log n)$. All summations below will only be over k such that $\hat{p}_k > 0$ or $p_k > 0$. It is easily verified that

$$\begin{aligned}\tilde{H} - \hat{H} &= -\sum_k \frac{\tilde{p}_k \log \tilde{p}_k}{1-(1-\tilde{p}_k)^n} - \hat{p}_k \log \hat{p}_k \\ &= \underbrace{-\sum_k \left[\frac{\hat{C}}{1-(1-\tilde{p}_k)^n} - 1\right] \hat{p}_k \log \hat{p}_k}_{D_n} \\ &\quad \underbrace{-\sum_k \frac{\hat{C}\hat{p}_k \log \hat{C}}{1-(1-\tilde{p}_k)^n}}_{R_n}\end{aligned}$$

To bound R_n we will use the $\mathcal{O}_{\mathrm{P}}\left(1/(\log n)^2\right)$ rate of $\hat{C}$ from Theorem 2.3. Note that $\hat{C}/n \le \hat{C}\hat{p}_k = \tilde{p}_k \le 1$ and by the decreasing nature of $1/[1-(1-\tilde{p}_k)^n]$

$$|R_n| \le \frac{|\log \hat{C}|}{1-(1-\hat{C}/n)^n} \sum_k \hat{p}_k = \frac{|\log \hat{C}|}{1-(1-\hat{C}/n)^n}$$

By Lemma 2.5, $\hat{C} \to 1$ almost surely and since $x_n \to 1$ implies $(1-x_n/n)^n \to e^{-1}$, the right side is $\sim |\log \hat{C}|/(1-e^{-1}) = \mathcal{O}_{\mathrm{P}}\left(1/(\log n)^2\right)$. As for D_n,

$$|D_n| \le -\sum_k \frac{|\hat{C}-1| + (1-\tilde{p}_k)^n}{1-(1-\tilde{p}_k)^n} \hat{p}_k \log \hat{p}_k$$

and since $\tilde{p}_k \geq \hat{C}/n$ whenever $\tilde{p}_k > 0$,

$$\begin{aligned} -\sum_k \frac{|\hat{C}-1|}{1-(1-\tilde{p}_k)^n}\hat{p}_k \log \hat{p}_k &\leq \frac{|\hat{C}-1|}{1-(1-\hat{C}/n)^n}\hat{H} \\ &\sim \frac{|\hat{C}-1|}{1-e^{-1}}\hat{H} \\ &= \mathcal{O}_{\mathrm{P}}\left(1/(\log n)^2\right) \end{aligned}$$

because $\hat{H}$ is consistent. The remaining part of D_n will require a bit more work and we will split it according to the size of $\hat{p}_k$. Let $\epsilon_n = \log n/n$. Then

$$\begin{aligned} -\sum_k \frac{(1-\tilde{p}_k)^n}{1-(1-\tilde{p}_k)^n}\hat{p}_k \log \hat{p}_k = &- \sum_{k:\hat{p}_k \leq \epsilon_n} \frac{(1-\tilde{p}_k)^n}{1-(1-\tilde{p}_k)^n}\hat{p}_k \log \hat{p}_k \\ &- \sum_{k:\hat{p}_k > \epsilon_n} \frac{(1-\tilde{p}_k)^n}{1-(1-\tilde{p}_k)^n}\hat{p}_k \log \hat{p}_k \end{aligned}$$

Similarly to our previous argument, $\frac{(1-\tilde{p}_k)^n}{1-(1-\tilde{p}_k)^n}$ is decreasing in $\tilde{p}_k$. So the second summation on the right side is

$$\begin{aligned} -\sum_{k:\hat{p}_k > \epsilon_n} \frac{(1-\tilde{p}_k)^n}{1-(1-\tilde{p}_k)^n}\hat{p}_k \log \hat{p}_k &\leq \frac{(1-\epsilon_n)^n}{1-(1-\epsilon_n)^n}\hat{H} \\ &= \mathcal{O}_{\mathrm{P}}\left(1/n\right) \end{aligned}$$

For the remaining summation, we use the fact that $\tilde{p}_k \geq \hat{C}/n$ and the monotonicity argument once more.

$$-\sum_{k:\hat{p}_k \leq \epsilon_n} \frac{(1-\tilde{p}_k)^n}{1-(1-\tilde{p}_k)^n}\hat{p}_k \log \hat{p}_k \leq -\frac{(1-\hat{C}/n)^n}{1-(1-\hat{C}/n)^n} \sum_{k:\hat{p}_k \leq \epsilon_n} \hat{p}_k \log \hat{p}_k$$

By the consistency of $\hat{C}$, the leading term converges to the constant $e^{-1}/(1-e^{-1})$ and can be ignored. Since $-\log \hat{p}_k \leq \log n$,

$$-\sum_{k:\hat{p}_k \leq \epsilon_n} \hat{p}_k \log \hat{p}_k \leq \log n \sum_{k:\hat{p}_k \leq \epsilon_n} \hat{p}_k$$

We split the summation once last time, but according to the size of p_k.

$$\log n \sum_{k:\hat{p}_k \leq \epsilon_n} \hat{p}_k \leq \log n \left[\sum_{k:p_k > 1/\sqrt{n}} \epsilon_n + \sum_{k:p_k \leq 1/\sqrt{n}} \hat{p}_k \right]$$
$$\leq \frac{(\log n)^2}{\sqrt{n}} + \log n \sum_{k:p_k \leq 1/\sqrt{n}} \hat{p}_k,$$

where we have used the fact that $|\{k : p_k > 1/\sqrt{n}\}| \leq \sqrt{n}$. Taking expectation, applying Lemma 2.6 and Markov's Inequality shows that

$$= \log n \sum_{k:p_k \leq 1/\sqrt{n}} \hat{p}_k = \mathcal{O}_{\mathrm{P}}\left(1/\log n\right)$$

The proof is complete because $(\log n)^2/\sqrt{n}$ is also $\mathcal{O}\left(1/\log n\right)$. ■

Chapter 3

Information Under Non-Stationarity

3.1 Introduction

Information estimates are frequently calculated using data from experiments where the stimulus and response are dynamic and time-varying (see for instance Hsu et al. (2004); Reich et al. (2001); Reinagel and Reid (2000); Nirenberg et al. (2001)). For mutual information to be properly defined, see for example Cover and Thomas (1991), the stimulus and response must be considered random, and when the estimates are obtained from time-averages, they should also be stationary and ergodic. In practice these assumptions are usually tacit, and information estimates, such as the *direct method* proposed by Strong et al. (1998), can be made without explicit consideration of the stimulus. This can lead to misinterpretation.

In this chapter we show that the direct method information estimate can be reinterpreted as the average divergence across time of the conditional response distribution from its overall mean; in the absence of stationarity and ergodicity:

1. information estimates do not necessarily estimate mutual information, but
2. potentially useful interpretations can still be made by referring back to the time-varying divergence.

Although our results are specialized to the direct method with the plug-in entropy estimator, they should hold more generally regardless of the choice of entropy estimator.

The fundamental issue concerns stationarity: methods that assume stationarity are unlikely to be appropriate when stationarity appears to be violated. In the non-stationary case, our second result should be of use, as would be other methods that explicitly consider the dynamic and non-stationary nature of the stimulus and response; see for instance Barbieri et al. (2004).

We begin with a brief review of the direct method and plug-in entropy estimator. This is followed by results showing that the information estimate can be recast as a time-average. This characterization leads us to the interpretation that the information estimate is actually a measure of variability of the stimulus conditioned response distribution. This observation is first made in the finite number of trials case, and then formalized by a theorem describing the limiting behavior of the information estimate as the number of trials tends to infinity. Following the theorem is discussion about the interpretation of the limit, and examples that illustrate the interpretation with a proposed graphical plot.

3.2 The Direct Method

In the direct method a time-varying stimulus is chosen by the experimenter and then repeatedly presented to a subject over multiple trials. The observed responses are conditioned by the same stimulus. Two types of variation in the response are considered:

1. variation across time (potentially related to the stimulus), and

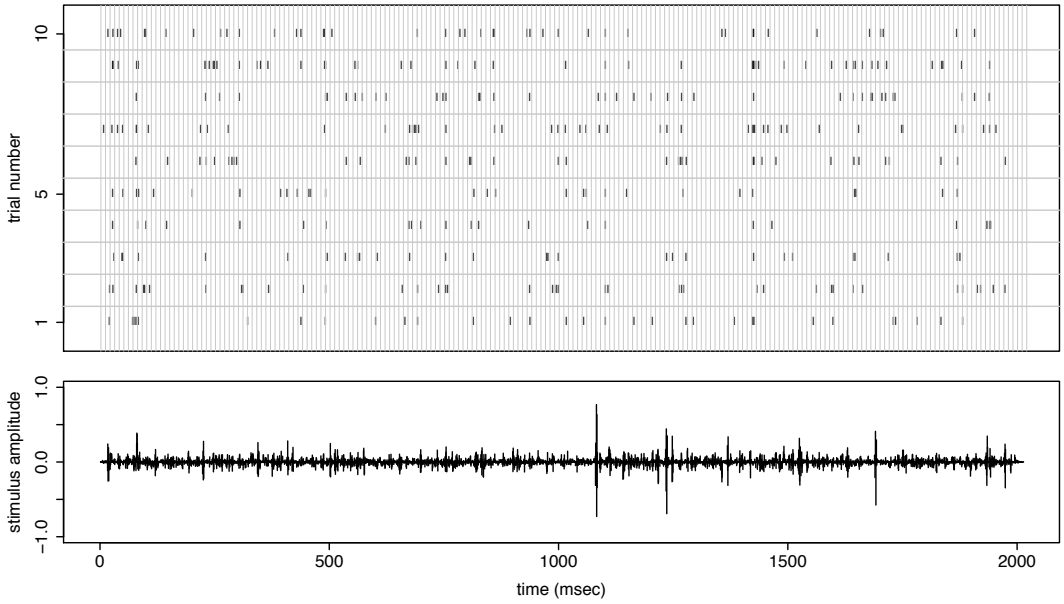
2. trial-to-trial variation.

Figure 3.1(a) shows an example of data from such an experiment. The upper panel is a raster plot of the response of a Field L neuron of an adult male Zebra Finch during synthetic song stimulation. The lower panel is a plot of the audio signal corresponding to the natural song. Details of the experiment can be found in Hsu et al. (2004).

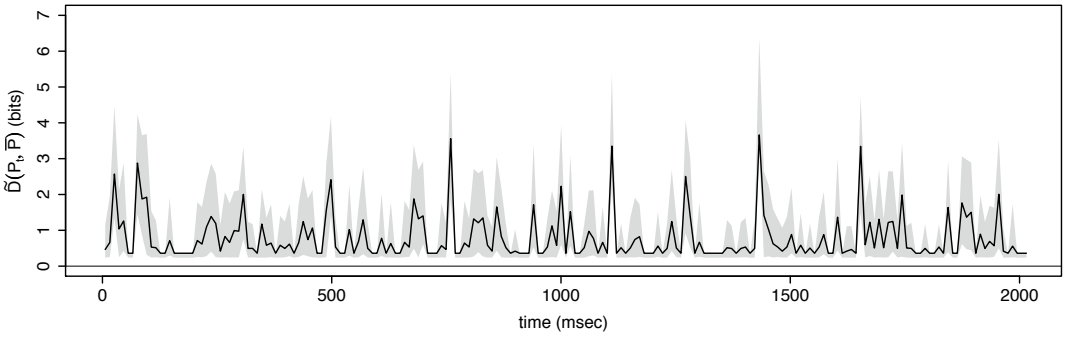
Let us consider the random process $\{S_t, X_t^k\}$ representing the value of the stimulus and response at time $t = 1, \ldots, n$ during trial $k = 1, \ldots, m$. The response is made discrete by dividing time into bins of size dt and then considering *words* (or patterns) of spike counts formed within intervals (overlapping or non-overlapping) of L adjacent time bins. The number of spikes that occur in each time bin become the letters in the words. X_t^k corresponds to these words, and may belong to a countably infinite set (because the number of spikes in a bin is theoretically unbounded). In the raster plot of Figure 3.1(a) the time bin size is $dt = 1$ millisecond, and the vertical lines demarcate non-overlapping words of length $L = 10$ time bins.

Given the responses $\{X_t^k\}$, the direct method considers two different entropies:

1. the *total entropy* H of the response, and



(a) Stimulus and response



(b) Divergence plot

Figure 3.1: (a) Raster plot of the response of the a Field L neuron of an adult male Zebra Finch (above) during the presentation of a synthetic audio stimulus (below) for 10 repeated trials. The vertical lines indicate boundaries of $L = 10$ millisecond (msec) words formed at a resolution of $dt = 1$ msec. The data consists of 10 trials, each of duration 2000 msecs. (b) The coverage adjusted estimate (solid line) of $D(P_t, \bar{P})$ from the response shown above with 10 msec words. Pointwise 95% confidence intervals are indicated by the shaded region and obtained by bootstrapping the trials 1000 times. The information estimate, 0.77 bits (per 10msec word, or 0.077 bits/msec), corresponds to the average value of the solid curve.

2. the local *noise entropy* H_t of the response at time t.

The total entropy is associated with the stimulus conditioned distribution of the response across all times and trials. The local noise entropy is associated with the stimulus conditioned distribution of the response at time t across all trials. These quantities are calculated directly from the neural response, and the difference between the total entropy and the average (over t) noise entropy is what Strong et al. (1998) call "the information that the spike train provides about the stimulus."

H and H_t depend implicitly on the length L of the words. Normalizing by L and considering large L leads to the total and local entropy rates that are defined to be $\lim_{L\to\infty} H(L)/L$ and $\lim_{L\to\infty} H_t(L)/L$, respectively, when they exist. The direct method of Strong et al. (1998) prescribed an extrapolation for estimating these limits, however they do not necessarily exist when the stimulus and response process are non-stationary. When there is stationarity, estimation of entropy for large L is potentially difficult, and extrapolation from a few small choices of L can be suspect. Since we are primarily interested in the non-stationary case, we do not address these issues and refer the reader to Kennel et al. (2005); Gao et al. (2006) for larger discussion on the stationary case. For notational simplicity, the dependence on L will be suppressed.

The plug-in entropy estimate Strong et al. (1998) proposed estimating H and H_t by plug-in with the corresponding empirical distributions:

$$\hat{P}(r) := \frac{1}{mn} \sum_{t=1}^{n} \sum_{k=1}^{m} 1_{\{X_t^k = r\}} \tag{3.2.1}$$

and

$$\hat{P}_t(r) := \frac{1}{m}\sum_{k=1}^{m} 1_{\{X_t^k = r\}}. \tag{3.2.2}$$

Note that $\hat{P}$ is also the average of $\hat{P}_t$ across $t = 1, \ldots, n$. So the direct method *plug-in* estimates[1] of H and H_t are

$$\hat{H} := -\sum_r \hat{P}(r) \log \hat{P}(r),$$

and

$$\hat{H}_t := -\sum_r \hat{P}_t(r) \log \hat{P}_t(r),$$

respectively. The direct method plug-in information estimate is

$$\hat{I} := \hat{H} - \frac{1}{n}\sum_{t=1}^{n} \hat{H}_t. \tag{3.2.3}$$

3.3 Interpretation of the Information Estimate

The direct method information estimate is not only the difference of entropies shown in Equation (3.2.3), but also a time-average of divergences. The empirical distribution of response across all trials and times Equation (3.2.1) is equal to the average of $\hat{P}_t$ over

[1] Strong et al. (1998) used the name *naive estimates*.

time. That is $\hat{P}(r) = n^{-1}\sum_{t=1}^{n}\hat{P}_t(r)$ and so

$$\begin{aligned}\hat{I} &= \hat{H} - \frac{1}{n}\sum_{t=1}^{n}\hat{H}_t \\ &= \frac{1}{n}\sum_{t=1}^{n}\sum_{r}\hat{P}_t(r)\log\hat{P}_t(r) - \sum_{r}\left[\frac{1}{n}\sum_{t=1}^{n}\hat{P}_t(r)\right]\log\hat{P}(r) \\ &= \frac{1}{n}\sum_{t=1}^{n}\sum_{r}\hat{P}_t(r)\log\hat{P}_t(r) - \frac{1}{n}\sum_{t=1}^{n}\sum_{r}\hat{P}_t(r)\log\hat{P}(r) \\ &= \frac{1}{n}\sum_{t=1}^{n}\sum_{r}\hat{P}_t(r)\log\frac{\hat{P}_t(r)}{\hat{P}(r)}. \end{aligned} \tag{3.3.1}$$

The quantity that is averaged over time in Equation (3.3.1) is the Kullback-Leibler divergence between the empirical time t response distribution $\hat{P}_t$ and the average empirical response distribution $\hat{P}$.

Since the same stimulus is repeatedly presented to the subject, and there is no evolution in the response, over multiple trials, a *repeated trial assumption* is natural:

- Conditional on the stimulus $\{S_t\}$ the m trials $\{S_t, X_t^1\}, \ldots, \{S_t, X_t^m\}$ are independent and identically distributed (i.i.d.).

Under this assumption $1_{\{X_t^1=r\}}, \ldots, 1_{\{X_t^m=r\}}$ are conditionally i.i.d. for each fixed t and r. Furthermore, the law of large numbers guarantees that as the number of trials m increases the empirical response distribution $\hat{P}_t(r)$ converges to its conditional expected value for each fixed t and r. Thus $\hat{P}_t(r)$ and $\hat{P}(r)$ can be viewed as estimates of $P_t(r|S_1, \ldots, S_n)$, defined by

$$P_t(r|S_1, \ldots, S_n) := \mathbb{P}(X_t^k = r|S_1, \ldots, S_n) = \mathbb{E}\{\hat{P}_t(r)|S_1, \ldots, S_n\},$$

and $\bar{P}(r|S_1, \ldots, S_n)$, defined by

$$\bar{P}(r|S_1, \ldots, S_n) := \frac{1}{n}\sum_{t=1}^{n}P_t(r|S_1, \ldots, S_n),$$

respectively. $\bar{P}$ is average response distribution across time $t = 1, \ldots, n$ conditional on the entire stimulus $\{S_1, \ldots, S_n\}$.

So the quantity that is averaged over time in Equation (3.3.1) should be viewed as a plug-in estimate of the Kullback-Leibler divergence between P_t and $\bar{P}$. We emphasize this by writing

$$\hat{D}(P_t||\bar{P}) := \sum_r \hat{P}_t(r) \log \frac{\hat{P}_t(r)}{\hat{P}(r)}.$$

This observation will be formalized by the theorem of the next section. For now we summarize the above with a proposition.

