To appear as: Chapter 36 of *The New Cognitive Neurosciences*, 4th edition
Editor: M. Gazzaniga. MIT Press, October, 2009.

Optimal Estimation in Sensory Systems

Eero P. Simoncelli

Center for Neural Science, and
Courant Institute of Mathematical Sciences,
New York University
New York, NY 10003

14 May 2009

Abstract: A variety of experimental studies suggest that sensory systems are capable of performing estimation or decision tasks at near-optimal levels. In this chapter, I explore the use of optimal estimation in describing sensory computations in the brain. I define what is meant by optimality and provide three quite different methods of obtaining an optimal estimator, each based on different assumptions about the nature of the information that is available to constrain the problem. I then discuss how biological systems might go about computing (and learning to compute) optimal estimates.

The brain is awash in sensory signals. How does it interpret these signals, so as to extract meaningful and consistent information about the environment? Many tasks require estimation of environmental parameters, and there is substantial evidence that the system is capable of representing and extracting very precise estimates of these parameters. This is particularly impressive when one considers the fact that the brain is built from a large number of low-energy unreliable components, whose responses are affected by many extraneous factors (e.g., temperature, hydration, blood glucose and oxygen levels).

The problem of optimal estimation is well studied in the statistics and engineering communities, where a plethora of tools have been developed for designing, implementing, calibrating and testing such systems. In recent years, many of these tools have been used to provide benchmarks or models for biological perception. Specifically, the development of signal detection theory led to widespread use of statistical decision theory as a framework for assessing performance in perceptual experiments. More recently, optimal estimation theory (in particular, Bayesian estimation) has been used as a framework for describing human performance in perceptual tasks.

Acknowledgements: Thanks to Alan Stocker, Martin Raphan, Mehrdad Jazayeri, and the section editors, Tony Movshon and Brian Wandell, for helpful comments and suggestions. This work was financially supported by the Howard Hughes Medical Institute, the National Institutes of Health (EY018003), and the Sloan-Swartz Center for Theoretical Visual Neuroscience at New York University, but the views expressed herein are my own.

In this chapter, I'll explore the use of optimal estimation in describing sensory computations in the brain. In the first half, I'll define what I mean by optimality, and develop three quite different formulations for obtaining an optimal estimator. In the second half, I'll ask how biological systems might go about computing optimal estimates. This is not intended as a complete review of this rich multi-disciplinary topic, and I apologize in advance to the many authors whose important contributions I've neglected to mention. Instead, my purpose is to clarify and resolve a number of myths and misunderstandings about optimal estimation, and to offer a personal perspective on the relationship between these concepts and the design and function of biological sensory systems.

1 Definition and formulations of optimal estimation

A common problem for systems that must interact with the world (including both biological organisms, and man-made devices) is that of obtaining estimates of environmental properties, x, from sensory measurements, m. An *estimator* is simply a deterministic function, $f(m)$ that maps measurements to values of the variable of interest. If x is a binary variable, then the estimator reduces to a *decision* function. Generally, the measurements are assumed to be corrupted by noise, which could arise from a number of sources, including the signal itself (e.g., the quantization of light into photons, when one is interested in knowing the light intensity), the transduction mechanism, or variability within the neurons that are transmitting and computing with this information (see (Faisal *et al.*, 2008) for a recent review of noise in the nervous system).

Our primary question is: how does an organism select and implement a good estimator, or (more optimistically) the best estimator? To address this, we'll have to state explicitly what we mean by "best". The traditional statistical formulation of the best estimator is the one that minimizes the average value of a pre-defined loss (cost) function, $L(x, f(m))$. The loss function specifies the cost of generating an estimated value of $f(m)$ when the true value is x. It is generally assumed to be positive, and equal to zero only when the estimate is equal to the true value.

1.1 Regression formulation

Suppose we wanted to build a machine that could perform optimal estimation of x, given a noisy measurement m[1]. We can imagine "training" this machine by showing it many signal-measurement pairs, $\{x_n, m_n\}$. Typically, we imagine that each measurement arises from its associated true value through some sort of noisy transformation. Figure 1(a) illustrates such a set of training data.

An estimator, f, attempts to invert the measurement process, mapping measurements m

[1]Throughout, x and/or m can be scalar-valued or vector-valued quantities. For example, m might represent responses of a population of neurons.

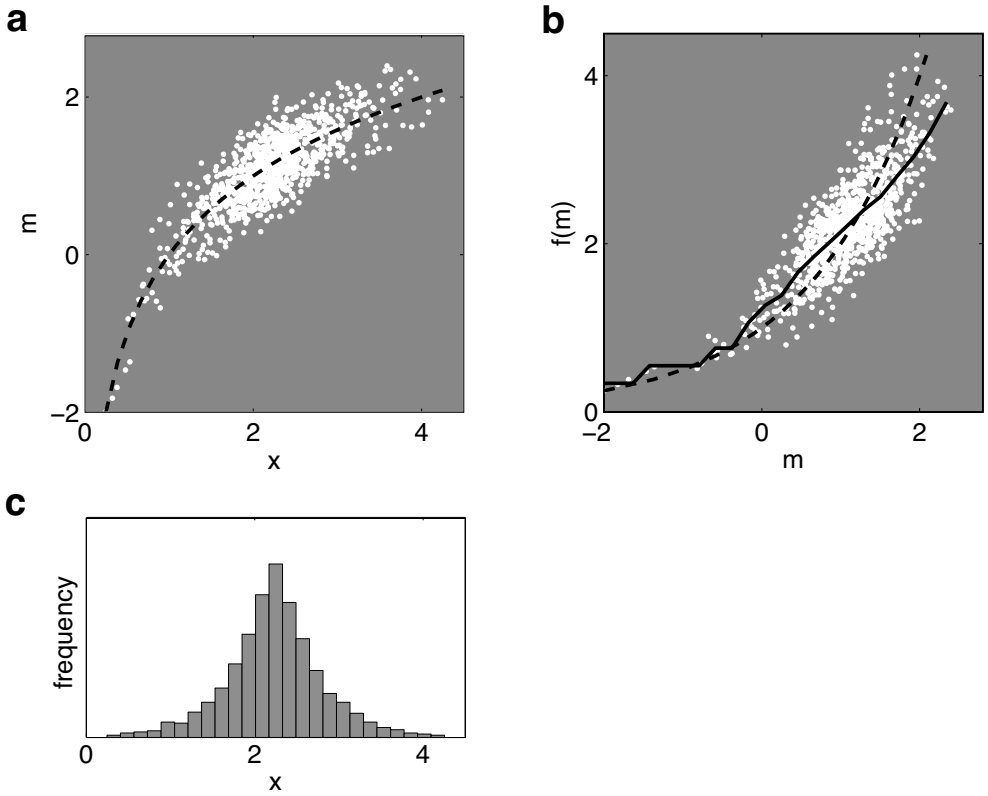

Figure 1. Regression formulation of the optimal estimation problem, illustrated for a one-dimensional signal and measurement. **(a)** The measurement process (also known as the *encoding* process). We assume a set of data pairs (plotted points), $\{x_n, m_n\}$, indexed by $n \in [1, 2, ...N]$, representing true signal values and associated noisy measurements. The dashed line indicates the average measurement as a function of the true signal value. *(b)* The estimation (or *decoding*) process. The estimator, $f(m)$ maps measurements back to estimated signal values. The optimal estimator (solid line) does this so as to minimize a specified loss function. Note that this need not be (and is generally not) the inverse of the average measurement function (dashed line). Note also that the optimal estimator will depend on the signal values that are included in the data set, which are summarized by the histogram shown in **(c)**.

