Combining Probabilistic Population Codes

Richard S. Zemel
University of Arizona
Tucson, AZ 85721 USA
zemelOu.arizona.edu

Peter Dayan
MIT
Cambridge, MA 02139 USA
dayan@psyche.mit.edu

Abstract

We study the problem of statistically correct inference in networks whose basic representations are population codes. Population codes are ubiquitous in the brain, and involve the simultaneous activity of many units coding for some low dimensional quantity. A classic example are place cells in the rat hippocampus: these fire when the animal is at a particular place in an environment, so the underlying quantity has two dimensions of spatial location. We show how to interpret the activity as encoding whole probability distributions over the underlying variable rather then just single values, and propose a method of inductively learning mappings between population codes that are computationally tractable and yet offer good approximations to statistically optimal inference. We simulate the method on some simple examples to prove its competence.

In a population code, information about some low-dimensional quantity (such as the position of a visual feature) is represented in the activity of a collection of units, each responding to a limited range of stimuli within this low-dimensional space. Strong evidence exists for this form of coding at the sensory input areas of the brain (*eg* retinotopic and tonotopic maps) as well as at the motor output level [Georgopoulos *et* a/., 1986]. Evidence is mounting that many other intermediate neural processing areas also use population codes [Tanaka, 1996]. Certain important questions about population codes have been extensively investigated, including how to extract an optimal underlying value [Salinas and Abbott, 1994; Snippe, 1996] and how to learn such representations [Kohonen, 1982]. However, two important issues have been almost ignored (with the important exception of [Anderson, 1994]). One is the treatment of population codes as encoding whole probability density functions (PDFs) over the underlying quantities rather than just a single value. PDFs can convey significant additional information, such as certainty *(eg* in the existence in an image of the relevant object), as well as the mean and variance *(eg* in its position). The other issue is how to perform inference in networks whose basic representations are population codes.

Zemel, Dayan, and Pouget [1997] have recently presented a general framework for the probabilistic interpretation of population codes in terms of PDFs. In this paper we apply this framework to all the population codes in a processing hierarchy, and suggest an inference method that approximates, in a quantifiable manner, Bayesian optimal methods of representing and combining the probability distributions.

We first discuss how to interpret PDFs from population codes, and then introduce our framework for combining these codes. We illustrate the techniques with an example based on a form of cue combination.

1 An Example

Consider the case of a hunter attempting to shoot a pheasant as it flies out of a tree. We'll assume that the hunter uses two cues, a visual cue concerning motion in the tree and an auditory cue based on rustling of the leaves, to estimate the pheasant's size and velocity. Based on this estimate, he selects a time and place to fire his shotgun.

The *combination* problem concerns how the two inputs should be combined to produce the output. In the simplest version of the combination problem for this example, visual motion is confined to one part of the tree, and the auditory signal directly corresponds to this visual signal. Here these two single-valued inputs (which we will term v and a) give rise to a single output, and the hunter confidently aims his shotgun (to location s).

Evidence exists that the two inputs and the output information in this example are each represented in neural population codes in some animals. That is, a fixed collection of neurons fire for each of the three variables of interest. The relevant visual input is represented by the activity of a population of motion detectors: in monkeys, a particular cortical area (MT) contains cells that selectively respond to motion of a particular velocity within a small range of visual locations. Similarly, the relevant auditory input is represented in a population of detectors tuned to particular frequencies and spatial locations in owl auditory cortex [Knudsen and Konishi, 1978]; the frequency may contain important information about the bird's size and speed. Directional motor output is also represented in a population code in monkey motor cortex [Georgopoulos *et al*., 1986].

Therefore even in the simple version of the problem, the brain does not directly represent the values v, a, and s, but instead represents each in a separate population code. The most straightforward way to solve this problem is to perform an intermediate step of extracting separate single values from the input population codes, combine these values, and then encode these into the motor output population code. However, this seems not to be the strategy actually implemented in the brain, where new population codes appear to be generated directly from old ones.

Another level of complexity is introduced into the problem when we consider that the inputs may be uncertain or ambiguous. For example, if the wind is blowing, then leaves may be moving all over the tree giving rise to multiple plausible motion hypotheses, while at the same time the auditory cues may be too faint to confidently estimate the motion. The experienced hunter may then be able to narrow down the set of candidate motions based on his knowledge of the combinations of auditory and visual cues, but he might not be able to confidently select a single value. Two additional problems are introduced in this more general case. First we must interpret a population code as representing a whole probability distribution over the underlying variable. And then the combination method must preserve the probabilistic information in the inputs. Thus the aim of a combination network is to infer a population code for the motor action that preserves the statistical relationship between the input and output probability distributions.