Proposition 3.1. *The information estimate is the time-average* $\hat{I} = \frac{1}{n} \sum_{t=1}^{n} \hat{D}(P_t||\bar{P})$.

This decomposition of the information estimate is analogous to the decomposition of mutual information that Deweese and Meister (1999) call the "specific surprise," while "specific information" is analogous to the alternative decomposition,

$$\hat{I} = \frac{1}{n} \sum_{t=1}^{n} [\hat{H} - \hat{H}_t]. \tag{3.3.2}$$

An important difference is that here the stimulus itself is a function of time and the decompositions are given in terms of time-dependent quantities. It is possible that these quantities can reveal dynamic aspects of the stimulus and response relationship. This will be explored further in Section 3.3.2 and Section 3.3.4.

3.3.1 What is being estimated?

There are two directions in which the amount of observed response data can be increased: length of time n, and number of trials m. The information estimate is the average

of $\hat{D}(P_t||\bar{P})$ over time, and may not necessarily converge as n increases. This could be due to $\{S_t, X_t^k\}$ being non-stationary and/or highly dependent in time. Even when convergence may occur, the presence of serial correlation. See the autocorrelation in Figure 3.2(b) for example. in $\hat{D}(P_t||\bar{P})$ can make assessments of uncertainty in $\hat{I}$ difficult.

Assuming that the stimulus and response process is stationary and not too dependent in time could guarantee convergence, but this could be unrealistic. On the other hand, the repeated trial assumption is appropriate if the same stimulus is repeatedly presented to the subject over multiple trials. It is also enough to guarantee that the information estimate converges as the number of trials m increases.

Theorem 3.2. *Suppose that P_t has finite entropy for all $t = 1, \ldots, n$. Then under the repeated trial assumption*

$$\lim_{m\to\infty} \hat{I} = H(\bar{P}) - \frac{1}{n}\sum_{t=1}^{n} H(P_t) = \frac{1}{n}\sum_{t=1}^{n}[H(\bar{P}) - H(P_t)] = \frac{1}{n}\sum_{t=1}^{n} D(P_t||\bar{P})$$

with probability 1, and in particular the following statements hold uniformly for $t = 1, \ldots, n$ with probability 1:

1. $\lim_{m\to\infty} \hat{H} = H(\bar{P})$,
2. $\lim_{m\to\infty} \hat{H}_t = H(P_t)$, *and*
3. $\lim_{m\to\infty} \hat{D}(P_t||\bar{P}) = D(P_t||\bar{P})$ *for* $t = 1, \ldots, n$,

where $D(P_t||\bar{P})$ is the Kullback-Leibler divergence defined by,

$$D(P_t||\bar{P}) := \sum_r P_t(r|S_1, \ldots, S_n) \log \frac{P_t(r|S_1, \ldots, S_n)}{\bar{P}(r|S_1, \ldots, S_n)},$$

and $H(P)$ is the entropy of the distribution P, defined by

$$H(P) := -\sum_r P(r) \log P(r).$$

See Section 3.5 for the proof. Note that if stationary and ergodicity do hold, then P_t for $t = 1, \ldots, n$ is also stationary and ergodic[2]. So its average, $\bar{P}(r)$, is guaranteed by the ergodic theorem to converge pointwise to $P(X_1^1 = r)$ as $n \to \infty$. Moreover, if X_1^1 can only take on a finite number of values, then $H(\bar{P})$ also converges to the marginal entropy $H(X_1^1)$ of X_1^1. Likewise, the average of the conditional entropy $H(P_t)$ also converges to the expected conditional entropy: $\lim_{n\to\infty} H(X_n^1|S_1, \ldots, S_n)$. So in this case the information estimate does indeed estimate mutual information.

However, the primary consequence of the theorem is that, in the absence of stationarity and ergodicity, the information estimate $\hat{I}$ does not necessarily estimate mutual information. The three particular statements show that the time-varying quantities $[\hat{H} - \hat{H}_t]$ and $\hat{D}(P_t||\bar{P})$ converge individually to the appropriate limits, and justify our assertion that the information estimate is a time-average of plug-in estimates of the corresponding time-varying quantities. Thus, the information estimate can always be viewed as an estimate of the time-average of either $D(P_t||\bar{P})$ or $[H(P) - H(P_t)]$—stationary and ergodic or not.

3.3.2 Time-averaged Divergence

The Kullback-Leibler Divergence $D(P_t||\bar{P})$ has a simple interpretation: it measures the dissimilarity of the time t response distribution P_t from its overall average $\bar{P}$. So as a function of time, $D(P_t||\bar{P})$ measures how the conditional response distribution varies across

[2] P_t and $\bar{P}$ are stimulus conditional distributions, and hence random variables potentially depending on $S_1, \ldots, S_n$.

time, relative to its overall mean. This can be seen in a more familiar form by considering the leading term of the Taylor expansion,

$$D(P_t||\bar{P}) = \frac{1}{2}\sum_r \frac{[P_t(r|S_1,\ldots,S_n) - \bar{P}(r|S_1,\ldots,S_n)]^2}{\bar{P}(r|S_1,\ldots,S_n)} + \cdots .$$

Thus, its average is in this sense a measure of the average variability of the response distribution.

It is, of course, possible that characteristics of the response are due to confounding factors rather than the stimulus. Furthermore, the presence of additional noise in either process would weaken a measured relationship between stimulus and response, compared to its strength if the noise were eliminated. Setting these concerns aside, the variation of the response distribution P_t about its average provides information about the relationship between the stimulus and the response. In the stationary and ergodic case, this information may be averaged across time to obtain mutual information. In more general settings averaging across time may not provide a complete picture of the relationship between stimulus and response. Instead, we suggest examining the time-varying $D(P_t||\bar{P})$ directly, via graphical display as discussed next.

3.3.3 Coverage Adjusted Estimation of $D(P_t||\bar{P})$

The plug-in estimate $\hat{D}(P_t||\bar{P})$ is an obvious choice for estimating $D(P_t||\bar{P})$, but it turns out that estimating $D(P_t||\bar{P})$ is akin to estimating entropy. Since the trials are conditionally i.i.d., the coverage adjustment method described in Chapter 2 can be used to improve estimation of $D(P_t||\bar{P})$ over the plug-in estimate. The main idea behind coverage adjustment is to adjust estimates for potentially unobserved values. This happens in two

places: estimation of P_t and estimation of $D(P_t||\bar{P})$. In the first case, unobserved values affect the amount of weight that $\hat{P}_t$, defined in Equation (3.2.2) in the main text, places on observed values. In the second case unobserved values correspond to missing summands when plugging $\hat{P}_t$ into the Kullback-Leibler divergence. See Chapter 2 for a more thorough explanation of these ideas. Let

$$N_t(r) := \sum_{k=1}^{m} 1_{\{X_t^k = r\}}.$$

The sample coverage, or total P_t-probability of observed values r, is estimated by $\hat{C}_t$ defined by

$$\hat{C}_t := 1 - \frac{\#\{r : N_t(r) = 1\} + .5}{m+1}.$$

The number in the numerator of the fraction refers to the number of singletons—patterns that were observed only once across the m trials at time t. Then the coverage adjusted estimate of P_t is the following shrunken version of $\hat{P}_t$:

$$\tilde{P}_t(r) = \hat{C}_t \hat{P}_t(r).$$

$\bar{P}$ is estimated by simply averaging $\tilde{P}_t$:

$$\tilde{P}(r) = \frac{1}{n} \sum_{t=1}^{n} \tilde{P}_t(r).$$

The coverage adjusted estimate of $D(P_t||\bar{P})$ is obtained by plugging $\tilde{P}_t$ and $\tilde{P}$ into the Kullback-Leibler divergence, but with an additional weighting on the summands according to the inverse of the estimated probability that the summand is observed:

$$\tilde{D}(P_t||\bar{P}) := \sum_r \frac{\tilde{P}_t(r)\{\log \tilde{P}_t(r) - \log \tilde{P}(r)\}}{1 - (1 - \tilde{P}_t(r))^m}.$$

The additional weighting is to correct for potentially missing summands. Confidence intervals for $D(P_t||\bar{P})$ can be obtained by bootstrap sampling entire trials, and applying $\tilde{D}$ to the bootstrap replicate data.

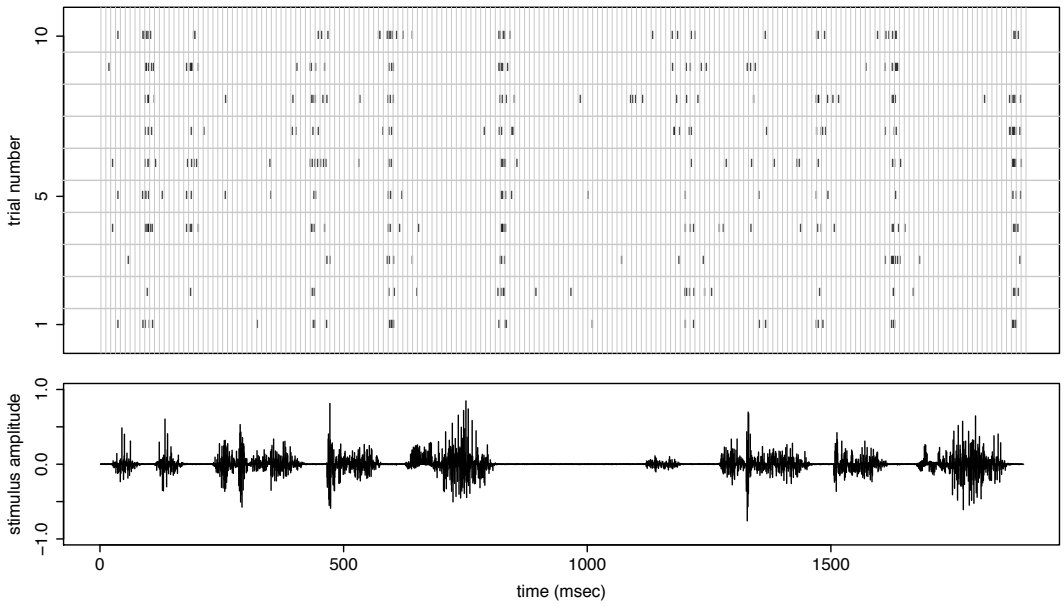
3.3.4 Plotting $D(P_t||\bar{P})$

Figure 3.1(a) and Figure 3.2(a) show the responses of the same Field L neuron of an adult male Zebra Finch under two different stimulus conditions. Details of the experiment and the statistics of the stimuli are described in Hsu et al. (2004). Panel (a) of the figures shows the stimulus and response data. In Figure 3.1(a) the stimulus is synthetic and stationary by construction, while in Figure 3.2(a) the stimulus is a natural song. Panel (b) of the figures shows the coverage adjusted estimate of the divergence $D(P_t||\bar{P})$ plotted as a function of time. 95% confidence intervals were formed by bootstrapping entire trials, i.e. an entire trial is either included in or excluded from a bootstrap sample.

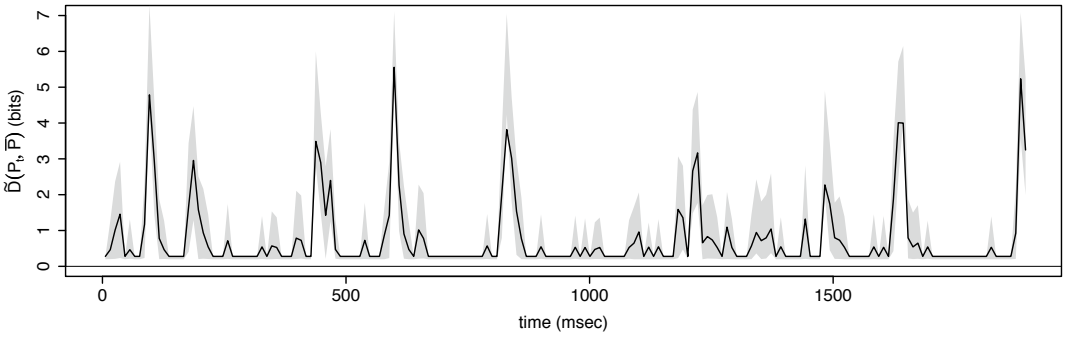
The information estimate going along with each Divergence plot is the average of the solid curve representing the estimate of $D(P_t||\bar{P})$. It is equal to 0.77 bits (per 10 millisecond word) in Figure 3.1(b) and 0.76 bits (per 10 millisecond word) in Figure 3.2(b). Although the information estimates are nearly identical, the two plots are very different.

In the first case, the stimulus is stationary by construction and it appears that the time-varying divergence is too. Its fluctuations appear to be roughly of the same scale across time, and its local mean is relatively stable. The average of the solid curve seems to be a fair summary.

In the second case the stimulus is a natural song. The isolated bursts of the time-



(a) Stimulus and response



(b) Divergence plot

Figure 3.2: (a) Same as in Figure 3.1(a), but in this set of trials the stimulus is a conspecific natural song. (b) The coverage adjusted estimate (solid line) of $D(P_t, \bar{P})$ from the response shown above. Pointwise 95% confidence intervals are indicated by the shaded region and obtained by bootstrapping the trials 1000 times. The information estimate, 0.76 bits (per 10 msec word or 0.076 bits/msec), corresponds to the average value of the solid curve.

varying divergence and relatively flat regions in Figure 3.2(b) suggest that the response process (and the divergence) is non-stationary and has strong serial correlations. The local mean of the divergence also varies strongly with time. Summarizing $D(P_t||\bar{P})$ by its time-average hides the time-dependent features of the plot.

More interestingly, when the divergence plot is compared to the plot of the stimulus in Figure 3.2(a), there is a striking coincidence between the location of large isolated values of the estimated divergence and visual features of the stimulus waveform. They tend to coincide with the boundaries of the bursts in the stimulus signal. This suggests that the spike train may carry information about the onset/offset of bursts in the stimulus. We discussed this with the Theunissen Lab and they confirmed from their STRF models that the cell in the example is an offset cell. It tends to fire at the offsets of song syllables—the bursts of energy in the stimulus waveform. They also suggested that a word length within the range of 30–50 milliseconds is a better match to the length of correlations in the auditory system. We regenerated the plots for words of length $L = 40$ (not shown here) and found that the isolated structures in the divergence plot became even more pronounced.

3.4 Conclusions

Estimates of mutual information, including the plug-in estimate, may be viewed as measures of the strength of the relationship between the response and the stimulus when the stimulus and response are jointly stationary and ergodic. Many applications, however, use non-stationary or even deterministic stimuli, so that mutual information is no longer well defined. In such non-stationary cases do estimates of mutual information become

meaningless? We think not, but the purpose of this chapter has been to point out the delicacy of the situation, and to suggest a viable interpretation of information estimates, along with the divergence plot, in the non-stationary case.

In using stochastic processes to analyze data there is an implicit practical acknowledgment that assumptions cannot be met precisely: the mathematical formalism is, after all, an abstraction imposed on the data; the hope is simply that the variability displayed by the data is similar in relevant respects to that displayed by the presumptive stochastic process. The "relevant respects" involve the statistical properties deduced from the stochastic assumptions. The point we are trying to make is that highly non-stationary stimuli make statistical properties based on an assumption of stationarity highly suspect; strictly speaking, they become void.

To be more concrete, let us reconsider the snippet of natural song and response displayed in Figure 3.2(a). When we look at the less than 2 seconds of stimulus amplitude given there, the stimulus is not at all time-invariant: instead, the stimulus has a series of well-defined bursts followed by periods of quiescence. Perhaps, on a very much longer time scale, the stimulus would look stationary. But a good stochastic model on a long time scale would likely require long-range dependence. Indeed, it can be difficult to distinguish non-stationarity from long-range dependence (Künsch, 1986), and the usual statistical properties of estimators are known to breakdown when long-range dependence is present (Beran, 1994). Given a short interval of data, valid statistical inference under stationarity assumptions becomes highly problematic. To avoid these problems we have proposed the use of the divergence plot, and a recognition that the "bits per second" summary is no longer mutual

information in the usual sense. Instead we would say that the estimate of information measures magnitude of variation of the response as the stimulus varies, and that this is a useful assessment of the extent to which the stimulus affects the response as long as other factors that affect the response are themselves time-invariant. In other deterministic or non-stationary settings the argument for the relevance of an information estimate should be analogous. Under stationarity and ergodicity, and indefinitely many trials, the stimulus sets that affect the response—whatever they are—will be repeatedly sampled, with appropriate probability, to determine the variability in the response distribution, with time-invariance in the response being guaranteed by the joint stationarity condition. This becomes part of the intuition behind mutual information. In the deterministic or non-stationary settings information estimates do not estimate mutual information, but they may remain intuitive assessments of strength of effect.

3.5 Proofs

We will use the following extension of the Lebesgue Dominated Convergence Theorem in the proof of Theorem 3.2.

Lemma 3.3. *Let f_m and g_m for $m = 1, 2, \ldots$ be sequences of measurable, integrable functions defined on a measure space equipped with measure μ, and with pointwise limits f and g, respectively. Suppose further that $|f_m| \leq g_m$ and $\lim_{m\to\infty} \int g_m \, d\mu = \int g \, d\mu < \infty$. Then*

$$\lim_{m\to\infty} \int f_m \, d\mu = \int \lim_{m\to\infty} f_m \, d\mu.$$

Proof. By linearity of the integral,

$$\liminf_{n\to\infty} \int (g + g_m) \, d\mu - \limsup_{n\to\infty} \int |f - f_m| \, d\mu = \liminf_{n\to\infty} \int (g + g_m) - |f - f_m| \, d\mu.$$

Since $0 \leq (g + g_m) - |f - f_m|$, Fatou's Lemma implies

$$\liminf_{n\to\infty} \int (g + g_m) - |f - f_m| \, d\mu \geq \int \liminf_{n\to\infty} (g + g_m) - |f - f_m| \, d\mu.$$

The limit inferior on the inside of the right-hand integral is equal to $2g$ by assumption. Combining with the previous two displays and the assumption that $\int g_m \, d\mu \to \int g \, d\mu$ gives

$$\limsup_{n\to\infty} \left| \int f d\mu - \int f_m d\mu \right| \leq \limsup_{n\to\infty} \int |f - f_m| d\mu \leq 0. \qquad \blacksquare$$

Proof of Theorem 3.2. The main statement of the theorem is implied by the three numbered statements together with Proposition 3.1. We start with the second numbered statement. Under the repeated trial assumption, $X_t^1, \ldots, X_t^m$ are conditionally i.i.d. given the stimulus

$\{S_t\}$. So Corollary 1 of Antos and Kontoyiannis (2001), can be applied to show that

$$\begin{aligned}\lim_{m\to\infty} \hat{H}_t &= \lim_{m\to\infty} -\sum_r \hat{P}_t(r)\log\hat{P}_t(r) \\ &= -\sum_r P_t(r|S_1,\ldots,S_n)\log P_t(r|S_1,\ldots,S_n) \\ &= H(P_t)\end{aligned} \tag{3.5.1}$$

with probability 1. This proves the first numbered statement.