back to signal values x (Fig. 1(b)). This mapping is deterministic: each measurement leads to a unique estimate. But if we hold the signal value fixed, and make a set of estimates (each arising from a different measurement), these estimates will fluctuate because of the variability in the underlying measurements. The optimal estimator is the one that minimizes the average loss over these examples:

$$f_{\text{opt}} = \arg\min_f \frac{1}{N} \sum_n L(x_n, f(m_n))$$

We will refer to this as a "regression" estimator - a special case is the linear regression solution, which arises when L is the squared error. An example optimal estimator is indicated by the solid line in Fig. 1(b). Note that this transformation is not the same as the inverse of the transformation to average measurements (i.e., the inverse of the dashed line shown in Fig. 1(a)).

Of course, the precision with which we can constrain the function f depends on how much data we have. Loosely speaking, the usual approach is to restrict f to be sufficiently simple (e.g., smooth, or defined by a small number of parameters) that the available data will constrain it properly. For example, the estimator shown in Fig. 1(b) was computed by binning the data (as a function of m), and computing the best estimate value for each bin. More formally, we might specify a restricted set of possible functions (denoted $\mathcal{F}$) from which the solution will be selected.[2] Finally, note that the solution we obtain will depend on the *distribution* of data. If the set of training examples includes many x values clustered in a particular region of the space, then the average loss will contain many terms from that region, and the optimization process will thus attempt to reduce the estimation errors there, typically at the expense of larger errors elsewhere. This suggests that the training examples should be selected so as to represent the distribution of values that might be encountered in the environment.

The regression formulation is appealing because it is simple and intuitive. Its primary limitation is that it requires *supervised* training. That is, obtaining an optimal estimator relies on a training set of noisy measurements, m_n, each accompanied by its corresponding correct signal value, x_n. Supervised learning for estimation and classification problems has been well-studied. A standard example is the problem of learning an input-output relationship with a simplified network of artificial neurons, for which the optimal solution may be obtained by back-propagation (essentially, a form of stochastic gradient descent on the objective function). But this requires large amounts of data, especially when learning muti-dimensional functions.

From the biological/behavioral perspective, a fully supervised training paradigm also seems implausible. Although most organisms absorb enormous amount of sensory data during their lifetimes, the information received regarding "correct" answers would seem to be relatively sparse. For example, consider the problem of estimating the distance to a nearby

[2] A more sophisticated solution would adjust the smoothness of the estimator adaptively to the amount of data. This issue of model complexity and its relationship to to learning from data is fundamental to the study of machine learning.

object based on visual input. We can compare our estimate to the one obtained by reaching out and touching the object. But the amount of this kind of feedback we receive seems vastly insufficient to train the enormous cascade of neurons that are involved in estimating distances from visual input. Similarly, optimization through natural selection (with surviving organisms passing preferred solutions to their offspring genetically) seems implausible, both because of the time required and because genetic material seems unlikely to contain sufficient information to encode even a fraction of the detailed connectivity of those neurons.

Instead, it seems that evolution has endowed the brain with powerful capabilities for *unsupervised* learning (based on noisy measurements alone), and that this is used to supplement and bolster the supervised learning that may be used in the relatively infrequent cases for which the correct answers are known. Unsupervised learning is a heavily studied topic in machine learning (e.g., Hinton & Sejnowski, 1999), and methods have been developed for learning patterns in data, mostly for purposes of optimal coding or clustering/categorization. Perhaps less well known is the fact that optimal estimators may also be written in unsupervised form. To explain this, I'll turn first to a probabilistic formulation of the problem.

1.2 Probabilistic (Bayesian) Formulation

When we describe optimality in terms of minimizing an objective function over a training data set, we usually have in mind that this set is representative of future data we will encounter. This notion may be formalized by describing both the training and future data as samples randomly drawn from a common probability distribution. The law of large numbers tells us that as the number of data pairs grows, the original regression objective function will converge to the expected value (mean) of the loss function, integrated over all possible combinations of x and m:

$$f_{\text{opt}} = \arg\min_f \iint P(x,m)\, L\left(x, f(m)\right)\, dx\, dm$$

Unlike the regression formulation, which is written directly in terms of data, the probabilistic formulation is written in terms of a continuous probability density. And since this formulation effectively results from assuming infinite amounts of data, the smoothness constraint that was necessary for selecting an estimator in the regression case is now optional.

The probabilistic objective function may be simplified by using the definition of conditional probability to rewrite the joint density as a product of the marginal density of m, and the conditional density of x given m (known as the *posterior* distribution):

$$f_{\text{opt}} = \arg\min_f \int P(m) \int P(x|m)\, L(x, f(m))\, dx\, dm.$$

If the estimator is unrestricted, then we may ignore the outer integral and optimize the estimator separately for each measurement value:

$$f_{\text{opt}}(m) = \arg\min_f \int P(x|m)\, L(x, f(m))\, dx.$$

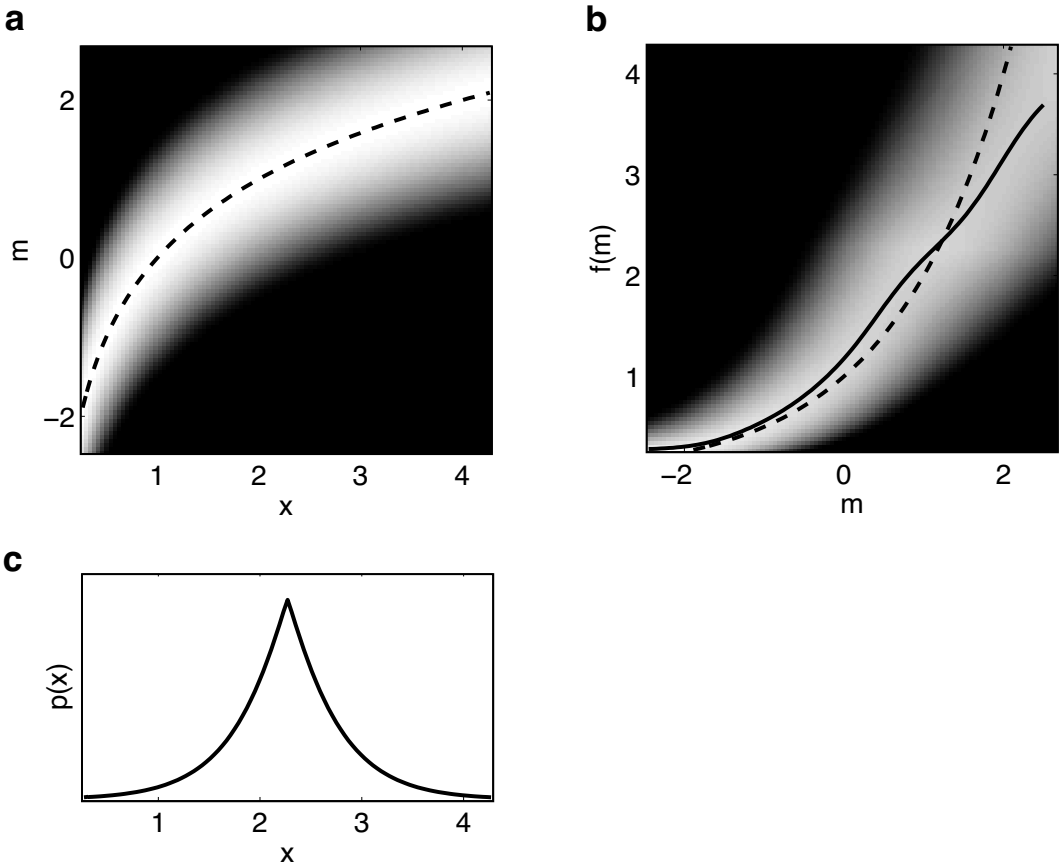

Figure 2. Bayesian formulation of the optimal estimation problem. **(a)** the measurement density, $P(m|x)$, shown as a grayscale image, where intensity indicates log probability. The dashed line indicates the mean of the density, as a function of x. **(b)** the posterior density, $P(x|m)$. Solid line indicates mean of the density, and dashed line indicates the (inverted) mean of the measurement density in (a). **(c)** the prior density, $P(x)$.

That is, for each measurement, the best estimate is the one that minimizes the expected value of the loss function over the posterior distribution for that measurement.

Finally, the posterior distribution may be rewritten in terms of densities that are more naturally associated with process from which the data arise. Specifically, we can the describe measurement noise using a conditional probability $P(m|x)$. This *measurement density* expresses the probability of m, for each value x of the signal. If we think of it the other way around, holding the measurement fixed and reading off a function of the signal, this is known as a *likelihood function*. Now Bayes' rule may be used to express the posterior in terms of the measurement density and the *prior* distribution $P(x)$, which expresses the probability of occurrance of value x in the world:

$$f_{\text{opt}}(m) = \arg\min_f \int \frac{P(m|x)P(x)}{P(m)} L(x, f(m))\, dx. \tag{1}$$

An example of the Bayesian solution, based on the same distributions used to generate the data in Fig. 1, is illustrated in Fig. 2. To provide some intuition, it is worth mentioning several well-known special cases.

Quadratic error (least squares) solution. The most common case used in the engineering community is the *least squares* loss function, $L(x, f(m)) = (x - f(m))^2$. In this case, the optimal estimate (which may be derived by differentiating the objective function and setting equal to zero) is simply the mean of the posterior:

$$f_{\mathrm{LS}}(m) = \int x\, P(x|m)\, dx.$$

It is worth mentioning that in the special case of a jointly Gaussian probability density over signal and measurement, this solution turns out to be a linear function of the measurement (the solution is the same as that of our next example).

Linear estimator, quadratic error. Now consider what happens when the estimator is restricted to be a linear function of the measurement. The *Linear Least Squares* solution is

$$f_{\mathrm{LLS}}(m) = \frac{\sigma_{xm}}{\sigma_{mm}}\, m.$$

The linear solution relies only the cross-correlation between signal and measurement, and between the measurement and itself, and not on full knowledge of the posterior density. This result extends naturally to multi-dimensional inputs or outputs.

Maximum probability solution Suppose the loss function penalizes all errors equally, except for the correct answer (which incurs no penalty). Then the solution is the maximum of the posterior density, known as the *maximum a posteriori* (MAP) estimator:

$$f_{\mathrm{MAP}}(m) = \arg\max_x P(x|m).$$

The probabilistic formulation expresses the estimation problem in terms of four natural ingredients:

- the prior, $P(x)$, which represents the probability of encountering different signal values in the world,
- the measurement density, $P(m|x)$, which represents the (probabilistic) relationship between the signal and measurement,
- the loss function, $L(x, f(m))$, which represents the cost of making errors,
- and the family of functions $\mathcal{F}$ from which the estimator is to be chosen (this ingredient may not be required for the Bayesian solution, which effectively operates under conditions of infinite data).

Note that although the regression solution of the previous section was developed directly from pairs of input-output data, it is also implicitly relying on these same ingredients. Specifically, it is effectively based on the joint probability density of signal and measurement,

which is equivalent to the product of the prior and likelihood. And, as stated in the previous section, it also requires the specification of a loss function and a family of functions from which the solution is to be drawn.

It is worth emphasizing the most obvious implication of this ingredient list, since it is often misunderstood: Optimality is not a fixed universal property of an estimator, but one that depends on each of these defining ingredients: Statements about optimality that do not fully specify the ingredients are therefore relying on hidden assumptions. For example, many authors assume that optimality implies that an estimator must be unbiased (that is, on average, computes the correct value). But many well-known optimal estimators exhibit bias (in fact, all of the estimator examples mentioned above can exhibit bias, depending on the specific choices of prior and measurement densities).[3]

Despite the appealing decomposition of the problem into intuitively sensible ingredients, there are drawbacks to the probabilistic formulation. In particular, the reliance of the regression solution on supervised training data has been replaced with reliance on knowledge of two abstract probability densities. Since the measurement density is a property of the sensory system, we might imagine learning it from a set of calibration measurements, or assuming that it is a fixed property of the device. On the other hand, the Bayesian formulation is often criticized for reliance on an unknown (and perhaps unknowable) prior distribution,[4] and this criticism is further inflamed by the many examples in the literature that introduce a prior as an ad-hoc function that may be freely chosen to make the solution tractable. In the view set out here, the prior is meant to capture the statistical structure of some aspect of the world, an assumption that is only slightly stronger than the assumption that the training data in the regression estimator are representative of future data that the system will need to process.