We will use Lemma 3.3 to prove the first numbered statement. For each r the law of large numbers asserts $\lim_{m\to\infty}\hat{P}_t(r) = P_t(r|S_1,\ldots,S_n)$ with probability 1. So for each r,

$$\lim_{m\to\infty} -\hat{P}_t(r)\log\hat{P}(r) = -P_t(r|S_1,\ldots,S_n)\log\bar{P}(r|S_1,\ldots,S_n)$$

and

$$\lim_{m\to\infty} -\hat{P}_t(r)\log\hat{P}_t(r) = -P_t(r|S_1,\ldots,S_n)\log P_t(r|S_1,\ldots,S_n)$$

with probability 1. Fix a realization where the above limits hold and let

$$f_m(r) := -\hat{P}_t(r)\log\hat{P}(r)$$

and

$$g_m(r) := -\hat{P}_t(r)[\log\hat{P}_t(r) - \log n].$$

Then for each r

$$\lim_{m\to\infty} f_m(r) = -P_t(r|S_1,\ldots,S_n)\log\bar{P}(r|S_1,\ldots,S_n) =: f(r)$$

and

$$\lim_{m\to\infty} g_m(r) = -P_t(r)[\log P_t(r) - \log n] =: g(r).$$

The sequence f_m is dominated by g_m because

$$\begin{aligned} 0 \leq -\hat{P}_t(r)\log\hat{P}(r) &= f_m(r) \\ &= -\hat{P}_t(r)[\log\sum_{u=1}^{n}\hat{P}_u(r) - \log n] \\ &\leq -\hat{P}_t(r)[\log\hat{P}_t(r) - \log n] \qquad (3.5.2) \\ &= g_m(r) \end{aligned}$$

for all r, where Equation (3.5.2) uses the fact that $\log x$ is an increasing function. From Equation (3.5.1) we also have that $\lim_{m\to\infty}\sum_r g_m(r) = \sum_r g(r)$. Clearly, f_m and g_m are summable. Moreover $H(P_t) < \infty$ by assumption. So

$$\sum_r g(r) = \sum_r -P_t(r)\log P_t(r) + \log n\sum_r P_t(r) = H(P_t) + \log n < \infty$$

and the conditions of Lemma 3.3 are satisfied. Thus

$$\lim_{m\to\infty}\sum_r -\hat{P}_t(r)\log\hat{P}(r) = \lim_{m\to\infty}\sum_r f_m(r) = \sum_r f(r) = \sum_r -P_t(r)\log\bar{P}(r). \qquad (3.5.3)$$

Averaging over $t = 1, \ldots n$ gives

$$\hat{H} = \lim_{m\to\infty}\sum_r -\hat{P}(r)\log\hat{P}(r) = \sum_r -\bar{P}(r)\log\bar{P}(r) = H(\bar{P}).$$

This proves the first numbered statement.

For the third numbered statement we begin with the expansions

$$\hat{D}(P_t||\bar{P}) = \sum_r \hat{P}_t(r)\log\hat{P}_t(r) - \hat{P}_t(r)\log\hat{P}(r).$$

and

$$D(P_t||\bar{P}) = \sum_r P_t(r)\log P_t(r) - P_t(r)\log\bar{P}(r).$$

The second numbered statement and Equation (3.5.3) imply

$$\lim_{m\to\infty} \sum_r \hat{P}_t(r) \log \hat{P}_t(r) - \hat{P}_t(r) \log \hat{P}(r) = \sum_r P_t(r) \log P_t(r) - \sum_r P_t(r) \log \bar{P}(r)$$

with probability 1. This proves the third numbered statement. ■

Part II

Regularized Regression

Chapter 4

Sparse Nonparametric Regression of V1 fMRI on Natural Images

4.1 Introduction

There are two main classes of problems in computational neuroscience: *encoding* and *decoding*. Encoding deals with the forward problem of modeling the neural response as a function of the sensory stimulus. Decoding is the inverse problem of predicting the stimulus from the neural response. In this chapter we focus on the encoding problem. Specifically, we model brain activity in area V1 of the human visual cortex, as measured by functional magnetic resonance imaging, in response to natural image stimuli.

4.1.1 Functional MRI

Functional magnetic resonance imaging (fMRI) provides an indirect measure of brain activity. Magnetic resonance imaging is based on how certain nuclei such as hydrogen

atoms align under magnetic fields. By modulating the field, these aligning behaviors can then be detected in a radio-frequency coil. When neurons in the brain are active, there is an increased oxygenated blood flow to that neuron. Oxygenated hemoglobin in the blood is diamagnetic and aligns against a surrounding magnetic field while deoxygenated hemoglobin is paramagnetic and aligns with a magnetic field. This difference in magnetic susceptibility is then captured by a sequence of magnetic resonance images that constitute the Blood Oxygen Level Dependent (BOLD) fMRI signal. The relationship between measured fMRI activity and the spiking activity of neurons is thus not direct. The fMRI images of the brain are typically taken at the rate of one or two per second, where each image comprises values in a three-dimensional grid of volume elements called voxels, each of size a few cubic millimeters.

4.1.2 Area V1

The visual cortex is located in the occipital lobe at the back of the brain, and consists of several areas including V1 through V5. The focus of this chapter is area V1, also called the primary visual cortex. V1 consists mainly of two types of neurons: simple cells and complex cells. An important characterization of a neuron is its spatial *receptive field*, which is the region of visual space where a stimulus can affect the firing response of that neuron. Further, it is only stimuli with specific properties that can affect the neural response. Hubel and Wiesel (1962), Daugman (1985) and others observed that stimuli with an elongated envelope of high luminance, flanked on both sides by an elongated envelope of low luminance, or vice vera, are the most excitatory stimuli for simple cells. As regards complex cells, Hubel and Wiesel (1962) and others observed that their spatial receptive

fields were larger than that of simple cells, and that their responses were mostly invariant to the spatial phase of stimuli.

4.1.3 The Data

The data set analyzed in this chapter was obtained from the fMRI experiment described in Kay et al. (2008). It consists of a total of 1,294 voxels, each of size 2mm x 2mm x 2.5mm, from area V1 of one human subject. The BOLD signals were recorded from voxels at a frequency of 1Hz using a 4T Varian MRI scanner. The sensory stimuli used in the experiment consisted of 20-by-20 degree grayscale natural images, masked by a circular aperture. (See Figure 4.1 for some instances.) The motivations behind studying responses to natural image stimuli are three-fold. Firstly, natural images are an efficient stimulus set for probing the visual system, because they are likely to evoke responses from both early visual areas as well as from more central, highly nonlinear visual areas. Secondly, if the neural response is nonlinear in the image stimuli, then the response to a complex stimulus such as a natural image cannot be expressed in terms of responses to simpler stimuli such as gratings. Finally, natural images are a strict subset of the set of all images and the brain has likely evolved to effectively respond to these. Two recent observations provide evidence of such adaptation. Olshausen and Field (1996) computed the optimal basis to encode natural images, and found these to resemble Gabor wavelets. Daugman (1985) had in turn observed that Gabor wavelets resembled the receptive fields of simple cell neurons in area V1. Thus, it would be most interesting to study the fMRI responses of the brain to the natural image stimuli that it has adapted to.

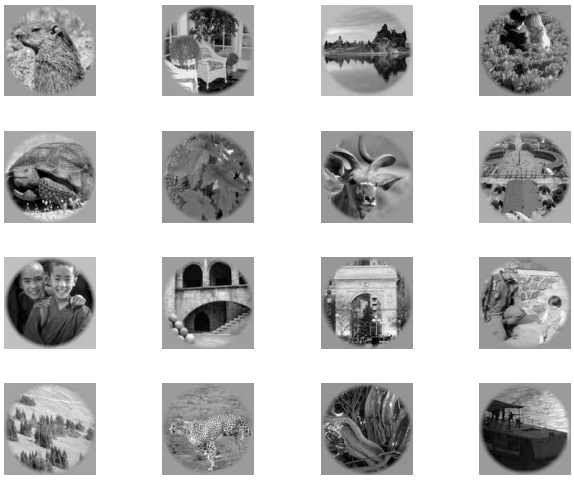


Figure 4.1: Some natural image stimuli.

The fMRI datasets for model estimation and model validation were collected separately. 1,750 natural images were used as stimuli for the model estimation data set, while a distinct set of 120 natural images were used for the model validation data set. The images were shown in a sequence of *trials*, each of which consisted of the image being flashed briefly 3 times during a 1 second display period, followed by a blank period of 3 seconds. For the estimation data set, each image was shown in two trials over the sequence of trials, while in the validation data set each image was shown in thirteen trials over the sequence. Further details of the fMRI experiment can be found in Kay et al. (2008).

Pre-processing

For each of the model estimation and validation dataset, at the end of the image trial sequences we obtain the fMRI signal as a time-series recorded at 1Hz, at each of the 1294 voxels. This raises two main pre-processing concerns. The first is that if each distinct image stimulus has a separate time-series BOLD response, also called a hemodynamic response

function, then the observed time-series signal consists of a superposition of all these separate response functions. It is thus necessary to "deconvolve" these separate response function signals from the observed combination signal. The second is that the fMRI signal has a smooth "drift" throughout the experiment, often at magnitude larger than task-related signal. The drift is due to varied factors such as the fMRI machine and steady variations in blood pressure. The basis-restricted separable model addresses both of these concerns (Dale, 1999; Kay et al., 2008). Deferring details to Section 4.8, at the end of the pre-processing, for each image stimulus, we obtain a scalar response amplitude at each voxel. In other words, both the model estimation and validation datasets consist of a set of image stimuli used as input, and as output for each image stimulus there is a scalar response amplitude at each of the 1,294 voxels.

4.2 Previous Work

4.2.1 Models of Area V1 Neurons

Daugman (1985) had observed that the receptive fields of simple cells resembled spatial *Gabor* wavelets. Interestingly, Gabor wavelets were also found to be an optimal basis for encoding natural images (Olshausen and Field, 1996). A Gabor wavelet can be written (up to centering and scaling) as the product of a bivariate Gaussian density and a spatial sinusoidal function (or cosine grating),

$$\phi(\vec{x}) \propto \exp\{-\omega^2(\vec{x}-\mu)^T K(\vec{x}-\mu)/2\}\cos(\omega\langle\theta,\vec{x}-\mu\rangle+\psi), \tag{4.2.1}$$

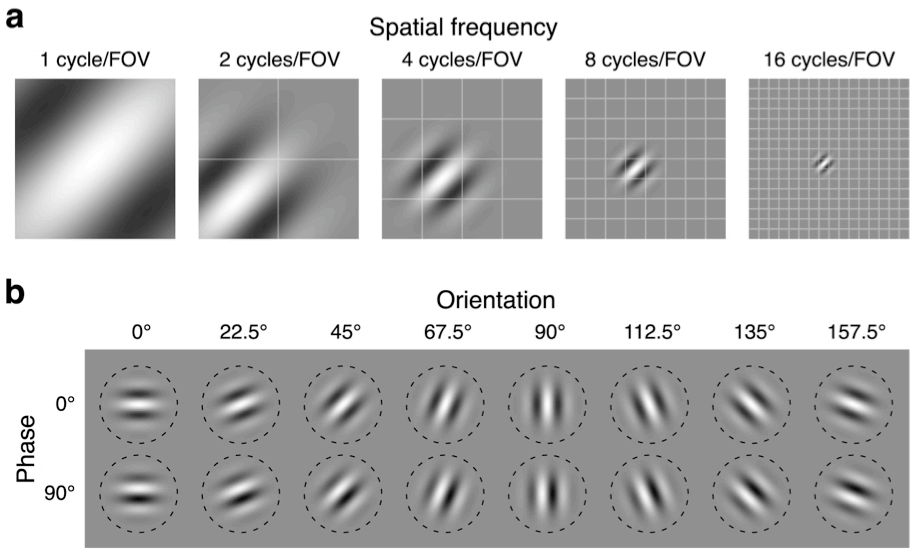


Figure 4.2: (a) Gabor wavelets identical up to location and frequency (in cycles per field-of-view). (b) Gabor wavelets with the same spatial frequency, but of differing orientations and phases.

where $\vec{x} = (x, y)$ are spatial coordinates, and $\langle \cdot, \cdot \rangle$ is the Euclidean inner product. The Gaussian density acts as a window that localizes the wavelet in space, while the sinusoid localizes the wavelet in the spatial Fourier domain. So the parameters (ω, θ, ψ) determine the wavelet's frequency, orientation, and phase, while the parameters (μ, K) determine the spatial location and envelope of the window. Figure 4.2 shows some examples of Gabor wavelets.

Simple cells

Hubel and Wiesel (1962), Movshon et al. (1978a) and others have shown that simple cell outputs are roughly linear in the image stimuli intensities. Daugman (1985) showed further that the corresponding set of linear combination weights, as a linear filter,

resemble Gabor wavelets. A simple cell is thus typically modeled as follows (Kay et al., 2008). Let I denote an image represented as a vector of pixel intensities. Suppose the number of pixels is d so that $I \in \mathbb{R}^d$. Denote by ϕ_j a mean 0, norm 1, Gabor wavelet sampled on a grid the size of the image, so that it too can be represented as vector in $\mathbb{R}^d$. Then the model simple cell neural activation given an image I is

$$X_j(I) = [\langle \phi_j, I \rangle]_+, \tag{4.2.2}$$

where $[\cdot]_+$ is a non-negative rectification, or thresholding. (See Figure 4.3.) Correspondingly, $X_j(I) = [\langle -\phi_j, I \rangle]_+$ gives the activation of the 180° spatial phase counterpart. The rectification obtains a firing rate from the internal signal of the neuron.

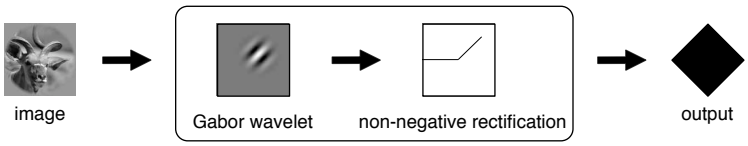


Figure 4.3: Simple cell model. The activation of a model simple cell given an image is the inner product of the image with a Gabor wavelet, followed by a non-negative rectification.

Complex cells

Complex cells were initially identified as being nonlinear in the image stimulus. Hubel and Wiesel (1962), Movshon et al. (1978b) and others observed that, like simple cells, complex cells are selective for spatial frequency and orientation of the stimulus, however their response is invariant to the stimulus' spatial phase. This property can be modeled by computing the sum of squares of the outputs of four simple cells that are identical up to spatial phase, followed by a monotonically increasing output nonlinearity $h : [0, \infty) \mapsto$

$[0, \infty)$ (Kay et al., 2008). Thus the activation of the model complex cell given an image I is defined as

$$X_j(I) = h\left([\langle \phi_j, I\rangle]_+^2 + [\langle -\phi_j, I\rangle]_+^2 + [\langle \phi'_j, I\rangle]_+^2 + [\langle -\phi'_j, I\rangle]_+^2\right), \tag{4.2.3}$$

where ϕ_j and ϕ'_j are Gabor wavelets identical up to phase (also called a quadrature pair; see Figure 4.4).

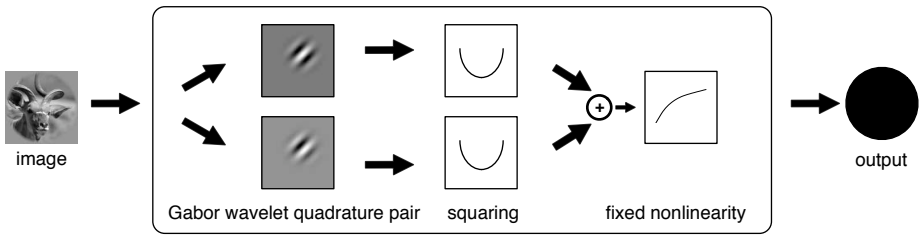


Figure 4.4: Complex cell model. The activation of a model complex cell given an image is the sum of squares of 4 simple cell model outputs, followed by a nonlinearity.

4.2.2 Linear Models

Let p_c denote the number of complex cell filters, $\{X_1, \ldots, X_{p_c}\}$. Kay et al. (2008) modeled the fMRI response $Y(I)$ to a stimulus image I as a linear model,

$$Y(I) = \beta_0 + \sum_{j=1}^{p_c} \beta_j X_j(I) + \epsilon. \tag{4.2.4}$$

In our experiments, we used as a benchmark a sparse variant of this model with the nonlinearity fixed to be $h(t) = \log(1 + \sqrt{t})$. The sparse variant constrains the set $J = \{j : \beta_j \neq 0\}$ to be small, and captures the empirical observation that the response at any voxel is receptive to stimuli in a small specific spatial region, so that the set of active complex cells at a single voxel would be sparse. Our choice of h was based preliminary investigations.

We also experimented with other regularized linear models, including ridge regression, but these had similar performance to the sparse linear model (data not shown). Intuitions regarding fMRI responses first led us to what we call the V-SPAM model, and then the V-iSPAM model, which in the next section we will present as instances of a single class of models called the V-SPAM framework.

4.3 The V-SPAM Framework

Our framework for predicting the voxel response amplitude builds on two intuitions about fMRI signals. The first is that each voxel corresponds to an approximately 2mm cube of neural tissue, and hence any measurement at the level of a voxel reflects the pooled activity of many neurons. Secondly, the fMRI BOLD signal is an indirect measurement of neural activity in terms of blood flows, so that neural activity is reflected in the fMRI signal only after being subjected to potentially nonlinear transformations.

The V-SPAM framework attempts to mimic this process by a two stage scheme. The first stage is a biologically-inspired hierarchical filtering scheme that consists of three distinct layers of artificial or model neurons, arranged hierarchically: simple cells, complex cells, and linear combinations of the complex cells (here called pooled-complex cells). The output of this filtering stage is then fed to the next "pooling" stage where transformations are applied to a sparse subset of the filter outputs before a final summation producing a prediction of the fMRI response. This process parallels the transformation from neural activity to hemodynamic signals.

4.3.1 Filtering Stage

The first two layers of the filtering stage comprise simple and complex cells, models for which were discussed in Section 4.2.1 and Section 4.2.1 respectively.