1.3 Unsupervised learning of optimal estimators

The Bayesian view assumes all ingredients of the problem, including the prior, are known. If the prior is meant to correctly represent the distribution of signal values in the world, it must presumably be learned from measurements. Engineers that need to design real systems generally follow one of several practical solutions: (1) directly measure the distribution of signal values that might be encountered by the device, and use a model of this empirical distribution for the prior; (2) assume a prior distribution of some parametric form, and then adjust the parameters so as to best explain the observed distribution of noisy measurements; or (3) assume an estimator of some parametric form, and adjust this directly to improve performance on observed data. The first solution requires a separate set of uncorrupted

[3] Bias can arise because the estimator is restricted to lie in a family that does not include the best solution. But it can also arise due to asymmetries in the posterior (e.g., due to the influence of a prior) or cost function, in which case it should be viewed as desirable. E.T. Jaynes described the insistence on unbiased estimators as a "pathology of orthodox statistics" (Jaynes, 2003).

[4] By presenting the Bayesian and regression formulations side by side, I hope to alleviate some of the tensions between the Bayesian and "frequentist" viewpoints. See (Jaynes, 2003) for further discussion.

signal measurements, and thus does not seem relevant to biological systems which are presumably only able to make measurements through the same noisy sensors from which they will be making their estimates. The second solution is generally known as *empirical Bayesian* estimation, since the prior is obtained from noisy training data. The third solution, as described, is simply the regression solution, which relies on supervised training data in order to measure and optimize the estimator performance. Remarkably, it can sometimes be rewritten in an unsupervised form. Below, I'll consider the second and third solutions in more detail.

The empirical Bayes formulation assumes a prior of a known parametric form[5], optimizes the parameters to fit the (noisy) measurements, and then uses this optimized prior to obtain the estimator (by minimizing Eq. (1)). The prior parameters are typically chosen to make the observed data as consistent as possible with the model, and this is usually achieved in practice by maximizing the probability of the observed data:

$$\theta_{\text{opt}} = \arg\max_{\theta} \prod_{n} \int P_\theta(x) P(m_n|x)\, dx.$$

Beyond the potential difficulties associated with computing this optimization, the introduction of this probabilistic cost function is a bit inconsistent, since the prior that best explains the data is not necessarily the one that will lead to the best estimator. Nevertheless, empirical Bayes solutions are often quite successful in situations where the data are sufficient to strongly constrain the prior parameters.

The third solution mentioned above can, in some cases, be obtained from unsupervised training (Raphan & Simoncelli, 2007). As such, we'll refer to it as *unsupervised regression*. The derivation of the general form is somewhat complex, but the simplest case (due to Charles Stein(Stein, 1981)) arises in the context of an additive Gaussian noise model and a squared-error loss function, and can be written quite simply. Stein showed that the mean squared error in this case may be rewritten in a form that depends only on the measurements, and not the signal:

$$\iint P(x,m)\,[x - f(m)]^2\; dx\; dm = \int P(m)\,[g(m)^2 + 2\sigma^2 g'(m) + \sigma^2]\, dm,$$

where $g(m) = f(m) - m$, and σ is the standard deviation of the additive Gaussian noise (assumed known). This remarkable result implies that the squared error may be approximated by averaging over measured (noisy) data, without knowledge of the correct answers (i.e., *unsupervised*) and with no assumption about the prior, $P(x)$. This implies that we can select an optimal estimator f (or equivalently, g) by minimizing the integrand above,

[5] A less well-known form of empirical Bayes estimation arises from rewriting the estimator directly in terms of the distribution of measurements, which can be approximated from the observed data (Miyasawa, 1956). This form may be derived from the prior-free estimator described previously (Raphan & Simoncelli, 2007).

	meas. $\{m_n\}$	sig. $\{x_n\}$	meas. prob. $P(m\|x)$	sig. prob. $P(x)$	loss fn. L	est. family $\mathcal{F}$
Regression	✓	✓			✓	parametric
Bayesian			✓	✓	✓	
Empirical Bayesian	✓		✓	parametric	✓	
Unsupervised (e.g., Stein) regression	✓		restricted		quadratic	parametric

Table 1. Ingredients required for specifying/learning various formulations of optimal estimator (see text for definitions). Columns correspond to: measurement data, signal data (paired with measurement data), measurement probability, signal (prior) probability, loss function, and family of functions from which the estimator is chosen. Checkmarks indicate that the ingredient is required but unrestricted. Unlabeled spaces indicates the ingredient is not needed. Note that the unsupervised regression estimator has been derived only for certain specific measurement densities, and squared-error loss (Raphan & Simoncelli, 2007).

averaged over a set of noisy measurement data. As with the original regression solution, the estimator must be restricted sufficiently (e.g., drawn from some parametric family) so that it can be constrained by the available data. Analogous expressions may be derived for a number of other measurement probabilities (Raphan & Simoncelli, 2007).

In both the empirical Bayesian and the unsupervised regression formulations, we have exchanged the supervised data pairs required by the standard regression solution for unsupervised data and a *known* (or previously calibrated) description of the measurement density, $P(m|x)$. We can thus view these solutions as a compromise between the data-oriented regression form and the more abstract Bayesian form. A summary of the ingredients required by each of the optimal estimators introduced thus for is provided in Table 1.

2 Optimal estimation in the brain

In this section, we ask how the optimal estimation formulations developed in the previous section can be used in modeling biological sensory systems, and how these models may be tested experimentally. These questions can be addressed at many levels, and in this short chapter, I will not attempt to provide a complete overview. Rather, I'll describe few published results, and try to explain what I see as some of the more important challenges that we currently face in this endeavor.

The concept that sensory perception arises through the fusion of incoming sensor measurements with one's prior experience is often attributed to Hermann von Helmholtz (von Helmholtz, 1925). Although his descriptions are qualitative in nature, and do not mention noise or loss functions, they do capture the essence of the Bayesian formulation described in the previous section. This interpretation of perception seems to have lain dormant from

von Helmholtz's day until the 1950's, when E.T. Jaynes, a statistically-minded physicist, submitted an article to the IRE Transactions on Information Theory, in which he proposed that Bayesian estimation might be used as a framework for modeling sensory transformations (Jaynes, 1957). It was rejected by the journal (on grounds that it was too speculative) and the concept appears to have lain dormant for another 30 years! In the interim, perceptual psychologists began using Signal Detection Theory as a framework for analyzing psychophysical data (Green & Swets, 1966), and for providing an upper bound on performance. This methodology often does not include explicit loss functions, and rarely includes a prior, but the formalization nevertheless represents an important step toward the optimal estimation framework.

2.1 Perceptual Bayesianism

In the 1980's and 90's there was a dramatic revival of the Bayesian methodology across many fields, and perceptual science was one of these. A variety of experiments have aimed to test optimality of human estimation judgements by comparing performance to an "ideal observer" model (e.g., Barlow, 1980; Geisler, 1989; Kersten, 1990; Knill *et al.*, 1990). A number of reviews document the activity to date (Knill & Richards, 1996; Maloney, 2002; Mamassian *et al.*, 2002; Kersten *et al.*, 2004; Körding, 2007), and this endeavor has been expanded by recent activity in "Neuroeconomics", a cross-disciplinary enterprise that aims to characterize decision-making processes with respect to prior probabilities and reward contingencies (Glimcher *et al.*, 2008).

What does it mean to say that a human subject is performing optimally? As I've emphasized in the first part of this chapter, the definition of the word *optimal* requires specification of a set of ingredients: the measurement probability, the prior, the loss function, and (in some cases) a family of estimators. Specifying these ingredients for a human observer performing a particular task is often difficult or impossible. For example, specifying the measurement probability requires knowledge of how the signal of interest is represented within the brain (including a specification of the noise). In some experiments, investigators have incorporated noise into the stimulus, which can provide insights into the properties of internal noise (and thus the measurement probability) (e.g., Pelli & Farell, 1999; Körding & Wolpert, 2004). The specification of an appropriate family of estimators should be determined by the set of computations that can potentially be performed by neurons, but we currently lack a detailed description of this set.

The loss function can pose more substantial difficulties. Subjects may differ inherently in the way they behave in an experimental situation (e.g., consider personality traits such as risk aversion vs. thrill-seeking). And even in cases where the investigator attempts to control for this by building a loss function directly into the design of the experiment (for example, by paying/penalizing subjects for correct/incorrect answers), one does not know whether or how the subject will learn and internalize a los function, what type of training (e.g., supervised vs. unsupervised) this would require, and how long it would take.

Last, consider the prior, for which one can ask the same questions as were asked for the the loss function (whether/how a subject internalizes it, what type of training is required, over what time scales). We might imagine that the observer operates according to a prior that was obtained over a relatively slow timescale (much longer than the duration of a typical experiment, say, on a developmental or evolutionary timescale). In this case, the investigator might attempt to measure it from the environment (or derive it from a model of the environment). At the other extreme, we might imagine that the subject's internal prior representation is quite flexible and that, over the duration of the experiment, they internalize the distribution of stimuli that have been presented. In fact, many experiments are designed so that the subject can learn the distribution of signal values over a set of training trials.

In the context of learning a prior, one apparent paradox seems worth mentioning. Adaptation to stimuli that persist for timescales of seconds or minutes has been found, for every sensory modality and for a wide range of stimulus configurations, to produce substantial changes in subsequent perception. For example, adaptation to a visual stimulus, say, of a given orientation, induces biases in the perceived orientation of subsequently viewed stimuli. These biases are generally *repulsive*: the perceived orientation of a post-adaptation stimulus is pushed away from that of the adaptor. If we were to interpret the adaptation as a means by which the system updates its prior probability distribution for orientation, we might expect that heavy exposure to the adapting stimulus would cause an increase in the internally represented prior probability of that stimulus, which should then lead to an *attractive* bias in subsequent perception! Thus, adaptation over these time scales appears to be inconsistent with learning of prior probabilities. An alternative interpretation, still within the Bayesian framework, is that adaptation effects correspond to a change in the likelihood function (Stocker & Simoncelli, 2006b).

Given the difficulty of specifying the ingredients of an optimal estimator, one can consider an alternative experimental approach for exploring optimal estimation theories of perception. Specifically, one can ask "for what choices of ingredients would the subject's behavior be considered optimal?". The trial-averaged estimates of a human subject do not place a sufficient constraint on the problem to answer this question. In particular, the average response of an estimator does not uniquely determine the prior, likelihood , and loss function that could have been used to define it. But if one assumes, say, a quadratic loss function, and measures not just the average response but the the full distribution of estimates, then it is possible to extract a prior (Paninski, 2006) or even the likelihood and prior (Stocker & Simoncelli, 2006a) from the psychophysical data.

Ultimately, it seems important to move beyond the initial question of *whether* an observer is optimal, to examine the prior and cost conditions under which the observer may be considered optimal, and the flexibility of that optimality. Given that the estimators must operate under changing conditions, we'd like to know: (1) which ingredients of the problem may be learned or adjusted, and what type of adjustments can be made (e.g., Körding & Wolpert, 2004)), (2) what type of learning is possible (e.g., supervised, unsupervised, direct verbal communication), (3) over what time scales this learning occurs, and (4) whether the

observer is able to switch between estimators (or ingredients of estimators, such as the prior) that have been previously learned (e.g., Körding & Wolpert, 2004; Maloney & Mamassian, 2009).

2.2 Physiological implementation

In addition to interpreting perception in the context of optimal estimation, we can also consider how such optimal computations might be implemented in the brain. The responses of sensory neurons are commonly described in terms of their selectivity to particular parameters of the stimulus. In most cases, no single neuron is responsible for encoding a stimulus parameter. Instead, the parameter is jointly represented by a population of neurons with different tuning properties, and thus any estimate of the parameter requires a combination of information across many cells, if not the whole population. Over the past 20 years, the theoretical neuroscience community has been exploring the means by which neural responses might be optimally "read out" in order to explain behavior (e.g., Georgopoulos *et al.*, 1986; Bialek *et al.*, 1991; Anderson & van Essen, 1994; Seung & Sompolinsky, 1993; Potters & W.Bialek, 1994; Salinas & Abbott, 1994; Sanger, 1996; Snippe, 1996; Shadlen *et al.*, 1996; Rieke *et al.*, 1997; Rieke & Baylor, 1998; Zhang *et al.*, 1998; Zemel *et al.*, 1998; Platt & Glimcher, 1999; Gold & Shadlen, 2000; Simoncelli, 2003; Pouget *et al.*, 2003; Bialek & van Steveninck, 2005; Jazayeri & Movshon, 2006).

It is worth noting that, despite my emphasis on the probabilistic formulation of the problem, computation of optimal estimates does not necessarily require that the brain explicitly represent or compute probabilities. As described in the previous section, given the four ingredients of the optimal estimation problem, an estimator is just a fixed deterministic function that maps noisy measurements to estimated values. If these ingredients are fixed and unchanging, the brain could learn to compute the optimal estimator using either regression, or one of the two unsupervised methods described previously, and would not need to explicitly calculate or represent probabilities!