Pooled-complex cell model

A pooled-complex cell is simply a linear combination of certain complex cells. Let $\{X_{j_1}, \ldots, X_{j_m}\}$ denote the activations of a set of m complex cells. The corresponding pooled-complex cell activation $Z_{j_1,\ldots,j_k}$ given an image I is then given by the simple linear combination,

$$Z_{j_1,\ldots,j_m}(I) = \sum_{r=1}^{k} X_{j_r}(I).$$

In Section 4.2.1 we saw how complex cells combine simple cells that share the same spatial frequency, location and orientation but have differing phases, so that the resulting complex cell is phase-invariant. We can further combine complex cells that share the same spatial location and frequency but differ in their orientations, so that the resulting *pooled-complex* cell is orientation-invariant. Figure 4.5 shows an instance of such an orientation-invariant pooled-complex cell. While cells with such receptive fields might have direct biological interpretation only in visual areas higher than V1, we have nonetheless included these in order to improve the representational power of the model. The reason why these cells improve representational power inspite of a final pooling stage will be clarified after the description of the final pooling stage.

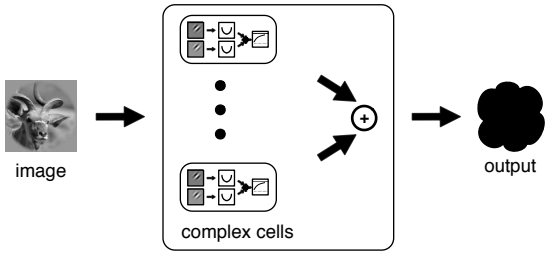


Figure 4.5: Pooled-complex cell model. Subsets of complex cells that share a common spatial location and frequency are summed.

4.3.2 Pooling Stage

In the final pooling stage a sparse subset amongst the set of all complex and pooled-complex cell outputs are nonlinearly transformed, and then summed to model the fMRI response Y. Let p_c and p_l denote the total number of complex and pooled-complex cells in the filtering stage respectively. Denote the corresponding outputs of the model complex cells by $\{X_1, \ldots, X_{p_c}\}$, and those of the model pooled-complex cells by $\{Z_1, \ldots, Z_{p_l}\}$. The fMRI response $Y(I)$ to image I is then modeled as,

$$Y(I) = \sum_{j=1}^{p_c} f_j(X_j(I)) + \sum_{r=1}^{p_l} g_r(Z_l(I)) + \epsilon, \tag{4.3.1}$$

where $\{f_j\}_{j=1}^{p_c}$ and $\{g_r\}_{r=1}^{p_l}$ are sets of unidimensional functions that have to be estimated from data. That only a sparse subset amongst the set of all filtering stage outputs are to be selected is captured by an additional assumption that the sets $J_1 = \{j \in \{1, \ldots, p_c\} : f_j \not\equiv 0\}$ and $J_2 = \{r \in \{1, \ldots, p_c\} : g_r \not\equiv 0\}$ are small.

We can now see why pooled-complex cells improve the representation capacity of the model. A nonlinear transformation of a pooled-complex cell—a linear combination of complex cells—cannot be expressed as a linear combinations of nonlinear transformations of individual complex cells alone. Thus the V-SPAM pooling Equation (4.3.1) can represent

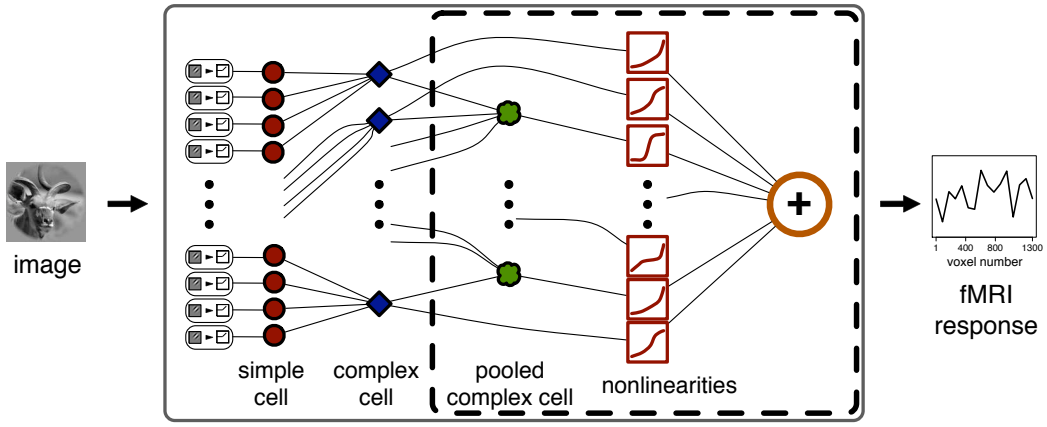


Figure 4.6: The V-SPAM model. The fMRI voxel response is modeled as the summation of nonparametric functions of complex and pooled-complex cell outputs. The connections and components in the dashed region are to be estimated from the data under the assumption that many of them are null.

a function than cannot be expressed without the pooled-complex cells.

4.3.3 V-SPAM Model

The V-SPAM model is an instance of the above framework where the functions $\{f_j\}_{j=1}^{p_c}$ and $\{g_r\}_{r=1}^{p_l}$ are allowed to be distinct, as shown in Figure 4.6. In Section 4.4.1, we will describe the class of nonparametric regression models called *sparse additive models (SPAM)* that model the response as an additive combination of nonparametric transformations of a sparse subset of the predictors. Thus the V-SPAM model relates the fMRI response as a sparse additive model of complex and pooled-complex cell outputs.

4.3.4 V-iSPAM Model

If we assume the nonlinear transformations of the individual neurons are the result of the same underlying causes, then we would at least want to constrain the individual trans-

formations to have the same form. In the V-iSPAM model, we assume that the functions are identical up to a scale factor and model the fMRI response $Y(I)$ to image I as

$$Y(I) = \sum_{j=1}^{p_c} \alpha_j f(X_j(I)) + \sum_{r=1}^{p_l} \beta_r f(Z_l(I)) + \epsilon, \tag{4.3.2}$$

where f is the unidimensional function that represents the identical nonlinear transformation, while the weights $\{\alpha_j\}_{j=1}^{p_c}$ and $\{\beta_r\}_{r=1}^{p_l}$ are individual to each model neuron. That a sparse subset of the model neurons is selected is reflected in an additional assumption that the set $J_1 = \{j \in \{1, \ldots, p_c\} : \alpha_j \neq 0\}$ and $J_2 = \{r \in \{1, \ldots, p_c\} : \beta_r \neq 0\}$ are small. Thus the individual neural responses are all transformed by the same function f, but there is a separate scalar gain control parameter for each neural output.

In Section 4.4.2 we will describe the class of nonparametric regression models called *identical sparse additive models* which model the response as a sparse additive combination of scalings of a single nonparametric transformation of the individual predictors. Thus, the V-iSPAM model relates the fMRI response as an identical sparse additive model of complex and pooled-complex cell outputs. Figure 4.7 summarizes the V-iSPAM model. The only difference from the V-SPAM model (Figure 4.6) is that the nonlinear transformations in the final pooling stage are the same except perhaps for scale.

4.4 Some Nonparametric Regression Models

4.4.1 Sparse Additive Models

Recall that in the final pooling stage of VSPAM, we simultaneously select a sparse subset of all predictors and learn nonparametric transformations of these selected predic-

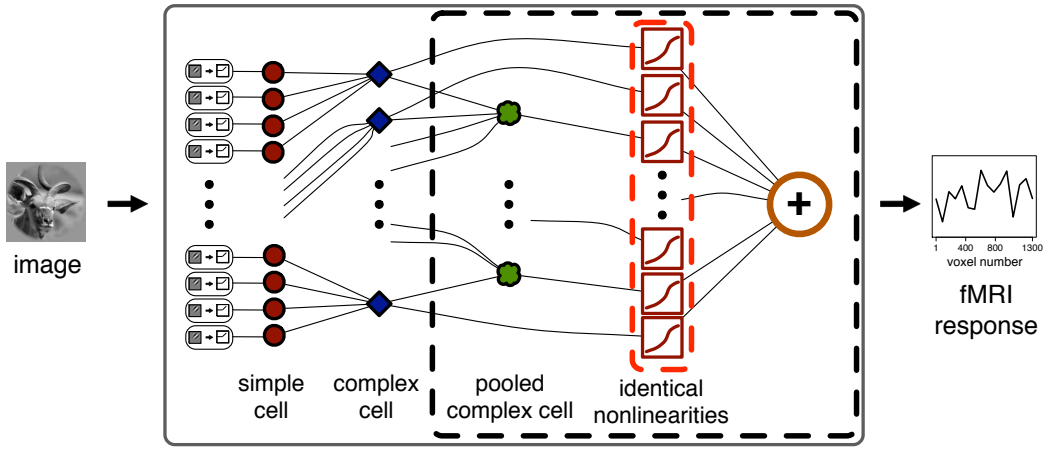


Figure 4.7: The V-iSPAM model. The fMRI voxel response is modeled as the summation of nonlinear functions of complex and pooled-complex cell outputs. The nonlinear functions share a common form. The connections and components in the dashed region are to be estimated from the data under the assumption that many of them are null.

tors. This is precisely the setting of sparse additive models (Ravikumar et al., 2007). Suppose $X = (X_1, \ldots, X_p) \in \mathbb{R}^p$ is a p-dimensional predictor, with marginal density functions $p_{X_1}(x), \ldots, p_{X_p}(x)$, and $Y \in \mathbb{R}$ is a real-valued, mean 0 response. In general multivariate nonparametric regression, the response Y is related to the predictor X as

$$Y = m(X) + \epsilon,$$

where $m : \mathbb{R}^p \mapsto \mathbb{R}$ is the regression function, and ϵ is independent zero mean noise, for instance $\epsilon \sim N(0, \sigma^2)$. In additive models, introduced by Hastie and Tibshirani (1999), the regression function is constrained to be an additive combination of functions of individual predictors,

$$Y = \sum_{j=1}^{p} f_j(X_j) + \epsilon,$$

so that $f_j \in L_2(P(X_j))$ is a uni-dimensional real-valued function. A sparse additive model (SPAM) (Ravikumar et al., 2007), constrains this further by assuming that the set $J =$

$\{j \in \{1, \ldots, p\} \,|\, f_j(X_j) \not\equiv 0\}$, of individual predictor functions that are not identically zero, is sparse. Suppose we are given n independent observations of the predictor and response, $\{(X_i, Y_i)\}_{i=1}^n$. In high-dimensional settings, where p is large relative to the sample size n, estimating even linear models such as $Y = \beta^\top X + \epsilon$ is challenging. However when the true coefficient vector β of the linear model is sparse, Wainwright (2006); Zhao and Yu (2006) and others have shown that the ℓ_1 penalized least squares estimator, also called the Lasso (Tibshirani, 1996) defined as,

$$\hat{\beta} = \arg\min_{\beta} \sum_i (Y_i - \beta^\top X_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|,$$

can estimate the linear model consistently under certain conditions even in high-dimensional settings. Sparse additive models (Ravikumar et al., 2007) extend these sparse linear models to the nonparametric domain.

Fitting Algorithm for Sparse Additive Models

Ravikumar et al. (2007) propose an iterative procedure to fit a sparse additive model to data that reduces the multivariate additive regression problem to a series of one-dimensional nonparametric regression problems. The procedure makes extensive use of nonparametric estimators known as *linear smoothers*. A nonparametric estimator $\hat{m}(x)$ of the regression function $m(x) = \mathbb{E}(Y|X = x)$ is called a linear smoother if $\hat{m}(x)$ is a linear combination of the training samples $Y_1, \ldots, Y_n$ for each fixed x. We will assume we are given a set of linear smoothers for regressing any response upon the individual predictors $\{X_j\}$. Consider predictor X_j in the sparse additive model. Given any response $Z \in \mathbb{R}$ and n samples $\{Z^{(i)}, X_j^{(i)}\}_{i=1}^n$, the smoothed response from a linear smoother would then

Input: Data (X_i, Y_i), regularization parameter λ.

Initialize $\hat{f}_j = 0$, for $j = 1, \ldots, p$.

Iterate until convergence:

For each $j = 1, \ldots, p$:

Compute the residual $\mathrm{R}_j(X, Y, \hat{f}_j) = Y - \sum_{k \neq j} \hat{f}_k(X_k)$ at the data points: $\hat{\mathrm{R}}_j = \{\mathrm{R}_j(X^{(i)}, Y^{(i)}, \hat{f}_j)\}_{i=1}^n$.

Estimate the conditional expectation $\mathrm{P}_j = \mathbb{E}[\mathrm{R}_j | X_j]$ by smoothing: $\hat{\mathrm{P}}_j(x) = \mathcal{S}_j(x)\hat{\mathrm{R}}_j$.

Set $\hat{s}_j{}^2 = n^{-1} \sum_{i=1}^n \hat{\mathrm{P}}_j{}^2(X_j^{(i)})$.

Soft-threshold: $\hat{f}_j = [1 - \lambda/\hat{s}_j]_+ \hat{\mathrm{P}}_j$.

Center: $\hat{f}_j \leftarrow \hat{f}_j - \mathrm{mean}(\hat{f}_j)$.

Output: Component functions $\hat{f}_j$ and estimator $\hat{m}(X) = \sum_j \hat{f}_j(X_j)$.

Figure 4.8: The SPAM Fitting Algorithm

be $\mathcal{S}_j(x)\hat{Z}$, where $\hat{Z}$ is a column vector of the n samples of the response Z and $\mathcal{S}_j(x)$ is a row-vector of the linear combination weights, that only depends on the samples of the predictor X_j. With this notation, we can summarize the sparse backfitting procedure of Ravikumar et al. (2007) for fitting a sparse additive model to data in Figure 4.8.

At each iteration the algorithm cycles through the predictors. At each predictor, it computes a residual for that predictor, nonparametrically regresses the residual onto that predictor, and soft thresholds the resulting function.

4.4.2 iSPAM: Identical Sparse Additive Models

In this section, we study a generalization of sparse additive models where the functions on the individual predictors are constrained to be identical up to scaling. To facilitate working with a function defined on all the predictor variables, let $\bar{P}(X)$ denote the uniform mixture of the marginal distributions of $X_1, \ldots, X_p$. The response Y given

$X_1, \ldots, X_p$ is modeled as

$$Y = \sum_{j=1}^{p} \alpha_j f(X_j) + \epsilon, \tag{4.4.1}$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and $f \in L_2(\bar{P}(X))$. So the conditional mean of Y given X is linear in the transformed values $f(X_1), \ldots, f(X_p)$. If f was fixed to be $f(x) = x$, this model would reduce to the usual linear regression model. However, we assume f to be unknown.

Algorithm, Population Version

The standard optimization problem for this regression model in the population setting is

$$\min_{\alpha \in \mathbb{R}^p, f \in L_2(\bar{P}(X))} \mathbb{E}(Y - \sum_{j=1}^{p} \alpha_j f(X_j))^2,$$

where the expectation is with respect to the noise ϵ and the predictor vector X. This optimization problem does not have a unique solution because if (α, f) is a solution then so is $(c\alpha, c^{-1}f)$ for any $c \neq 0$. This obstacle can be overcome by imposing constraints on α and f. First we fix the scale of the function f by constraining the norm of f. Then the weights $\{\alpha_j\}$ serve as the scale of the effect of each predictor, and we can thus impose sparsity on the set of active predictors by an ℓ_1 constraint on the weights $\{\alpha_j\}$. This leads to the constrained optimization problem

$$\begin{aligned} \inf_{\alpha \in \mathbb{R}^p, f \in L_2(\bar{P}(X))} \quad & \mathbb{E}(Y - \sum_{j=1}^{p} \alpha_j f(X_j))^2 \\ \text{s.t.} \quad & \sum_j |\alpha_j| \leq C_1 \\ & \sum_j \mathbb{E}(f^2(X_j)) \leq C_2, \end{aligned}$$

and the corresponding penalized Lagrangian form is

$$\inf_{\alpha\in\mathbb{R}^p, f\in L_2(\bar{P}(X))} L(f,\alpha;\lambda,\mu) \equiv \frac{1}{2}\mathbb{E}(Y-\sum_j \alpha_j f(X_j))^2 + \lambda\sum_j |\alpha_j| + \frac{1}{2}\mu\sum_j \mathbb{E}(f^2(X_j)). \tag{4.4.2}$$

For each $j \in \{1,\ldots,p\}$, denote the residual of the identical sparse additive model with the j-th predictor removed as $\mathrm{R}_j(X,Y,f) = Y - \sum_{k\neq j} f(X_k)$. Then a solution (f^*,α^*) of the problem Equation (4.4.2) can be characterized by the following theorem.

Theorem 4.1 (Characterization of population optima)**.** *An optimal solution* $(f^*,\alpha^*) \in (L_2(\bar{P}(X))\times\mathbb{R}^p)$ *of the regularized population likelihood problem Equation* (4.4.2) *is a fixed point of the following joint updates:*

$$f(x) = \sum_j \left\{ \frac{\alpha_j\, p_{X_j}(x)}{\sum_k (\alpha_k^2+\mu)\, p_{X_k}(x)} \right\} \mathbb{E}(R_j | X_j = x) \tag{4.4.3a}$$

$$\alpha = \arg\min_{\alpha\in\mathbb{R}^p} \mathbb{E}(Y - \sum_{j=1}^p \alpha_j f(X_j))^2 + \lambda\sum_j |\alpha_j| \tag{4.4.3b}$$

A closer look at the updates

The update Equation (4.4.3b) for the weights is just an ℓ_1-regularized least squares problem. Let us now look at the updates for the function f. For each $j = 1,\ldots,p$, let w_j denote the function

$$w_j(x) = \frac{\alpha_j\, p_{X_j}(x)}{\sum_{k=1}^p (\alpha_k^2+\mu)\, p_{X_k}(x)}.$$

Then ((4.4.3a)) can be rewritten as,

$$f(x) = \sum_{j=1}^p w_j(x)\, \mathbb{E}(Y - \sum_{k\neq j} \alpha_k f(X_k) | X_j = x). \tag{4.4.4}$$

Linear combination of backfitted functions Consider the classical additive model, $Y = \sum_j f_j(X_j) + \epsilon$. Hastie and Tibshirani (1999) proposed the following iterative algorithm, called backfitting, for solving the corresponding least squares problem in the population setting. They proposed iterating over the predictors $j = 1, \ldots, p$ with the update,

$$f_j(x) \leftarrow \mathbb{E}(Y - \sum_{k \neq j} f_k(X_k) | X_j = x).$$

Comparing this with equation Equation (4.4.4), we see that the function update is equivalent to a two-staged procedure that first computes distinct backfitted functions at each predictor, as with vanilla additive models, and then sets the function f to a weighted linear combination of these individual predictor functions. Thus denoting $\tilde{f}_j(x) = \mathbb{E}(Y - \sum_{k \neq j} \alpha_k f(X_k) | X_j = x)$, the function update Equation (4.4.4) can be written as

$$f(x) = \sum_{k=1}^{p} w_j(x)\, \tilde{f}_j(x), \tag{4.4.5}$$

whereas vanilla backfitting would merely compute the functions $\tilde{f}_j$ at each predictor X_j.

Convergence of the function updates The update from equation Equation (4.4.4) can be rewritten as

$$f(x) = \sum_{j=1}^{p} w_j(x) \mathbb{E}(Y | X_j = x) - \sum_{k=1}^{p} \sum_{j \neq k} w_j(x) \alpha_k \mathbb{E}(f(X_k) | X_j = x). \tag{4.4.6}$$

Let $b \in L_2(\bar{P}(X))$ denote the function $b(x) = \sum_{j=1}^{p} w_j(x) \mathbb{E}(Y | X_j = x)$. For each $j, k \in \{1, \ldots, p\}$, $j \neq k$, let $\mathrm{P}_{jk} : L_2(\bar{P}(X)) \mapsto L_2(\bar{P}(X))$ be the linear operator defined as $(\mathrm{P}_{jk} f)(x) = \mathbb{E}(f(X_k) | X_j = x)$. With this notation, equation Equation (4.4.6) can be

written as,

$$f = b - \sum_{k=1}^{p} \alpha_k (\sum_{j \neq k} w_j \mathrm{P}_{jk}) f$$

$$= b - \mathcal{A} f,$$

where $\mathcal{A} : L_2(\bar{P}(X)) \mapsto L_2(\bar{P}(X))$ is the linear operator $\mathcal{A} = \sum_{k=1}^{p} \alpha_k (\sum_{j \neq k} w_j \mathrm{P}_{jk})$. Note that if $f^{(0)} = 0$, and $f^{(t+1)} = b - \mathcal{A} f^{(t)}$, it follows by induction that

$$f^{(t)} = \sum_{j=0}^{t-1} (-1)^j \mathcal{A}^j b.$$

We prove in the Appendix that

Lemma 4.2. *For all* $f \in L_2(\bar{P}(X))$

$$\int |(\mathcal{A} f)(x)|^2 \bar{P}(dx) < \frac{p||\alpha||^2}{\mu} \int |f(x)|^2 \bar{P}(dx)$$

and so $\mathcal{A}$ *is a contraction if* $\mu > p||\alpha||^2$.

Thus, if we set μ large enough so that $\mu > p||\alpha||^2$, then $(I + \mathcal{A})$ is non-singular and

$$f^{(t)} = (I + \mathcal{A})^{-1}(I - (-1)^t \mathcal{A}^t) b \to (I + \mathcal{A})^{-1} b$$

in $L_2(\bar{P}(X))$ as $t \to \infty$. So the iterates $f^{(t)}$ converge to the solution of the linear system $f = b - \mathcal{A} f$.

4.4.3 Empirical iSPAM Algorithm

Theorem 4.1 derives the true function f and the weights α as fixed points of a set of updates. However these updates depend on population quantities such as expectations

with respect to the population distribution, and the true predictor densities. Given samples $\{(X^{(i)}, Y^{(i)})\}_{i=1}^n$, we can then formulate a data version of these updates by plugging in sample-based counterparts of these population quantities. Thus, we estimate the conditional expectation of the j-th residual $\mathrm{R}_j(X, Y, \hat{f}) = Y - \sum_{k \neq k} \hat{\alpha}_j \hat{f}_k(X_k)$ given the j-th predictor, $\mathbb{E}(R_j | X_j)$, by smoothing. Let $\hat{\mathrm{R}}_j = (\mathrm{R}_j(X^{(1)}, Y^{(1)}, \hat{f}), \ldots, \mathrm{R}_j(X^{(n)}, Y^{(n)}, \hat{f}))^T$ be a column vector of the j-th residual evaluated at the data points. Then as in Section 4.5, assuming the smoother for predictor X_j is a linear smoother and denoting $\mathcal{S}_j(x)$ as the row-vector of the smoother weights that only depend on the samples of the predictor X_j, we estimate $\mathbb{E}(R_j | X_j = x)$ by $\mathcal{S}_j(x)\, \hat{R}_j$. We estimate the density $p_{X_j}(x)$ of the j-th predictor from the samples by any unidimensional density estimator, $\hat{p}_{X_j}(x)$. For instance, the kernel density estimate with kernel K and bandwidth h is given as

$$\hat{p}_{X_j}(x) = \frac{1}{n\, h} \sum_{i=1}^{n} K\left(\frac{X_j^{(i)} - x}{h} \right).$$

Finally, for the update Equation (4.4.3b) of the weights α, collect $\hat{F}_j = \{\hat{f}(X_j^{(i)})\}_{i=1}^n$, the function estimates at each predictor, as columns of a design matrix $\hat{F}$. Let $\hat{Y} = (Y^{(1)}, \ldots, Y^{(n)})^T$. Plugging in the empirical for the population expectation in the update Equation (4.4.3b), we then obtain α as,

$$\alpha = \arg\min_{\alpha \in \mathbb{R}^p} (\hat{Y} - \hat{F}\alpha)^2 + \lambda \mid \alpha \mid_1,$$

which is an instance of the Lasso (Tibshirani, 1996) with design matrix $\hat{F}$, response vector $\hat{Y}$ and penalty λ. Figure 4.9 gives the final data version of the updates Equation (4.4.3).

Input: Data $\{(X^{(i)}, Y^{(i)})\}_{i=1}^n$, regularization parameters λ, μ.

Initialize $\hat{f} = 0$. For $j = 1, \ldots, p$,

Set weight $\hat{\alpha}_j = 1$, and

Estimate the density $p_{X_j}(x)$ by a unidimensional density estimator: $\hat{p}_{X_j}(x)$.

Iterate until convergence:

Optimize for $\hat{f}$:

Iterate until convergence:

For each $j = 1, \ldots, p$:

Compute the residual $\mathrm{R}_j(X, Y, \hat{f}) = Y - \sum_{k \neq k} \hat{\alpha}_j \hat{f}_k(X_k)$ at the data points:

$$\hat{\mathrm{R}}_j = \{\mathrm{R}_j(X^{(i)}, Y^{(i)}, \hat{f})\}_{i=1}^n.$$

Estimate the conditional expectation $\mathrm{P}_j = \mathbb{E}[\mathrm{R}_j | X_j]$ by smoothing:

$$\hat{\mathrm{P}}_j(x) = \mathcal{S}_j(x)\, \hat{\mathrm{R}}_j.$$

Set $\hat{f}(x) = \sum_{j=1}^p \left\{ \frac{\hat{\alpha}_j \, \hat{p}_{X_j}(x)}{\sum_{k=1}^p (\hat{\alpha}_k^2 + \mu) \, \hat{p}_{X_k}(x)} \right\} \hat{\mathrm{P}}_j(x)$.

Optimize for α:

Collect $\hat{F}_j = \{\hat{f}(X_j^{(i)})\}_{i=1}^n$, the function estimates at each predictor, as columns of a design matrix $\hat{F}$.

Solve $\hat{\alpha} = \mathrm{Lasso}(\hat{Y}, \hat{F}, \lambda)$.

Output: Common function $\hat{f}$ and weights $\{\hat{\alpha}_j\}$.

Figure 4.9: The iSPAM Fitting Algorithm

4.5 Fitting

The previous sections described our V-SPAM and V-iSPAM models, and algorithms to fit these to data. Here, we provide further specifics on how we fit these models to our dataset.

A separate V-SPAM model was fit at each of the 1,294 voxels using the training set of 1,750 images and the evoked fMRI response amplitudes. Recall how the artificial neuron models in the filtering stage of the V-SPAM model were based on Gabor wavelets. We constructed a pyramid of these wavelets according to equation Equation (4.2.1), using five different spatial frequencies, 8 orientations and 2 phases on a grid of size 128×128 pixels. The wavelets were centered and scaled to have mean 0 and norm 1. At each of the spatial frequencies $\omega \in \{1, 2, 4, 8, 16, 32\}$ cycles per field of view, the wavelets were positioned evenly on a $\omega \times \omega$ grid covering the image. All combinations of the 8 orientations and 2 phases occured at each of the $\omega \times \omega$ positions. In total, the Gabor wavelet pyramid consisted of 21,840 wavelets plus 1 constant wavelet providing an intercept term for the model.

Using these wavelets, we constructed model simple and complex cells according to equations Equation (4.2.2) and ((4.2.3)) respectively thus obtaining 10,921 model complex cells. Recall the intuition behind the sparsity assumption in our models, which was based on the empirical observation that the response at any voxel is receptive to stimuli in a small spatial region, so that among the set of all complex cells the active set at any voxel would be small. We thus performed an intermediate *pre-screening* in order to eliminate complex cell outputs unrelated to a voxel response, and also to reduce the computational complexity

of successive stages of fitting. We computed the correlation of the response of each complex cell with the evoked voxel response, using the 1,750 images in the training set, and retained only the top k complex cells for each voxel. In pilot studies we found empirically that $k = 100$ was enough to give good statistical and computational performance (data not shown). Using this reduced number of complex cells, we constructed orientation invariant pooled-complex cells as described in Section 4.3.1.

With these neuronal filters in place, the V-SPAM model was then fit using the sparse additive model fitting algorithm in Figure 4.8. Given any image, the corresponding outputs of the model neurons were the predictors of the sparse additive model, while the evoked voxel response amplitude was its response. The 1,750 images and evoked fMRI responses in the model estimation dataset thus provided 1,750 training samples for the sparse additive model. The sparse additive model fitting algorithm also required as input two other quantities: a smoother (S_j in Figure 4.8) and a regularization parameter λ. We used the Gaussian kernel smoother with plug-in bandwidth, and chose λ according to the Akaike information criterion (AIC).

Similarly, the V-iSPAM model was fit using the iSPAM fitting algorithm and the 1,750 training samples of the neuronal filter outputs and the corresponding fMRI response amplitudes. The fitting algorithm required as input two regularization parameters λ and μ. The μ parameter effectively fixes the scale of the function, but must be large enough to ensure convergence of the function updates, as was shown in Lemma 4.2. We found that fixing it to 0.5 was enough to ensure numerical convergence across all voxels. We set the ℓ_1-regularization penalty λ by five-fold cross-validation. Finally, the iSPAM fitting

algorithm also required as input a unidimensional density estimator, for which we used the kernel density estimator, with the Gaussian kernel and plug-in bandwidth. For the unidimensional smoothing procedures, as before we used Gaussian kernel regression.

For comparison, we used the same data to fit a sparse linear pooling model. As described in Section 4.2, a linear pooling model aims to predict each voxel's response as a linear combination of all 10,921 complex cell outputs. This model, with the fixed transformation $h(t) = \sqrt{t}$, was used in earlier work with this data set (Kay et al., 2008). As benchmark, we used a sparse variant of this model with the fixed transformation $h(t) = \log(1 + \sqrt{t})$. The selection of this transformation was supported by investigations performed after (Kay et al., 2008) that indicated a slight improvement over the earlier results. The coefficients of the model were estimated by L2 Boosting (Bühlmann and Yu, 2003) with the stopping criterion determined by 5-fold cross-validation.

4.6 Results

For each voxel, we evaluated the fitted V-SPAM and V-iSPAM models by computing the *predictive* R^2 (squared correlation) of the predicted and actual fMRI responses evoked by each of the 120 images in the validation set.

4.6.1 Prediction

Figure 4.10(a) shows scatterplots comparing the performance of the three fitted models on the validation dataset. Both the fitted V-SPAM and the fitted V-iSPAM models are seen to provide a large improvement over the sparse linear model for many of the voxels.

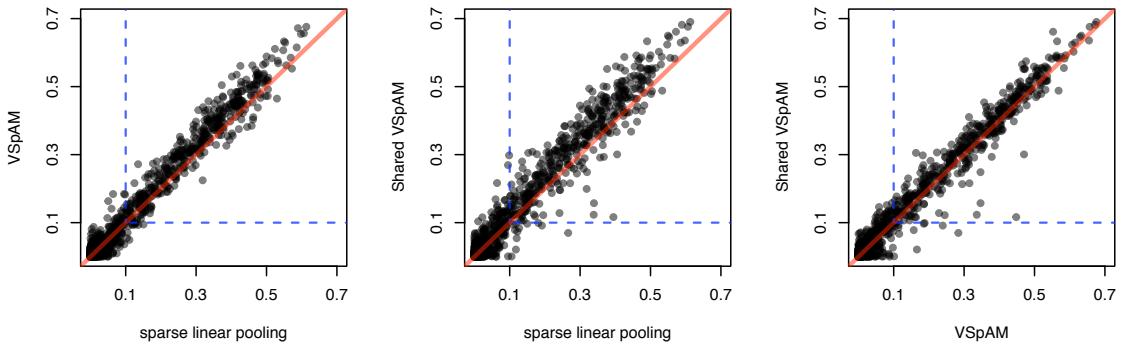
The inset region in the scatterplots contains the voxels for which both models had some predictive power ($R^2 \geq 0.1$). For those voxels, the average relative improvement of the non-parametric models over the sparse linear model was about 14% (standard deviation 14%) for V-SPAM and 16% (standard deviation 21%)for V-iSPAM. Figure 4.10(b) shows histograms of the relative improvement across voxels for each of the 3 comparisons.

The fitted V-SPAM and fitted V-iSPAM models have comparable predictive power. On average the latter has a slight edge over the former with an mean relative improvement of 5.6% (standard deviation 16%) on voxels with some predictive power. This may be due to additional regularization provided by the identical nonlinearity constraint of the iSPAM fit. However, the right-most scatterplot in Figure 4.10(a) shows that there are a few voxels where the fitted V-iSPAM model predicts very poorly compared to V-SPAM.

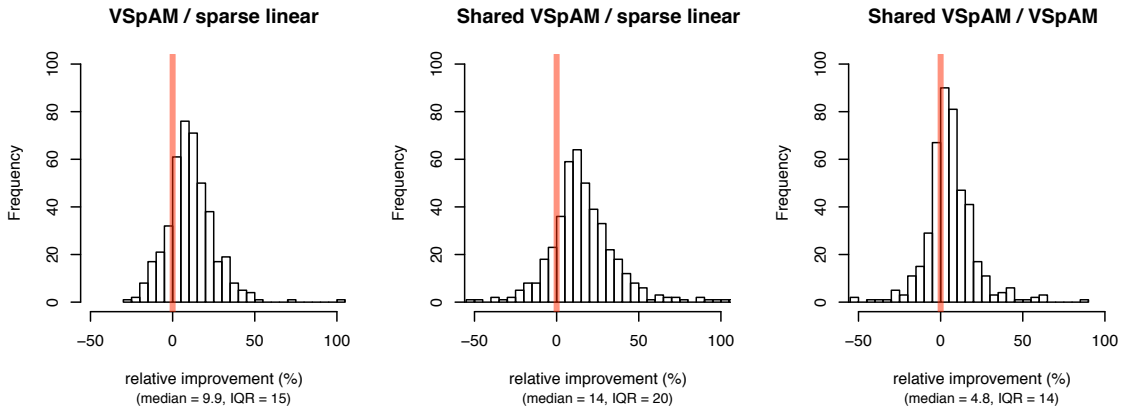
4.6.2 Nonlinearities

One of the potential advantages of the V-SPAM model over other approaches is that it can reveal novel nonlinear tuning and pooling properties. Figure 4.11(a) illustrates some of these functions estimated for a typical voxel with high predictive power (R^2 of 0.63). These correspond to the nonlinearities appearing in the final stage of the V-SPAM model (see Figure 4.6).

Here the horizontal axis is the input in standard units of the corresponding model complex or pooled-complex cell outputs, and the vertical axis is the output in standard units of predicted responses. For this voxel, these are the 4 largest (ranked by L^2 norm) nonlinearities. All four of these nonlinearities are compressive, that is they saturate and



(a) Predictive R^2



(b) Relative performance

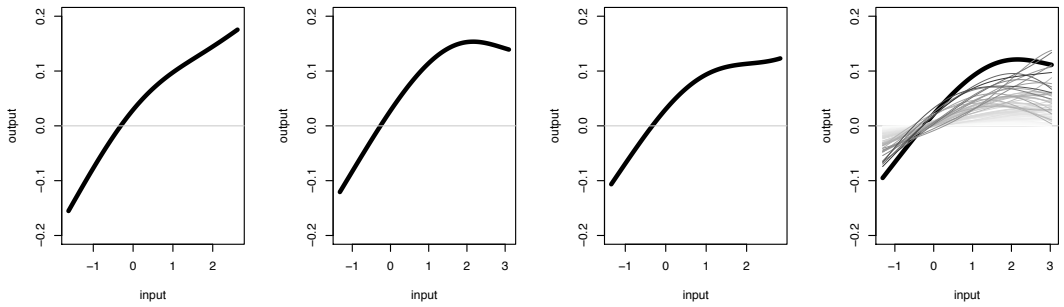
Figure 4.10: Comparisons of the predictive power of the 3 different fitted models. (a) Scatterplot of the predictive R^2 on the validation dataset for each of 1,294 voxels. The V-SPAM and the V-iSPAM models have similar prediction performance, and provide improvement over the sparse linear model for many of the voxels. (b) Relative performance (calculated as the ratio of the predictive R^2's minus 1) for the voxels contained in the inset regions in (a).

become less responsive to the input after a certain threshold. Figure 4.11(b) shows the nonlinearity estimated for the fitted V-iSPAM model for the same voxel. This too has a saturating form. The estimated nonlinearities had similar forms across voxels. Figure 4.12 shows transformations at ten other voxels with high predictive power as estimated by the V-iSPAM model.

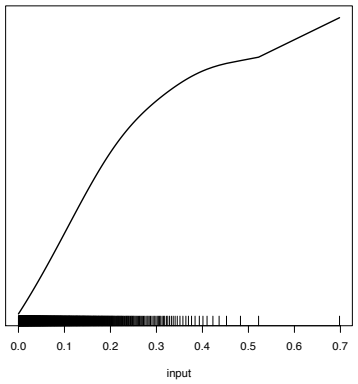
4.7 Qualitative Aspects

4.7.1 The Saturation Effect

The standard models for simple and complex cells in Section 4.2.1 and Section 4.2.1 only capture a simplified version of actual neural behavior. For instance, Maffei and Fiorentini (1973) showed that simple cell outputs exhibit saturation with increasing input light intensities. Movshon et al. (1978b) and others have shown, complex cells exhibit nonlinearities and saturative behavior as well. Interestingly, as shown in the previous section the nonlinear transformations of the standard complex cell outputs, automatically learned in the pooling stage, were saturating as well. This indicates that previous fixed transformations, e.g. $h(t) = \log(1 + \sqrt{t})$ is not compressive enough. However, due to the nature of the fMRI BOLD signal it is difficult to determine if this is an effect of nonlinearity in the electrophysiological transformation from image stimulus to neuronal output, or the transformation from neuronal output to BOLD signal. In either case, it is plausible that the saturation in the estimated nonlinearity reflects the fact that neural activity consumes resources in the brain and that it is impossible for the activation to be potentially unbounded, as would be



(a) Four largest V-SPAM nonlinearities



(b) V-iSPAM nonlinearity

Figure 4.11: Estimated nonlinearities for a voxel with high predictive power (R^2: 0.63). (a) The 4 largest (ranked by L^2 norm) are shown left to right by the thick lines. The grey lines in the rightmost figure show the rest of nonlinearities. (b) The nonlinearity estimated by the V-iSPAM model for the same voxel (plotted with unstandardized inputs). A rug plot of the corresponding $X_i's$ from the training set is plotted along the x-axis.

suggested by a linear relationship.