In general, we imagine that the ingredients of the optimal estimation problem *do* change: the loss function is typically task-dependent, the prior may change gradually (or even suddenly, for example, when the observer moves into a different environment), and the measurement probability may also change (due to physiological changes in the neural substrate). But even under these conditions, the solution need not explicitly require the calculation or representation of probabilities. If the prior and/or measurement probabilities are parametric, then the optimal estimator is just a function that depends on those parameters. In this case, the parameters may be computed from previous measurements, either through supervised regression, or using the empirical Bayesian formulation. Again, this does not require explicit representations of probabilities.

Although explicit probability representation is not required for optimal estimation, much published work assumes it. A simple means of encoding probabilities, suggested by a number of authors (e.g., Hinton, 1992; Simoncelli, 1993; Foldiak, 1993; Anderson & van Essen,

1994; Sanger, 1996; Gold & Shadlen, 2000; Weiss & Fleet, 2002; Eliasmith & Anderson, 2002; Sahani & Dayan, 2003; Simoncelli, 2003; Barber *et al.*, 2003), assumes that the firing rates of each neuron directly represents the probability (or the log probability) of a particular stimulus parameter value. In this view, the population encodes (either directly, or through a set of linear basis functions) a probability distribution over the parameter. Downstream neurons could explicitly compute an estimate from this population, by computing, for example, the population mean (a weighted sum) or peak (a "winner-takes-all"). For example, suppose that the firing rates in a given neural population represent the posterior probability evaluated at a set of different signal values. The mean of the density may be computed as a weighted sum over the responses (the weights will be determined by both the signal values at which the posterior is sampled, and also the portion of the signal space covered by each neuron). Or subsequent stages could operate on the posterior information, postponing the explicit determination of an estimate until it is needed. In either case, the prior in this model can be adjusted by changing the gain on each of the neurons.

The explicit representation of uncertainty, through the breadth and shape of the population response, along with the possibility of linear readout rules, are conceptually appealing features of this framework. But a detailed model of this form needs to address the inconsistency of directly representing probability values with neural responses that are noisy (e.g., see Sahani & Dayan, 2003). In addition, the responses of many visual neurons do not seem consistent with direct representation of posterior probability. For example, the shape of orientation tuning curves in area V1 neurons is preserved under changes in stimulus contrast. But lowering the stimulus contrast results in larger variance in the estimation of orientation, which implies a broadening of the posterior.

A widely followed alternative formulation represents probabilities implicitly, using the noisy responses of a population of neurons (e.g., Seung & Sompolinsky, 1993; Salinas & Abbott, 1994; Zhang *et al.*, 1998; Zemel *et al.*, 1998; Pouget *et al.*, 2003; Jazayeri & Movshon, 2006; Ma *et al.*, 2006).[6] Consider a population of N neurons whose responses represent the measurement in an optimal estimation problem. Suppose the mean firing rate of each neuron is determined by a tuning function $f_n(x)$, where x is the stimulus variable of interest. Suppose also that the number of spikes emitted by each of these neurons in a unit time interval to any given stimulus are statistically independent, and follow a Poisson distribution. Then the joint likelihood function is the product of the individual Poisson probabilities:

$$P(\vec{r}|x) = \prod_n \frac{f_n(x)^{r_n}}{r_n!} e^{-f_n(x)}$$

where $\vec{r}$ is a vector containing the population spike counts.

At this point, one might use unsupervised (or supervised) regression to learn the optimal estimator from data without specifying or learning a prior probability. Alternatively, we can follow the Bayesian formulation, multiplying the likelihood by a prior $P(x)$, dividing

[6] Note that this formulation is encoding the uncertainty due to the noise in the population response, but not the noise or structural ambiguities in the input (e.g., Sahani & Dayan, 2003).

by $P(\vec{r})$, and taking the (negative) log to write the log-posterior density:

$$-\log P(x|\vec{r}) = \sum_n \log(r_n!) - \log P(\vec{r}) - \sum_n r_n \log f_n(x) + \sum_n f_n(x) - \log P(x)$$

Now consider the MAP estimator. The first two terms do not contain the stimulus variable, x, so we can drop them, arriving at an objective function that can be maximized over x to obtain the estimate:

$$E(x) = -\sum_n r_n \log f_n(x) + \sum_n f_n(x) - \log P(x) \tag{2}$$

The first term is a sum of the observed spike counts, weighted by the log tuning curve value of each neuron (Zhang *et al.*, 1998; Jazayeri & Movshon, 2006). The second term is the sum of the tuning curves, and the third is the (negative) log of the prior. Much of the previous work on population coding has focused on the special case of orientation representation in V1 neurons, or selectivity for motion direction in MT neurons, and in these cases the prior over orientation is typically assumed to be constant, as is the sum of the tuning curves (e.g., Zemel *et al.*, 1998; Jazayeri & Movshon, 2006) These assumptions allows one to ignore the last two terms, and the resulting log-posterior objective function reduces to a simple weighted sum of spike counts, consistent with earlier proposals for linear readout (Bialek *et al.*, 1991; Anderson & van Essen, 1994; Rieke *et al.*, 1997). Note that later stages of processing (i.e., the estimator) presumably would not have access to the tuning curves, and thus could not compute the linear weights directly by taking the log. Instead the proper weights could again be learned using unsupervised (or supervised) regression.

The linear representation of log probability is especially convenient for fusing independent sources of information (e.g., accumulating evidence over time (Jazayeri & Movshon, 2006; Beck *et al.*, 2008), or combining evidence from multiple modalities (Ma *et al.*, 2006)). In these cases, one wishes to multiply the associated probability densities, which can be done by simply adding spike counts. Another common operation one would like to perform on probability densities is to integrate them, either for purposes of computing the probability of a particular event or hypothesis, or for reducing a density on multiple variables to one on a subset of those variables (known as *marginalization*). Given the log posterior representation considered here, this would require exponentiating the spike counts, summing over the relevant neurons, and then taking a log.

Rather than assuming a constant prior and sum of tuning curves in the objective function of Eq. (2), a more general solution could embed the prior into the measurements by arranging that the sum of the tuning curves is equal to the log of the prior:

$$\sum_n f_n(x) = \log P(x)$$

(note, a precise form of this proposal would need to limit the the smallest probability that could be represented). Under these conditions, the last two terms cancel each other, and the full log-posterior may again be computed as a linear function (i.e., a weighted sum) of the spike counts. This is effectively a strategy for embedding the prior into the measurements,

and is more efficient than encoding priors with spiking responses of another set of neurons (Ma *et al.*, 2006). This solution would require that the brain adjust the tuning curves so as to sum to the log prior, which is essentially a resource allocation problem: neurons should be adjusted so as to properly "cover" the distribution of inputs, assigning more resources (i.e., a higher total spiking response, which corresponds to an expenditure of more metabolic energy) to inputs that occur more frequently. This adjustment could be achieved by changing either their response gains, the overlap between their tuning curves, the widths of their tuning curves, or some combination of these.