4.7.2 Estimated Receptive Fields and Tuning Curves

Figure 4.13 shows the spatial receptive-fields (RF's) and joint frequency and orientation tuning curves estimated using the V-iSPAM model for 3 voxels. These voxels were chosen because they had high predictive power (R^2's of 0.65, 0.59, and 0.63, respectively from left to right) and so were modeled accurately. The upper row of the figure shows the spatial RF of each voxel. The intensity at each location in the spatial RF represents the standardized predicted response of the voxel to an image stimulus consisting of a single pixel at that location. The spatial RF's of these voxels are clearly localized in space, consistent with the known retinotopic organization of V1 and previous fMRI results (Wandell et al., 2007). The lower row of Figure 4.13 shows the joint frequency and orientation tuning properties of these same 3 voxels. Here the tuning curves were estimated by computing the predicted response of the fitted voxel model to cosine gratings of varying orientation (degrees) and spatial frequency (cycles/field of view). All of the voxels are tuned to spatial frequencies above about 8 cycles/field of view, while orientation tuning varies from voxel to voxel. The joint spatial frequency and orientation tuning of all 3 voxels appears to be non-separable (i.e. their orientation tuning is not a constant function of frequency). The tuning curves for the V-SPAM model were similar and are not shown.

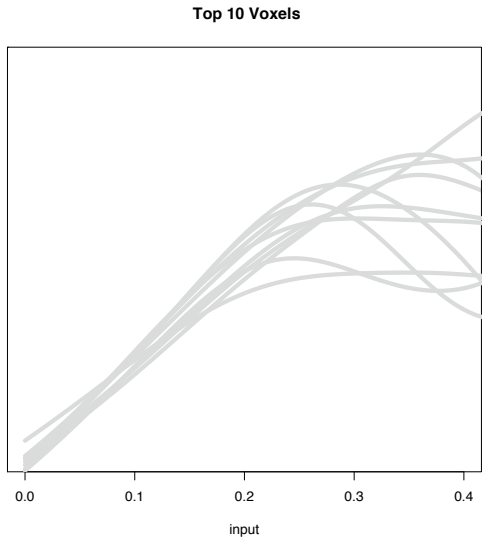


Figure 4.12: Non-linearities estimated by the V-iSPAM model for ten other voxels with high predictive power.

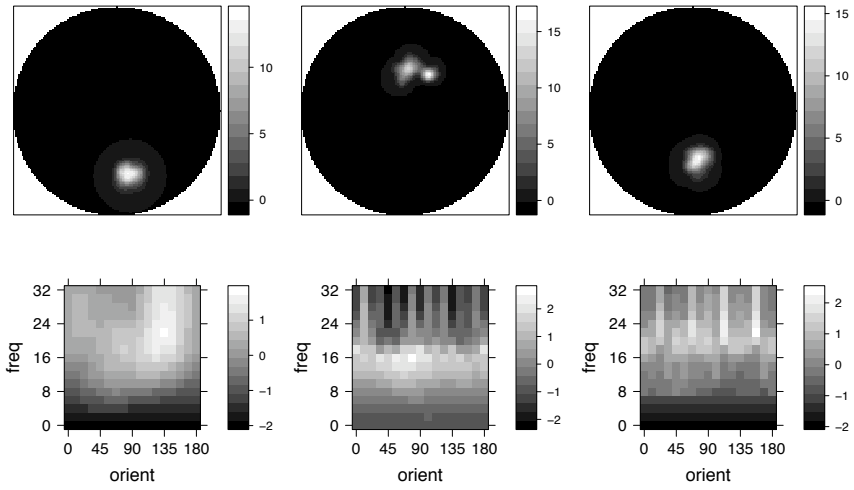


Figure 4.13: (upper) Spatial receptive-fields (RF's) and (lower) joint frequency and orientation tuning curves estimated by the V-iSPAM model for 3 voxels with high predictive power (R^2's of 0.65, 0.59, 0.63, left to right). Each location in the spatial RF shows the standardized predicted response of the voxel to an image consisting of a single pixel at that location. The tuning curves show the standardized predicted response of the voxel to cosine gratings of varying orientation (degrees) and spatial frequency (cycles/field of view).

4.7.3 Flat Map

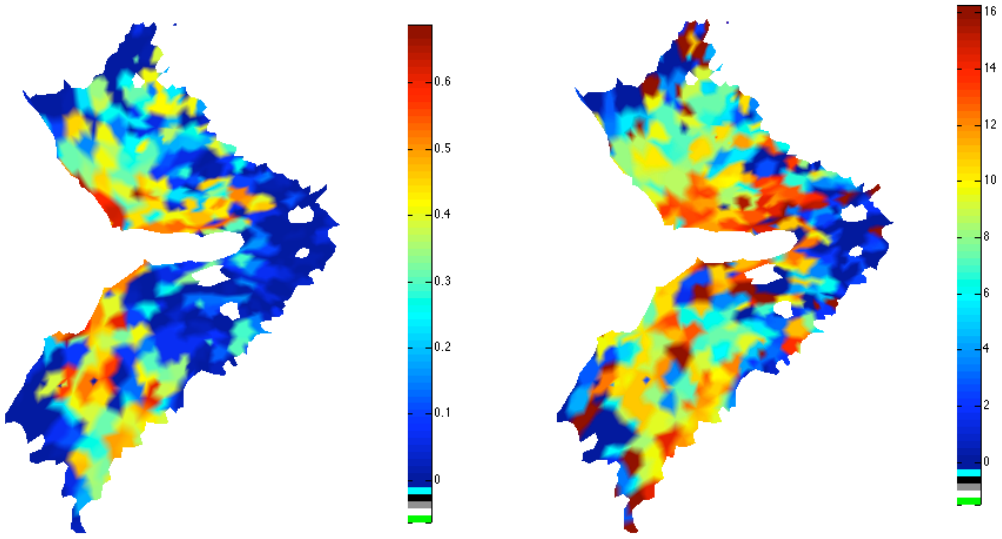
The V-SPAM framework models the fMRI response at each voxel separately. However, voxels are located in space and properties and it is of potential interest to determine whether the fitted models reflect the proximity of voxels in any properties. The surface of the brain is actually a convoluted surface–like a deflated beachball, and so a notion of voxel proximity should take this into account.

Flat maps provide a two-dimensional view of the surface of the brain Kay et al. (2008) and can be used for visualizing properties of the fitted voxel models. In the construction of a flat map, anatomical MRI data are acquired from the subject, and then specialized algorithms are used to reconstruct the cortical surface of each subject. The cortical surfaces are then computationally flattened before projecting data onto the surface. One basic property of the fitted voxels is their predictive R^2. Figure 4.14(a) shows these values projected onto the flat map. The smooth gradation of color indicates that nearby voxels have similar predictive power.

Another property of a fitted voxel provided by V-iSPAM model is the single function that is identical across predictor variables. The variation of these functions across the cortical surface is of potential scientific interest. One summary of the function is its mean derivative. To make this summary comparable across voxels, we standardized the fitted functions to have SD 1 for each voxel. The derivatives of the function can then be estimated using standard curve-fitting tools. We used smoothing splines with the smoothness parameter chosen by cross-validation. The mean derivative was then estimated by averaging over the aggregate of the active predictor variables in the training set. Figure 4.14(b) shows

the absolute value of the estimated mean derivative of the functions. There are similarities with Figure 4.14(a), however the gradation of values seems to be much smoother for the mean slope.

Both flat maps suggest that proximity does play a role in the properties of the voxel. Taking proximity into account, i.e. borrowing strength, in the fitting could be fruitful topic of investigation in the future.



(a) Predictive R^2 (b) Estimates of $|\int f'(x)d\bar{P}(x)|$

Figure 4.14: Flat map

4.8 Additional Details

Let $Y(t) \in \mathbb{R}$ denote the fMRI time-series signal. Let Let $h_i(t)$ denote the hemodynamic response function for the i-th image stimulus. And let $X_i(t)$ be the delta function indicating whether the trial at time t is showing the image i. (Recall that the same image is shown more than once over the sequence of image trials.) Let $D(t)$ and $\epsilon(t)$ denote the drift and the noise terms respectively. Finally, let $A(t) * B(t)$ denote the convolution of the series $A(t)$ and $B(t)$. With this notation, the fMRI signal $Y(t)$ is modeled as,

$$Y(t) = \sum_{i=1}^{n} X_i(t) * h_i(t) + D(t) + \epsilon(t),$$

where n is the number of distinct images. The basis-restricted separable model assumes further that the hemodynamic response functions have the same form except for scale, so that $h_i(t) = c_i\, h(t)$, for $i = 1, \ldots, n$. Thus the weight $c_i \in \mathbb{R}$ is distinct for each image while $h(t) \in \mathbb{R}$ as the form of the hemodynamic response function is common to all images. This common form $h(t)$ in turn is modeled as a linear combination of Fourier basis functions (consisting of a constant function and sine and cosine functions with 1, 2, and 3 cycles),

$$h(t) = \sum_{j=1}^{n_h} p_j L_j(t),$$

where n_h is the number of Fourier basis functions, and $\{p_j\}_{j=1}^{n_h}$ are their coefficients. The drift term is steady enough to be modeled as a linear combination of polynomial functions (of degrees 0 through 3),

$$D(t) = \sum_{k=1}^{n_d} b_k S_k(t),$$

where n_d is the number of polynomial regressors and $\{b_k\}_{k=1}^{n_d}$ are their coefficients. Combining the pieces, the fMRI signal $Y(t)$ is given as,

$$Y(t) = \sum_{i=1}^{n} X_i(t) * \Big(c_i \sum_{j=1}^{n_h} p_j L_j(t)\Big) + \Big(\sum_{k=1}^{n_d} b_k S_k(t)\Big) + \epsilon(t).$$

Once the parameters $\{c_i\}_{i=1}^{n}, \{p_j\}_{j=1}^{n_h}, \{b_k\}_{k=1}^{n_d}$ are fit to the observed fMRI signal $Y(t)$, we obtain a distinct hemodynamic response amplitude $c_i \in \mathbb{R}$ for each image stimulus at each voxel. Further details of this preprocessing can be found in Kay et al. (2008).

4.9 Proofs

Theorem 4.1

Proof. Consider optimizing problem (4.4.2) over the function f while holding the scale parameters α fixed. A first order expansion of the objective L around some $f \in L_2(P())$ with increment $\epsilon\eta \in L_2(P())$ is given as,

$$L(f + \epsilon\eta) = L(f) + \epsilon \sum_j \mathbb{E}\left[(\alpha_j^2 + \mu) f(X_j) - \alpha_j R_j)\eta(X_j)\right] + O(\epsilon^2).$$

The first variation of L around f in the direction of η can thus be written as

$$\begin{aligned}
\partial L(f;\eta) &= \lim_{\epsilon \to 0} [L(f + \epsilon\eta) - L(f)]/\epsilon \\
&= \sum_j \mathbb{E}\left\{[(\alpha_j^2 + \mu) f(X_j) - \alpha_j R_j]\eta(X_j)\right\} \\
&= \sum_j \int \left\{(\alpha_j^2 + \mu) f(x) - \mathbb{E}(\alpha_j R_j | X_j = x)\right\} \eta(x) P(X_j = x) dx \\
&= \int \left\{ \sum_j \left\{(\alpha_j^2 + \mu) f(x) - \mathbb{E}(\alpha_j R_j | X_j = x)\right\} P(X_j = x) \right\} \eta(x) dx,
\end{aligned}$$

which is linear in the increment η. Note that $L_2(P())$ has the inner-product

$$\langle f, g\rangle = \int_0^1 f(x)g(x)dx.$$

The Fréchet derivative of L at f is thus given by

$$\partial L(f)(x) = \sum_j \left\{(\alpha_j^2 + \mu)f(x) - \mathbb{E}(\alpha_j R_j | X_j = x)\right\} P(X_j = x),$$

since the first variation can be written as $\partial L(f;\eta) = \langle \partial L(f), \eta\rangle$. Setting the derivative to zero yields the following stationary condition:

$$f(x) \quad = \quad \frac{\sum_j \alpha_j \mathbb{E}(R_j | X_j = x) P(X_j = x)}{\sum_j (\alpha_j^2 + \mu) P(X_j = x)}. \tag{4.9.1}$$

Now consider optimizing the objective (4.4.2) over α, while holding f fixed. This is just an ℓ_1-regularized least squares problem, where the predictors are transformed by the function f. Thus the optimal scale parameters α when f is fixed would satisfy

$$\alpha \quad = \quad \arg\min_{\alpha \in \mathbb{R}^p} \mathbb{E}(Y - \sum_{j=1}^p \alpha_j f(X_j))^2 + \lambda \sum_j |\alpha_j|. \tag{4.9.2}$$

The optimal solution (f^*, α^*) satisfies these two stationary conditions (4.9.1), ((4.9.2)), and equivalently is a fixed point of the updates in the theorem. ■

Lemma 4.2

Proof. Let $W(x) = \sum_k (\alpha_k^2 + \mu) p_{X_k}(x)$ and fix $\mu > 0$ and x such that the density of $\bar{P}$ is positive at x.

$$\begin{aligned} |(\mathcal{A}f)(x)|^2 &= \frac{1}{[W(x)]^2} \left[\sum_k \alpha_k \sum_{j \neq k} \alpha_j p_{X_j}(x)(\mathrm{P}_{jk} f)(x)\right]^2 \\ &\leq \frac{1}{[W(x)]^2} \left[\sum_{jk} |\alpha_k \alpha_j p_{X_j}(x)(\mathrm{P}_{jk} f)(x)|\right]^2. \end{aligned}$$

Applying the Cauchy-Schwarz inquality to the sum over jk, and then Jensen's Inequality gives

$$\begin{aligned}|(\mathcal{A}f)(x)|^2 &\leq ||\alpha||^2 \frac{\sum_j \alpha_j^2 p_{X_j}(x)}{[W(x)]^2} \sum_{jk} p_{X_j}(x) |(\mathrm{P}_{jk} f)(x)|^2 \\ &\leq ||\alpha||^2 \frac{\sum_j \alpha_j^2 p_{X_j}(x)}{[W(x)]^2} \sum_{jk} p_{X_j}(x) (\mathrm{P}_{jk} f^2)(x).\end{aligned}$$

Note that $W(x) = \sum_k (\alpha_k^2 + \mu) p_{X_k}(x) > \sum_k \alpha_k^2 p_{X_k}(x)$ and so

$$\begin{aligned}|(\mathcal{A}f)(x)|^2 &< \frac{||\alpha||^2}{\sum_k (\alpha_k^2 + \mu) p_{X_k}(x)} \sum_{jk} p_{X_j}(x) (\mathrm{P}_{jk} f^2)(x) \\ &\leq \frac{||\alpha||^2}{\mu \sum_k p_{X_k}(x)} \sum_{jk} p_{X_j}(x) (\mathrm{P}_{jk} f^2)(x).\end{aligned}$$

Combine this with the definitions of $\bar{P}$ and P_{jk} to conclude that

$$\begin{aligned}\int |(\mathcal{A}f)(x)|^2 \, \bar{P}(dx) &= \frac{1}{p} \int |(\mathcal{A}f)(x)|^2 \sum_k p_{X_k}(x) \, dx \\ &< \frac{1}{p} \frac{||\alpha||^2}{\mu} \sum_j \sum_k \int p_{X_j}(x) \mathbb{E}(f^2(X_k) | X_j = x) \, dx \\ &= p \frac{||\alpha||^2}{\mu} \int |f(x)|^2 \bar{P}(dx).\end{aligned}$$

■

Chapter 5

High Dimensional Analysis of Ridge Regression

5.1 Introduction

Consider the linear model

$$Y = X\beta + \epsilon \tag{5.1.1}$$

where $X = Z\Sigma^{1/2}$ is a $n \times p$ matrix of predictors with $\mathbb{E}X = 0$, $\mathbb{E}X^T X = \Sigma$, and ϵ is a noise vector whose coordinates are i.i.d. with mean 0 and variance σ^2. The least squares estimate of β,

$$\hat{\beta}_{ols} = (X^T X)^{-1} X^T Y \,,$$

is known to be unstable—its value can change dramatically with small perturbations of Y—when the columns of X are highly colinear (Hoerl and Kennard, 1970). Moreover, it becomes invalid if $p > n$; for example, in the preceding chapter we considered $n = 1,331$

natural images and $p = 10,921$ features for each of the n images.

Ridge regression was introduced by Hoerl and Kennard (1970) to address the estimation of β in the highly colinear setting. Its roots are in chemometrics, but it is also related to work in numerical analysis due to Tikhonov (1943). The ridge regression estimate of β is given by,

$$\hat{\beta}_{ridge} = (X^T X + kI)^{-1} X^T Y \,,$$

and corresponds to the penalized least squares problem,

$$\min_b ||Y - Xb||^2 + kb^T b \,.$$

Thus, it modifies the ordinary least squares criterion by adding a penalty on the size of the coordinates of b. Larger k's encourage solutions with smaller norms $||b||^2$.

Hoerl and Kennard (1970) proved a *ridge existance theorem* which guarantees the existence of a k such that the mean squared estimation of ridge regression is smaller than that of ordinary least squares. That is, there exists k such that

$$\mathbb{E}||\hat{\beta}_{ridge} - \beta||^2 < \mathbb{E}||\hat{\beta}_{ols} - \beta||^2 \,.$$

Unfortunately, their theorem does not show how to choose such a k on the basis of the data, nor does it guarantee that a data based selection of k will be successful. That deficiency was one source of much of the early criticism of ridge regression (see Conniffe and Stone (1973) for example). More recent studies have shown that cross-validation can be a successful strategy for choosing k, but care must be taken (Burr and Fry, 2005). Another troubling aspect was that unlike $\hat{\beta}_{ols}$, the ridge estimate is no longer equivariant under rescaling of X. Smith and Campbell (1980) attributed this to the fact that $\hat{\beta}_{ridge}$ is a Bayes estimate

with an implicit prior. So it should not be surprising that rescaling X leads to different implicit prior distributions for β. Much of the research on ridge regression during the 1970s is summarized by Draper and van Nostrand (1979).

Power ridge regression was suggested in chemometrics (Sommers, 1964) and introduced into the statistics literature by Goldstein and Smith (1974). The power ridge estimate of β is given by,

$$\hat{\beta}_{power} := [X^T X + k(X^T X)^q]^{-1} X^T Y \,, \tag{5.1.2}$$

where the power q can be any number. It includes ridge regression as a special case with $q = 0$. It was dismissed by Hoerl and Kennard (1975) as "unfruitful," and yet has come back in more recent literature (Frank and Friedman, 1993; Obenchain, 1995; Burr and Fry, 2005). The works of Gibbons (1981) and Obenchain (1995) are interesting because they explicitly considered the problem of selecting q using the data, whereas Frank and Friedman (1993) had considered it fixed a priori. The simulation results of Burr and Fry (2005) suggest that selection of q from a few candidates by cross-validation can be successful.

5.2 Generalized Ridge Regression

Replacing the penalty $\beta^T \beta$ in ridge regression by a general quadratic penalty $b^T G b$, with $G \geq 0$ (G non-negative definite), leads to the penalized least squares problem,

$$\min_b \left\{ ||Y - Xb||^2 + k b^T G b \right\} \,. \tag{5.2.1}$$

Note that whereas ridge regression is invariant under rotation of X, generalized ridge regression is in general not, because the matrix G allows for the possibility of anisotropic

penalization of b. Ridge regression is a special case with $G = I$. Another special case is given by $G = (X^T X)^q$. In that case G is stochastic and different choices of q lead to different members of the power ridge family of estimators.

5.2.1 Duality of Penalization and Transformation

A related family of estimators is given by the penalized least squares problem,

$$\min_b \left\{ ||Y - XCb||^2 + k||b||^2 \right\} , \tag{5.2.2}$$

where C is a non-negative definite matrix. Equation (5.2.1) and Equation (5.2.2) are equivalent when G and C are restricted to be positive definite (or simply positive) via the mapping $C = G^{-1/2}$. Actually, Equation (5.2.2) is just ridge regression with the (linearly) transformed predictors,

$$\tilde{X} = XC .$$

This makes Equation (5.2.2) more useful to work with because it allows for the possibility of dimension reduction of the predictors in addition to anisotropic penalization of b. For example, C could be the composition $C = HG^{-1/2}$, where H is a dimension reducing projection and GB is a positive definite penalty matrix as in Equation (5.2.1). (For brevity, we will refer to positive definite matrices as positive and similarly for non-negative definite matrices.)

Since Equation (5.2.2) is equivalent to ridge regression with the transformed predictor $\tilde{X}$, its solution in terms of the original predictor (in the linear model of Equation (5.1.1)) is given by

$$\hat{\beta} = C(\tilde{X}^T \tilde{X} + kI)^{-1} \tilde{X}^T Y = C(C^T X^T X C + kI)^{-1} C^T X^T Y .$$

Dual Form If $k > 0$ or if $\tilde{X}\tilde{X}^T$ is non-singular, then we can write $\hat{\beta}$ in dual form as,

$$\hat{\beta} = C\tilde{X}^T(\tilde{X}\tilde{X}^T + kI)^{-1}Y = AX^T(XAX^T + kI)^{-1}Y\,, \tag{5.2.3}$$

where $A = CC^T$. This can be seen by considering the singular value decomposition of $\tilde{X}$. Saunders et al. (1998) gave a formal derivation from an optimization point of view and considered Equation (5.2.3) as a starting point for the kernelization of ridge regression. (They proposed replacing $\tilde{X}\tilde{X}^T$ by an arbitrary kernel matrix). Moreover, the dual form of ridge regression can be computationally more time and space efficient than the standard form when $p \gg n$ because Equation (5.2.3) can be solved for different values of k simultaneously by computing the singular value decomposition of an $n \times n$ matrix, rather than a $p \times p$ matrix; while the predicted value at a point $\tilde{x} = xC$ is given (via some algebra) by the equation

$$x\hat{\beta} = \tilde{x}\tilde{X}^T(Y - \tilde{X}\hat{\beta})/k\,.$$

Thus, the prediction equation is determined by an n-vector of inner products $\tilde{x}\tilde{X}^T$ and the n-vector of residuals $Y - X\hat{\beta}$.

5.2.2 Elliptical Constraints and the Choice of Penalty/Transformation

There is a natural correspondence between non-negative A and ellipsoids in $\mathbb{R}^p$. Associated to each $b \in \mathbb{R}^p$ is the rank 1 matrix bb^T. Then the set $\{b : bb^T \leq A\}$ is a (solid) ellipsoid in $\mathbb{R}^p$. When A is non-singular it is equivalent to $\{b : b^TA^{-1}b \leq 1\}$. The principal axes of the ellipsoid are given by the eigenvectors of A and the squared radii by the eigenvalues of A. We will occasionally abuse notation by referring to A as both a matrix and as an ellipsoid. The penalized least squares problem (Equation (5.2.1)) corresponds to the

Lagrangian form of a least squares problem with the elliptical constraint $b \in \{b : bb^T \leq cA\}$ for some $c > 0$. In that context, generalized ridge regression also appeared in the work of Kuks and Olman (1971, 1972). They considered the estimation of the one-dimensional parameter $a^T\beta$ under the meansquared error $\mathbb{E}[(a^T\hat{\beta} - a^T\beta)^2 \mid X]$ subject to the constraint that $\beta\beta^T \leq A/\sigma^2$.

Clearly, the choice of A has an effect. For example, the choice of q in power ridge regression (Equation (5.1.2)) corresponds to choosing A from the family $\{(X^TX)^{-q} : q\}$. There $q = 0$ corresponds to a sphere, while $q \neq 0$ correspond to proper ellipsoids provided X^TX is not a multiple of the identity. As q decreases to $-\infty$, the major axes of $(X^TX)^{-q}$ expand while its minor axes shrink—producing increasingly eccentric ellipsoids. The same is true as q increases to $+\infty$, except the role of the major and minor axes switches.

So different choices of q can lead to dramatically different ellipsoids. However, it is not immediately clear whether different ellipsoids will lead to dramatically different estimates. Furthermore, it would be useful to understand what properties are desirable for A. These concerns are addressed in the next section.

5.3 Prediction Error

Let us fix the notation in this section, taking $A = CC^T$ as in Equation (5.2.3), and define the generalized ridge estimator in dual form as

$$\hat{\beta}_{(k,A)} := AX^T(XAX^T + kI)^{-1}Y\,. \tag{5.3.1}$$

We are interested in the mean squared prediction error,

$$\text{mspe}(k, A, \beta) := \mathbb{E}\{||X\hat{\beta}_{(k,A)} - X\beta||^2 \mid X\},$$

that depends on the unknown parameter β. To understand the effect of the choice of A, consider the worst case error when β belongs to the ellipsoid B,

$$\text{MMSPE}(k, A, B) := \max_{\beta:\beta\beta^T \leq B} \text{mspe}(k, A, \beta)\,.$$

We will analyze $\text{MMSPE}(k, A, B)$ as k, A, and B vary.

5.3.1 Bias and Variance Decomposition

The mean squared prediction error has a well-known bias-variance decomposition. Firstly, we can expand the error $X\hat{\beta}_{(k,A)} - X\beta$ in the following way.

$$\begin{aligned} X\hat{\beta}_{(k,A)} - X\beta &= XAX^T(XAX^T + kI)^{-1}(X\beta + \epsilon) - X\beta \\ &= -k(XAX^T + kI)^{-1}X\beta + XAX^T(XAX^T + kI)^{-1}\epsilon\,. \end{aligned}$$

The first term on right hand side of the last line is related to the bias of $\hat{\beta}_{(k,A)}$, while the second term is related to the variance. Since ϵ has mean 0 and is independent of X, the squared bias and variance are given by

$$||\mathbb{E}\{X\hat{\beta}_{(k,A)} - X\beta)|X\}||^2 = k^2 \,\text{tr}\left\{(XAX^T + kI)^{-2}X\beta\beta^T X^T\right\}$$

and

$$\mathbb{E}\{||X\hat{\beta}_{(k,A)} - \mathbb{E}(X\hat{\beta}_{(k,A)}|X)||^2 \mid X\} = \sigma^2 \,\text{tr}\left\{(XAX^T + kI)^{-2}(XAX^T)^2\right\}\,,$$

respectively. Thus, the mean squared prediction error is

$$\begin{aligned} \text{mspe}(k, A, \beta) = k^2 \,&\text{tr}\left\{(XAX^T + kI)^{-2}X\beta\beta^T X^T\right\} \\ &+ \sigma^2 \,\text{tr}\left\{(XAX^T + kI)^{-2}(XAX^T)^2\right\}\,. \end{aligned} \tag{5.3.2}$$

5.3.2 MSPE under Ideal Conditions

When B is known, we may choose $A = B$ and hope that MMSPE(k, A, A) is small. The following theorem gives a lower bound for MMSPE in this ideal setting.

Theorem 5.1.

$$\min_k \text{MMSPE}(k, A, A) \geq \sigma^2 \operatorname{tr}\left\{(XAX^T + \sigma^2 I)^{-1}(XAX^T)\right\} . \tag{5.3.3}$$

The quantity in the lower bound in the theorem appears in Bayesian analysis. When β has a $\mathcal{N}(0, A^{-1})$ prior distribution and ϵ is also $\mathcal{N}(0, \sigma^2 I)$, then the posterior variance of $X\beta$ given (X, Y) is equal to the above upper bound. However, in our current setting $\beta\beta^T \leq A$ and so the $\mathcal{N}(0, A^{-1})$ prior does not apply. Interestingly, the lower bound depends only on the eigenvalues of XAX^T. In a later section we will use a result in random matrix theory to show that under certain assumptions on X it really only depends on the eigenvalues of $\Sigma^{1/2} A \Sigma^{1/2}$.

5.3.3 MSPE under Misspecification

What is the effect of the choice of A on the mean squared prediction error when $A \neq B$? The following theorem gives upper and lower bounds on MMSPE that depend on a geometric relationship between the ellipsoids A and B. To make this more clear, we reparameterize the regularization parameter k as $k = \sigma^2/\alpha$ to make this more clear. It includes Theorem 5.1 as a special case.

Theorem 5.2. *Let* $k = \sigma^2/\alpha$.

I. *Assume that* $A > 0$. *If* $\alpha > 0$ *then*

$$\begin{aligned} \mathrm{MMSPE}(\sigma^2/\alpha, A, B) \leq \sigma^2 \operatorname{tr}\left\{(XAX^T + (\sigma^2/\alpha)I)^{-1}(XAX^T)\right\} \\ + \sigma^2 \operatorname{tr}\left\{[B(A\alpha)^{-1} - I]^+\right\} \end{aligned} \tag{5.3.4}$$

and

$$\begin{aligned} \mathrm{MMSPE}(\sigma^2/\alpha, A, B) \geq \sigma^2 \operatorname{tr}\left\{(XAX^T + (\sigma^2/\alpha)I)^{-1}(XAX^T)\right\} \\ - \sigma^2 \operatorname{tr}\{[B(A\alpha)^{-1} - I]^-\}, \end{aligned} \tag{5.3.5}$$

where $\operatorname{tr}\{[(\cdot)]^+\}$ *(resp.* $\operatorname{tr}\{[(\cdot)]^-\}$*) denotes the sum of the absolute values of the positive (resp. negative) eigenvalues of the matrix* $(\cdot)$.

II. *Assume only that* $A \geq 0$.

- *If* $\alpha \geq \alpha^* = \inf\{\alpha : B \leq A\alpha\}$ *then*

$$\mathrm{MMSPE}(\sigma^2/\alpha, A, B) \leq \sigma^2 \operatorname{tr}\left\{(XAX^T + (\sigma^2/\alpha)I)^{-1}(XAX^T)\right\}. \tag{5.3.6}$$

- *If* $\alpha \leq \alpha_* = \sup\{\alpha : A\alpha \leq B\}$ *then*

$$\mathrm{MMSPE}(\sigma^2/\alpha, A, B) \geq \sigma^2 \operatorname{tr}\left\{(XAX^T + (\sigma^2/\alpha)I)^{-1}(XAX^T)\right\}. \tag{5.3.7}$$

The geometric interpretation of α, α_*, and α^* is illustrated in Figure 5.1. The inequalities in the theorem are sharp, because if $A = B$ and $\alpha = 1$ then the lower and upper bounds coincide. In fact that shows that there is actually equality in Theorem 5.1. There are two parts to the bounds. The first part measures the size of A, while the second measures the distance from A to B.

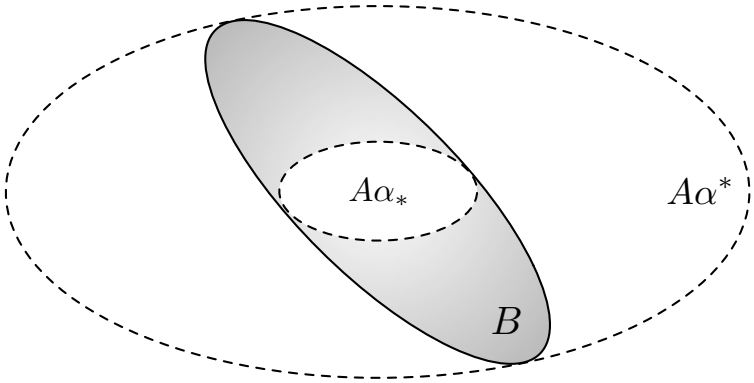


Figure 5.1: Ellipsoids in Theorem 5.2. $A\alpha^*$ is the smallest ellipsoid in the 1-dimensional family $\{A\alpha : \alpha \geq 0\}$ containing the ellipsoid B, while $A\alpha_*$ is the largest contained in B.

5.3.4 Evaluation of the MSPE bounds with Random Matrix Theory

The prediction bounds depend on the function,

$$\mathrm{V}(Q,s) := \mathrm{tr}\left\{(sXQX^T + I)^{-1}(XQX^T)\right\},$$

where $Q \geq 0$ and $s \geq 0$. $\mathrm{V}(Q,s)$ depends only on the eigenvalues of XQX^T. Let $\lambda_i(XQX^T)$ denote the eigenvalues of XQX^T, and define the *empirical spectral distribution* of XQX^T to be the empirical distribution F_n^Q of the eigenvalues. That is, let

$$F_n^Q(x) := \frac{1}{n}\sum_x 1_{\{\lambda_i(XQX^T)\leq x\}}.$$

Then

$$\mathrm{V}(Q,s) = \int \frac{x}{1+sx}\, dF_n^Q(x).$$

So the lower bound in Theorem 5.1 is

$$\min_k \mathrm{MMSPE}(k,A,B) \geq \mathrm{V}(A, 1/\sigma^2).$$

While the upper and the lower bounds in Theorem 5.2 can be written as

$$\mathrm{MMSPE}(\sigma^2/\alpha, A, B) \leq \mathrm{V}(A\alpha, 1/\sigma^2) + \sigma^2\,\mathrm{tr}\{[B(A\alpha)^{-1} - I]^+\}$$

and

$$\mathrm{MMSPE}(\sigma^2/\alpha, A, B) \geq \mathrm{V}(A\alpha, 1/\sigma^2) - \sigma^2\,\mathrm{tr}\{[B(A\alpha)^{-1} - I]^-\},$$

respectively.

Random matrix theory can be used to evaluate $\mathrm{V}(Q,s)$ under some conditions on X and Q. The following is derived from a reformulation of a result of Silverstein (1995). It shows that under certain conditions on X, when p and n are large and of comparable size,

$\mathrm{V}(Q,s)$ essentially only depends on the eigenvalues of $\Sigma^{1/2}Q\Sigma^{1/2}$. Moreover, it shows how to (approximately) calculate $\mathrm{V}(Q,s)$.

Theorem 5.3. *Assume*

1. *$Z = X\Sigma^{-1/2}$ has i.i.d. entries with mean 0, variance $1/n$, and finite fourth moment;*
2. *$p, n \to \infty$ in such a way that $p/n \to \gamma \in [0,\infty)$;*
3. *Q is a random matrix independent of X;*
4. *The empirical spectral distribution F_n^Q of $\Sigma^{1/2}Q\Sigma^{1/2}$ converges almost surely to a non-random limit F.*

Then with probability 1

$$\lim_{p,n\to\infty} \frac{1}{n}\mathrm{V}(Q,s) = \frac{1}{s}(1-\eta)\,, \tag{5.3.8}$$

where $\eta \in [0,1]$ satisfies the equation

$$1-\eta = \gamma\left[1 - \int \frac{1}{1+s\eta x} dF(x)\right]. \tag{5.3.9}$$

A simple consequence of the theorem is that $\mathrm{V}(Q,s)$ is asymptotically upper bounded by an easy to compute function of $\Sigma^{1/2}Q\Sigma^{1/2}$. That is shown in this corollary.

Corollary 5.4.

$$\lim_{p,n\to\infty} \frac{1}{n}\mathrm{V}(Q,s) \le \gamma \int \frac{x}{1+sx} dF(x) \tag{5.3.10}$$

5.4 Proofs

The proofs of the bounds depend on the following expansion of $\text{mspe}(k, A, B)$.

$$\begin{aligned}
\text{mspe}(k, A, \beta) &= k^2 \operatorname{tr}\left\{(XAX^T + kI)^{-2} X\beta\beta^T X^T\right\} \\
&\quad + \sigma^2 \operatorname{tr}\left\{(XAX^T + kI)^{-2}(XAX^T)^2\right\} \\
&= k^2 \operatorname{tr}\left\{(XAX^T + kI)^{-2} X\beta\beta^T X^T\right\} \\
&\quad + \sigma^2 \operatorname{tr}\left\{(XAX^T + kI)^{-2}(XAX^T + kI - kI)(XAX^T)\right\} \\
&= \sigma^2 \operatorname{tr}\left\{(XAX^T + kI)^{-1}(XAX^T)\right\} \qquad (5.4.1) \\
&\quad + k \operatorname{tr}\left\{(XAX^T + kI)^{-2} X[k\beta\beta^T - \sigma^2 A]X^T\right\} .
\end{aligned}$$

The next lemma is also used in the proofs of the upper and lower bounds. It lower bounds $\text{MMSPE}(k, A, B)$ by averaging over $\text{mspe}(k, A, \beta)$ when β is distributed on the surface of the ellipsoid B.