In the probabilistic representation described above, a single optimal estimate can be computed by appropriately combining information over the entire population. Implementing this in the brain would presumably require creation of a redundant population of neurons to linearly recode the implicit representation into an explicit one (that is, a population whose responses equal the posterior or log posterior), from which a maximum (or mean) could be selected. Although this sort of explicit estimation has been assumed by many of the previously mentioned publications, it seems to me wasteful of neural resources, and not robust to the additional noise that would be introduced by neurons computing and representing the estimate. It seems more likely that the brain leaves the representation probabilistic and implicit, performing further calculations in a way that is consistent with this (e.g. Ma *et al.*, 2006). Taking this *principle of delayed estimation* to its extreme, we could hypothesize that estimates need only be made explicit when the information reaches the motor system, and the animal most execute an action (e.g., reaching out to grasp an object). At that point, the estimate is "computed" by a bone, which responds by moving according to the collective activity of all the muscle fibers that are pulling on it! One case that may be an exception to this is that of a binary estimate (i.e., a decision), for which the log-posterior need only be computed at two different values (rather than a continuum). Experimental evidence suggests that the firing rates of neurons in parietal cortex may represent such values directly (e.g., Platt & Glimcher, 1999; Gold & Shadlen, 2000).

3 Conclusions

Optimal estimation provides a formal framework for investigating and interpreting perceptual capabilities. The definition of the optimal estimator does not specify a fixed universal function, but depends on four fundamental ingredients: the ensemble of input signals over which it is to be optimal, the (probabilistic) relationship between the signal and the measurement, the cost of mis-estimation, and the family of estimation functions from which the solution is to be chosen. In a biological system, these ingredients may change over time (especially the prior and loss function) and thus presumably need to be learned and continually updated based on recent input. I've reviewed three different formulations for developing an optimal estimator, each making different assumptions about the means by which these ingredients are obtained.

The basic formalism presented here is highly oversimplified. In particular, I've side-stepped

several important features of sensory systems that need to be incorporated in a full solution:

- Sensory computations occur in cascades of neural populations, and each of these presumably performs some transformations on the signals received from its afferents, and introduces additional noise. The designation of a particular population as the "measurement" is thus somewhat artificial.
- Sensory computations occur over time. Many optimal estimation problems may be rewritten in a form that can be computed incrementally (the classical solution in statistical signal processing is known as the "Kalman filter"), and such solutions have been used to model temporal aspects of neural processing (e.g., Rao & Ballard, 1997; Denève *et al.*, 2007)
- Many sensory inference computations depend on information beyond the measurements and the prior, such as inputs from multiple sensory modalities, feedback in the form of attentional signals, or from emotional or cognitive centers, etc. It might be possible to formalize these effects as a contextual form of prior (Jaynes, 2003).

In summary, the challenge for future research is to develop optimal estimation solutions that may be plausibly mapped onto brain architecture, that are flexible and adaptive, that may be cascaded (with additional noise introduced at each stage of the cascade), and that may be learned in a primarily unsupervised setting. The time seems ripe for this: A long tradition of rigorous study in statistical inference has been developed and refined into engineering tools. The use and development of these has been recently accelerated by new methods and algorithms arising in the machine learning community. Coupling these with a new generation of experimental measurement technologies (especially those for obtaining responses from groups or populations of neurons simultaneously) leads me to conclude that we're on the verge of fundamental advances in our understanding of sensory processing.

References

Anderson, C. & van Essen, D. (1994). Neurobiological computational systems. In *IEEE World Congress on Computational Intelligence*. IEEE Press, New York.

Barber, M. J., Clark, J., & Anderson, C. H. (2003). Neural representation of probabilistic information. *Neural Computation*, *15*(8), 1843–1864.

Barlow, H. B. (1980). The absolute efficiency of perceptual decisions. *Phil. Trans of the Royal Society, London B*, *290*, 71–82.

Beck, J., Ma, W. J., Kiani, R., Hanks, T., Churchland, A., Roitman, J., Shadlen, M., Latham, P., & Pouget, A. (2008). Probabilistic population codes for bayesian decision making. *Neuron*, *60*(6), 1142–1152.

Bialek, W., Rieke, F., de Ruyter van Steveninck, R. R., & Warland, D. (1991). Reading a neural code. *Science*, *252*, 1854–1857.

Bialek, W. & van Steveninck, R. d. (2005). Features and dimensions: Motion estimation in fly vision. Technical Report qbio/0505003, http://arxiv.org/.

Denève, S., Duhamel, J.-R., & Pouget, A. (2007). Optimal sensorimotor integration in recurrent cortical networks: A neural implementation of Kalman filters. *The Journal of Neuroscience*, *27*(21), 5744–5756.

Eliasmith, C. & Anderson, C. H. (2002). *Neural Engineering: Computation, Representation, and Dynamics in Neurobiological Systems*. MIT Press, Cambridge, MA.

Faisal, A. A., Selen, L. P. J., & Wolpert, D. M. (2008). Noise in the nervous system. *Nature Reviews Neuroscience*, *9*, 292–303.

Foldiak, P. (1993). The 'ideal homunculus': statistical inference from neural population responses. In Eeckmann, F. H. & Bower, J. M., editors, *Computation and Neural Systems*, chapter 9, pages 55–60. Kluwer Academic Publishers, Norwell, MA.

Geisler, W. S. (1989). Ideal-observer theory in psychophysics and physiology. *Physica Scripta*, *39*, 153–160.

Georgopoulos, A. P., Schwartz, A. B., & Kettner, R. E. (1986). Neuronal population coding of movement direction. *Nature*, *233*, 1416–1419.

Glimcher, P. W., Camerer, C., Poldrack, R., & Fehr, E., editors (2008). *Neuroeconomics: Decision Making and the Brain*. Academic Press, London.

Gold, J. I. & Shadlen, M. N. (2000). Representation of a perceptual decision in developing oculomotor commands. *Nature*, *404*, 390–394.

Green, D. M. & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. Wiley, New York.

Hinton, G. & Sejnowski, T. J., editors (1999). *Unsupervised Learning - Foundations of Neural Computation*. The MIT Press, Camridge, MA.

Hinton, G. E. (1992). How neural networks learn from experience. *Scientific American*.