Lemma 5.5.

$$\begin{aligned}
\text{MMSPE}(k, A, B) &\geq \sigma^2 \operatorname{tr}\left\{(XAX^T + kI)^{-1}(XAX^T)\right\} \\
&\quad + k \operatorname{tr}\left\{(XAX^T + kI)^{-2} X[kB - \sigma^2 A]X^T\right\} .
\end{aligned}$$

Proof. Let U be a random variable distributed uniformly on the sphere $\mathbb{S}_{p-1} = \{u \in \mathbb{R}^p : ||u|| = 1\}$ and let $\delta = B^{1/2}U$. Then $\delta\delta^T \leq B$ and with Equation (5.4.1),

$$\begin{aligned}
\text{MMSPE}(k, A, B) = \max_{\beta:\beta\beta^T \leq B} \text{mspe}(k, A, \beta) &\geq \text{mspe}(k, A, \delta) \\
&= \sigma^2 \operatorname{tr}\left\{(XAX^T + kI)^{-1}(XAX^T)\right\} \\
&\quad + k \operatorname{tr}\left\{(XAX^T + kI)^{-2} X[k\delta\delta^T - \sigma^2 A]X^T\right\} .
\end{aligned}$$

Since the distribution of U is rotationally invariant, $\mathbb{E}UU^T = I$ and so $\mathbb{E}\delta\delta^T = B$. Thus, by taking expectation over the distribution of U we have

$$\begin{aligned}\mathrm{MMSPE}(k, A, B) \geq \sigma^2 \operatorname{tr}\left\{(XAX^T + kI)^{-1}(XAX^T)\right\} \\ + k \operatorname{tr}\left\{(XAX^T + kI)^{-2}X[kB - \sigma^2 A]X^T\right\} .\end{aligned}$$ ■

Theorem 5.1

Proof. After applying Lemma 5.5 with $B = A$, we have

$$\begin{aligned}\mathrm{MMSPE}(k, A, B) \geq \sigma^2 \operatorname{tr}\left\{(XAX^T + kI)^{-1}(XAX^T)\right\} \\ + k \operatorname{tr}\left\{(XAX^T + kI)^{-2}X[kA - \sigma^2 A]X^T\right\} .\end{aligned}$$

Let λ_j be the eigenvalues of XAX^T. The right side of the above lower bound is equal to

$$\sum_j \frac{\sigma^2\lambda_j(\lambda_j + k) + k(k - \sigma^2)\lambda_j}{(\lambda_j + k)^2} = \sum_j \frac{\lambda_j(k^2 + \sigma^2\lambda_j)}{(\lambda_j + k)^2} .$$

Differentiating with respect to k we see that the above is minimized when

$$\sum_j \frac{k\lambda_j(\lambda_j + k) - \lambda_j(k^2 + \sigma^2\lambda_j)}{(\lambda_j + k)^3} = 0 .$$

That occurs when $k = \sigma^2$. Thus,

$$\mathrm{MMSPE}(k, A, B) \geq \sigma^2 \operatorname{tr}\left\{(XAX^T + \sigma^2 I)^{-1}(XAX^T)\right\} .$$ ■

We will use the following lemma in the proof of Theorem 5.2.

Lemma 5.6. *Suppose that $k \geq 0$. Then the singular values of*

$$D = kC^T(CC^T + kI)^{-2}C$$

are bounded above by 1.

Proof. Let s_i be the singular values of C. Then the singular values of D are

$$\frac{ks_i^2}{(s_i^2+k)^2} = \left(\frac{k}{s_i^2+k}\right)\left(\frac{s_i^2}{s_i^2+k}\right) \leq 1\,. \qquad \blacksquare$$

Theorem 5.2

Proof. We begin with proving Equation (5.3.4). From Equation (5.4.1),

$$\begin{aligned}\mathrm{MMSPE}(k,A,B) &= \max_{\beta:\beta\beta^T\leq B} \mathrm{mspe}(k,A,B)\\ &\leq \sigma^2 \operatorname{tr}\left\{(XAX^T+kI)^{-1}(XAX^T)\right\}\\ &\quad + k\operatorname{tr}\left\{(XAX^T+kI)^{-2}X[kB-\sigma^2A]X^T\right\}\,. \qquad (5.4.2)\end{aligned}$$

Since $A>0$, we may write the above as

$$\mathrm{MMSPE}(k,A,B) \leq \sigma^2 \operatorname{tr}\left\{(XAX^T+kI)^{-1}(XAX^T)\right\} + \operatorname{tr}(DE)\,, \qquad (5.4.3)$$

where

$$D = kA^{1/2}X^T(XAX^T+kI)^{-2}XA^{1/2}$$

and

$$E = kA^{-1/2}BA^{-1/2} - \sigma^2 I\,.$$

Using the spectral theorem, decompose E as the difference $E = E^+ - E^-$ where E^+ and E^- are both positive matrices. Then

$$\operatorname{tr}(DE) = \operatorname{tr}(DE^+) - \operatorname{tr}(DE^-) = \sum_j s_j(DE^+) - \sum_j s_j(DE^-)\,, \qquad (5.4.4)$$

where $s_1(\cdot) \geq s_2(\cdot) \geq \ldots$ are the singular values of the matrix. Since D is positive,

$$\operatorname{tr}(DE) \leq \sum_j s_j(DE^+)\,.$$

Applying a trace inequality of Von Neumann (see Mirsky 1975 or Theorem IV.2.5 in Bhatia 1997) and then Lemma 5.6 applied to D gives

$$\sum_j s_j(DE^+) \leq \sum_j s_j(D)s_j(E^+) \leq \sum_j s_j(E^+)\,. \tag{5.4.5}$$

Thus,

$$\operatorname{tr}(DE) \leq \sum_j s_j(E^+)$$

and together with Equation (5.4.3) we have shown that

$$\begin{aligned} \operatorname{MMSPE}(k, A, B) \leq \sigma^2 \operatorname{tr}\left\{(XAX^T + kI)^{-1}(XAX^T)\right\} \\ + \operatorname{tr}\{[kA^{-1/2}BA^{-1/2} - \sigma^2 I]^+\}\,. \end{aligned} \tag{5.4.6}$$

Then taking $k = \sigma^2/\alpha$ completes the proof of Equation (5.3.4), because $A^{-1/2}BA^{-1/2}$ and BA^{-1} have the same eigenvalues.

For Equation (5.3.5), we will use a similar argument as above, but begin with Lemma 5.5.

$$\operatorname{MMSPE}(k, A, B) \geq \sigma^2 \operatorname{tr}\left\{(XAX^T + kI)^{-1}(XAX^T)\right\} + \operatorname{tr}(DE)\,, \tag{5.4.7}$$

where D and E are as defined above. Similarly to Equation (5.4.4) and Equation (5.4.5),

$$\begin{aligned} \operatorname{tr}(DE) \geq -\operatorname{tr}(DE^-) &\geq -\sum_j s_j(DE^-) \\ &\geq -\sum_j s_j(E^-) = \operatorname{tr}\{[kA^{-1/2}BA^{-1/2} - \sigma^2 I]^-\}\,. \end{aligned}$$

Letting $k = \sigma^2/\alpha$ completes the proof of Equation (5.3.5).

In the second part of the theorem we only assume that $A \geq 0$. Note that $B \leq A\alpha^*$ implies that in Equation (5.4.2),

$$k \operatorname{tr}\left\{(XAX^T + kI)^{-2}X[kB - \sigma^2 A]X^T\right\} \leq 0$$

when $k = \sigma^2/\alpha \leq \sigma^2/\alpha^*$. Similarly, $B \geq A\alpha_*$ implies that in Lemma 5.5,

$$k \operatorname{tr}\left\{(XAX^T + kI)^{-2}X[kB - \sigma^2 A]X^T\right\} \geq 0$$

when $k = \sigma^2/\alpha \geq \sigma^2/\alpha_*$. ■

Theorem 5.3

Proof.

$$\begin{aligned}\frac{1}{n}\,\mathrm{V}(Q,s) &= \frac{1}{n}\operatorname{tr}\left\{(sXQX^T + I)^{-1}(XQX^T)\right\} \\ &= \int \frac{x}{1+sx}\; dF_n^Q(x) \\ &= \frac{1}{s}\left[1 - \int \frac{1}{1+sx}\; dF_n^Q(x)\right].\end{aligned}$$

$s \mapsto \int \frac{1}{1+sx}\; dF_n^Q(x)$ is the η-transform of F_n^Q (Silverstein and Tulino, 2006) and according to Theorem 3 of Silverstein and Tulino (2006),

$$\int \frac{1}{1+sx}\; dF_n^Q(x) \overset{\text{a.s.}}{\to} \eta\,,$$

where η satisfies

$$\gamma = \frac{1-\eta}{1 - \left[\int \frac{1}{1+s\eta x}\; dF(x)\right]}\,.$$

■

Bibliography

A. Antos and I. Kontoyiannis. Convergence properties of functional estimates for discrete distributions. *Random Structures and Algorithms*, 19:163–193, 2001.

J. Ashbridge and I. B. J. Goudie. Coverage-adjusted estimators for mark-recapture in heterogeneous populations. *Communications in Statistics-Simulation*, 29:1215–1237, 2000.

R. Barbieri, L. M. Frank, D. P. Nguyen, M. C. Quirk, V. Solo, M. A. Wilson, and E. N. Brown. Dynamic analyses of information encoding in neural ensembles. *Neural Computation*, 16(2):277–307, 2004.

G. P. Basharin. On a statistical estimate for the entropy of a sequence of independent random variables. *Theory of Probability and its Applications*, 4:333–336, 1959.

J. Beran. *Statistics for long-memory processes*. Chapman & Hall Ltd., 1994.

R. Bhatia. *Matrix Analysis*, volume 169 of *Graduate texts in mathematics*. Springer-Verlag, New York, 1997.

A. Borst and F. E. Theunissen. Information theory and neural coding. *Nature Neuroscience*, 2(11):947–957, 1999.

P. Bühlmann and B. Yu. Boosting with the L2 loss: Regression and classification. *Journal of the American Statistical Association*, 98(462):324–339, 2003.

T. L. Burr and H. A. Fry. Biased regression: The case for cautious application. *Technometrics*, 47:284–296, 2005.

A. Chao and T.-J. Shen. Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environmental and Ecological Statistics*, 10:429–443, 2003.

D. Conniffe and J. Stone. A critical view of ridge regression. *The Statistician*, 22(3):181–187, Sep 1973.

T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley & Sons, Hoboken, New Jersey, 1991.

A. M. Dale. Optimal experimental design for event-related fmri. *Hum. Brain Mapp.*, 8: 109–114, 1999.

J. G. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical Society of America A*, 2(7):1160–1169, July 1985.

M. R. Deweese and M. Meister. How to measure the information gained from one symbol. *Network: Computation in Neural Systems*, Jan 1999.

N. R. Draper and R. C. van Nostrand. Ridge regression and James–Stein estimation: Review and comments. *Technometrics*, 21(4):451–466, Nov 1979.

I. Frank and J. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, Jan 1993.

Y. Gao, I. Kontoyiannis, and E. Bienenstock. From the entropy to the statistical structure of spike trains. *Information Theory, 2006 IEEE International Symposium on*, pages 645–649, July 2006. doi: 10.1109/ISIT.2006.261864.

D. Gibbons. A simulation study of some ridge estimators. *Journal of the American Statistical Association*, 76(373):131–139, Mar 1981.

M. Goldstein and A. F. M. Smith. Ridge-type estimators for regression analysis. *Journal of the Royal Statistical Society, Series B*, 36(2):284–291, 1974.

I. J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3/4):237–264, 1953.

T. Hastie and R. Tibshirani. *Generalized additive models*. Chapman & Hall Ltd., 1999.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, Feb 1970.

A. E. Hoerl and R. W. Kennard. A note on a power generalization of ridge regression. *Technometrics*, 17(2):269, Jan 1975.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

A. Hsu, S. M. N. Woolley, T. E. Fremouw, and F. E. Theunissen. Modulation power and

phase spectrum of natural sounds enhance neural encoding performed by single auditory neurons. *The Journal of Neuroscience*, 24(41):9201–9211, 2004.

D. Hubel and T. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160:106–154, Jan 1962.

K. N. Kay, T. Naselaris, R. J. Prenger, and J. L. Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, 2008.

M. B. Kennel, J. Shlens, H. D. I. Abarbanel, and E. J. Chichilnisky. Estimating entropy rates with bayesian confidence intervals. *Neural Computation*, 17(7):1531–1576, 2005.

J. Kuks and V. Olman. Minimax linear estimation of regression coefficients. I. *Proceedings of the Estonian Academy of Sciences*, 20(1):480–482, 1971.

J. Kuks and V. Olman. Minimax linear estimation of regression coefficients. II. *Proceedings of the Estonian Academy of Sciences*, 21(1):66–72, 1972.

H. Künsch. Discrimination between monotonic trends and long-range dependence. *Journal of Applied Probability*, 23(4):1025–1030, Jan 1986.

M. Mächler and P. Bühlmann. Variable length Markov chains: Methodology, computing and software. Technical Report 104, ETH Zurich, 2002.

D. M. MacKay and W. S. McCulloch. The limiting information capacity of a neuronal link. *Bulletin of Mathematical Biophysics*, 14:127–135, 1952.

L. Maffei and A. Fiorentini. The visual cortex as a spatial frequency analyzer. *Vision Research*, 13:1255–1267, 1973.

D. McAllester and R. E. Schapire. On the convergence rate of Good–Turing estimators. In *Proceedings 13th Annual Conference on Computational Learning Theory*, pages 1–6, Stanford University, 2000. Morgan Kaufmann, San Francisco.

G. Miller. Note on the bias of information estimates. In H. Quastler, editor, *Information Theory in Psychology: Problems and Methods II-B*, pages 95–100. Free Press, Glencoe, IL, 1955.

L. Mirsky. A trace inequality of John von Neumann. *Monatshefte für Mathematik*, 79(4): 303–306, 1975.

J. Movshon, I. D. Thompson, and D. J. Tolhurst. Spatial summation in the receptive fields of simple cells in the cat's striate cortex. *J. Physiol. (Lond)*, 283:53—77, 1978a.

J. Movshon, I. D. Thompson, and D. J. Tolhurst. Receptive field organization of complex cells in the cat's striate cortex. *J. Physiol. (Lond)*, 283:79—99, 1978b.

I. Nemenman, W. Bialek, and R. de Ruyter van Steveninck. Entropy and information in neural spike trains: Progress on the sampling problem. *Physical Review E*, 69(5):056111, 2004.

S. Nirenberg, S. M. Carcieri, A. L. Jacobs, and P. E. Latham. Retinal ganglion cells act largely as independent encoders. *Nature*, 411(6838):698–701, Jun 2001. doi: 10.1038/35079612.

R. L. Obenchain. Maximum likelihood ridge regression. *Stata Technical Bulletin*, 28:22–36, Jan 1995.

B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, June 1996. doi: 10.1038/381607a0.

A. Orlitsky, N. P. Santhanam, K. Viswanathan, and J. Zhang. On modeling profiles instead of values. In *Conference on Uncertainty in Artificial Intelligence*, pages 426–435, Banff, Canada, 2004.

L. Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15: 1191–1253, 2003.

D. H. Perkel and T. H. Bullock. Neural coding: A report based on an NRP work session. *Neurosciences Research Program Bulletin*, 6:219–349, 1968.

P. Ravikumar, H. Liu, J. Lafferty, and L. Wasserman. SPAM: Sparse additive models. *Neural Information Processing Systems*, 2007.

D. S. Reich, F. Mechler, and J. D. Victor. Formal and attribute-specific information in primary visual cortex. *Journal of Neurophysiology*, 85(1):305–318, 2001.

P. Reinagel and R. C. Reid. Temporal coding of visual information in the thalamus. *Journal of Neuroscience*, 20(14):5392–5400, 2000.

F. Rieke, R. de Ruyter van Steveninck, and W. Bialek. *Spikes: Exploring the Neural Code*. MIT Press, Cambridge, Massachussetts, 1997.

H. E. Robbins. Estimating the total probability of the unobserved outcomes of an experiment. *Annals of Mathematical Statistics*, 39(1):256–257, 1968.

C. Saunders, A. Gammerman, and V. Vovk. Ridge regression learning algorithm in dual variables. In *Proceedings of the 15th International Conference on Macine Learning*, 1998.

C. E. Shannon. The mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 1948.

J. W. Silverstein. Strong convergence of the empirical distribution of eigenvalues of large dimensional random matrices. *Journal of Multivariate Analysis*, 55(2):331–339, 1995.

J. W. Silverstein and A. M. Tulino. Theory of large dimensional random matrices for engineers. *Spread Spectrum Techniques and Applications*, Jan 2006.

G. Smith and F. Campbell. A critique of some ridge regression methods. *Journal of the American Statistical Association*, 75(369):74–81, Mar 1980.

R. W. Sommers. Sound application of regression analysis in chemical engineering. unpublished paper presented at the American Institute of Chemical Engineers Symposium on Avoiding Pitfalls in Engineering Applications of Statistical Methods, 1964.

S. P. Strong, R. Koberle, R. de Ruyter van Steveninck, and W. Bialek. Entropy and information in neural spike trains. *Physical Review Letters*, 80(1):197–200, 1998.

F. E. Theunissen, S. V. David, N. C. Singh, A. Hsu, W. E. Vinje, and J. L. Gallant. Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network*, 12(3):289–316, 2001.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58, No. 1:267–288, 1996.

A. N. Tikhonov. On the stability of inverse problems. *C.R. (Doklady) Acad. Sci. URSS (N.S.)*, 39:176–179, 1943.

J. D. Victor. Asymptotic bias in information estimates and the exponential (bell) polynomials. *Neural Computation*, 12:2797–2804, 2000.

M. J. Wainwright. Sharp thresholds for high-dimensional and noisy recovery of sparsity using the Lasso. Technical Report 709, UC Berkeley, May 2006. To appear in IEEE Trans. Info. Theory.

B. A. Wandell, S. O. Dumoulin, and A. A. Brewer. Visual field maps in human cortex. *Neuron*, 56(2):366–383, 2007.

N. Wiener. *Cybernetics: or Control and Communication in the Animal and the Machine*. John Wiley & Sons, 1948.

D. Wolpert and D. Wolf. Estimating functions of probability distributions from a finite set of samples. *Physical Review E*, 52(6):6841–6853, 1995.

A. J. Wyner and D. Foster. On the lower limits of entropy estimation. *Unpublished manuscript*, 2003.

S. Zahl. Jackknifing an index of diversity. *Ecology*, 58:907–913, 1977.

P. Zhao and B. Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2567, 2006.