Jaynes, E. T. (1957). How does the brain do plausible reasoning? Technical Report 421. Reprinted in: Maximum-Entropy and Bayesian Methods in Science and Engineering. G.J. Erickson and C.R. Smitt (eds), vol. 1, pp. 1-23, 1988, Kluwer Academic Publishers.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge, UK.

Jazayeri, M. & Movshon, J. A. (2006). Optimal representation of sensory information by neural populationse. *Nature Neuroscience*, *9*(5).

Kersten, D. (1990). Statistical limits to image understanding. In Blakemore, C., editor, *Vision: Coding and efficiency*, pages 32–44. Cambridge University Press.

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as bayesian inference. *Annual Review of Psychology*, pages 271–304.

Knill, D. & Richards, W. (1996). *Perception as Bayesian Inference*. Cambridge University Press.

Knill, D. C., Field, D., & Kersten, D. (1990). Human discrimination of fractal images. *J. Opt. Soc. Am. A*, *7*, 1113–1123.

Körding, K. P. (2007). Decision theory: What "should" the nervous system do? *Science*, *318*(5850), 606–610.

Körding, K. P. & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*(6971), 244–247.

Ma, W. J., Beck, J. M., Latham, P. E., & Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, *9*, 1432–1438.

Maloney, L. T. (2002). Statistical decision theory and biological vision. In Heyer, D. & Mausfeld, R., editors, *Perception and the Physical World: Psychological and Philosophical Issues in Perception*, chapter 6, pages 145–189. Wiley, New York.

Maloney, L. T. & Mamassian, P. (2009). Bayesian decision theory as a model of visual perception: Testing Bayesian transfer. *Visual Neuroscience*, *26*(1), 147–155.

Mamassian, P., Landy, M. S., & Maloney, L. T. (2002). Bayesian modeling of visual perception. In Rao, R., Lewicki, M. ., & Olshausen, B., editors, *Probabilistic Models of the Brain: Perception and Neural Function*, chapter 1, pages 13–36. MIT Press, Cambridge, MA.

Miyasawa, K. (1956). An empirical Bayes estimator of the mean of a normal population. *Bull. Inst. Internat. Statist.*, *38*, 181–188.

Paninski, L. (2006). Nonparametric inference of prior probabilities from Bayes-optimal behavior. In Weiss, Y., Schölkopf, B., & Platt, J., editors, *Adv. in Neural Information Processing Systems*, volume 18, pages 1067–1074. MIT Press, Cambridge, MA.

Pelli, D. G. & Farell, B. (1999). Why use noise? *J. Optical Societyof America A*, *16*, 647–653.

Platt, M. L. & Glimcher, P. W. (1999). Neural correlates of decision variables in parietal cortex. *Nature*, *400*, 233–238.

Potters, M. & W.Bialek (1994). Statistical mechanics and visual signal processing. *J. Physics I France*, *4*, 1755–1775.

Pouget, A., Dayan, P., & Zemel, R. S. (2003). Inference and computation with population codes. *Ann. Rev. Neuroscience*, *26*, 381–410.

Rao, R. P. N. & Ballard, D. H. (1997). Dynamic model of visual recognition predicts neural response properties in the visual cortex. *Neural Computation*, *9*, 721–763.

Raphan, M. & Simoncelli, E. P. (2007). Learning to be Bayesian without supervision. In Schölkopf, B., Platt, J., & Hofmann, T., editors, *Adv. Neural Information Processing Systems 19*, volume 19, pages 1145–1152, Cambridge, MA. MIT Press.

Rieke, F. & Baylor, D. (1998). Single photon detection by rod cells of the retina. *Rev. Modern Physics*, *70*, 1027–1036.

Rieke, F., Warland, D., de Ruyter van Steveninck, R. R., & Bialek, W. (1997). *Spikes: Exploring the Neural Code*. MIT Press, Cambridge, MA.

Sahani, M. & Dayan, P. (2003). Doubly distributional population codes: Simultaneous representation of uncertainty and multiplicity. *Neural Computation*, *15*, 2255–2279.

Salinas, E. & Abbott, L. F. (1994). Vector reconstruction from firing rates. *J. Computational Neuroscience*, *1*, 89–107.

Sanger, T. D. (1996). Probability density estimation for the interpretation of neural population codes. *J. Neurophysiology*, *76*(4), 2790–2793.

Seung, H. S. & Sompolinsky, H. (1993). Simple models for reading neural population codes. *Proc. National Academy of Sciences*, *90*, 10749–10753.

Shadlen, M. N., Britten, K. H., Newsome, W. T., & Movshon, J. A. (1996). A computational analysis of the relationship between neuronal and behavioral responses to visual motion. *jneurosci*, *16*(4), 1486–1510.

Simoncelli, E. P. (1993). *Distributed Analysis and Representation of Visual Motion*. PhD thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA. Also available as MIT Media Laboratory Vision and Modeling Technical Report #209.

Simoncelli, E. P. (2003). Local analysis of visual motion. In Chalupa, L. M. & Werner, J. S., editors, *The Visual Neurosciences*, chapter 109, pages 1616–1623. MIT Press.

Snippe, H. P. (1996). Parameter extraction from population codes: A critical assessment. *Neural Computation*, *8*(3), 511–530.

Stein, C. M. (1981). Estimation of the mean of a multivariate Normal distribution. *Annals of Statistics*, *9*(6), 1135–1151.

Stocker, A. A. & Simoncelli, E. P. (2006a). Noise characteristics and prior expectations in human visual speed perception. *Nature Neuroscience*, *9*(4), 578–585.

Stocker, A. A. & Simoncelli, E. P. (2006b). Sensory adaptation within a Bayesian framework for perception. In Weiss, Y., Schölkopf, B., & Platt, J., editors, *Adv. Neural Information Processing Systems (NIPS*05)*, volume 18, pages 1291–1298, Cambridge, MA. MIT Press.

von Helmholtz, H. (1925). *Treatise on Physiological Optics*, volume III. Optical Society of America, New York.

Weiss, Y. & Fleet, D. J. (2002). Velocity likelihoods in biological and machine vision. In Rao, R., Olshausen, B., & Lewicki, M., editors, *Probabilistic Models of the Brain: Perception and Neural Function*, pages 81–100. MIT Press.

Zemel, R. S., Dayan, P., & Pouget, A. (1998). Probabilistic interpretation of population codes. *Neural Computation*, *10*, 403–430.

Zhang, K., Ginzburg, I., McNaughton, B., & Sejnowski, T. J. (1998). Interpreting neuronal population activity by reconstruction: Unified framework with application to hippocampal place cells. *J. Neurophysiology*, *79*, 1017–1044.