2 Theory

The basic theory underlying the combination of population codes is extremely simple. Population codes use the *explicit* activities $\mathbf{r} = \{r_i\}$ of multiple cells (as in area MT) to code information about the value of an *implicit* underlying variable x (such as the direction and speed of motion of the leaves). We are interested in the case that the activities r code a whole probability distribution over the underlying variable:

$$\mathcal{P}[\mathbf{x}|\mathbf{r}]. \qquad \text{(i)}$$

Consider the example of the hunter. Activities $\{r_i^v\}$ and $\{r_j^a\}$ represent probability distributions over the motion position and velocity based on the visual and auditory signals respectively. We will assume that afferent information in the different modalities is independent. The activities $\{r_k^s\}$ will represent a probability distribution over the corresponding required position $\mathbf{s}$ of the shotgun according to the equivalent of Equation 1.

Two computational operations are required to produce appropriate $\mathbf{r}^s$: information from the different modalities must be integrated and then expressed in appropriate coordinates. These operations have to respect the statistical semantics of $\mathbf{r}^v$ and $\mathbf{r}^a$. We use an underlying analysis-by-synthesis statistical model as in the Helmholtz machine [Hinton *et al.*, 1995]. In such a model, inference is based on the analysis or recognition inverse to a probabilistic synthetic or generative model that specifies probability distributions $\mathcal{P}[\mathbf{v},\mathbf{a}|\mathbf{s}]$ over the visual motion signal $\mathbf{v}$ and auditory pattern $\mathbf{a}$ given the shotgun location $\mathbf{s}$.

Given true probability distributions $\mathcal{P}[\mathbf{v}|\omega]$ and $\mathcal{P}[\mathbf{a}|\omega]$ over the visual and auditory information (here ω represents the underlying information available to the hunter), recognition requires calculating:

$$\mathcal{P}[\mathbf{s}|\omega] = \int_{\mathbf{v},\mathbf{a}} \mathcal{P}[\mathbf{v}|\omega]\mathcal{P}[\mathbf{a}|\omega]\mathcal{P}[\mathbf{s}|\mathbf{v},\mathbf{a}]d\mathbf{v}d\mathbf{a} \qquad (2)$$

$$\propto \mathcal{P}[\mathbf{s}] \int_{\mathbf{v},\mathbf{a}} \mathcal{P}[\mathbf{v}|\omega]\mathcal{P}[\mathbf{a}|\omega]\mathcal{P}[\mathbf{v},\mathbf{a}|\mathbf{s}]d\mathbf{v}d\mathbf{a} \qquad (3)$$

where $\mathcal{P}[\mathbf{s}]$ is the prior distribution over $\mathbf{s}$. Equation 3 establishes the standard by which inferences about the distribution over $\mathbf{s}$ should be judged.

We have therefore reduced the computational problem to one of mapping activities $\mathbf{r}^v$ and $\mathbf{r}^a$ into activities $\mathbf{r}^s$ for which $\mathcal{P}[\mathbf{s}|\mathbf{r}^s]$ from Equation 1 is a good approximation to the integration in Equation 3, where $\mathcal{P}[\mathbf{v}|\omega]$ is what $\mathbf{r}^v$ represents (according to Equation 1) and $\mathcal{P}[\mathbf{a}|\omega]$ is what $\mathbf{r}^a$ represents. Figure 1 illustrates the generative and recognition operations, showing the activities, the distributions that they represent, and the various probabilistic relationships.

The remaining questions concern how activities $\mathbf{r}$ specify distributions as in Equation 1, and how $\mathbf{r}^v$ and $\mathbf{r}^a$ are actually combined to produce $\mathbf{r}^s$. We describe two models for Equation 1: a model based on a standard form of function approximation, kernel density estimation (KDE) [Anderson, 1994], and an extension to the conventional statistical population coding model that is designed to handle any form of PDF [Zemel *et al.*, 1997]. Both models form estimates $\hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x})$ of $\mathcal{P}[\mathbf{x}|\omega]$ based on $\mathbf{r}$.

2.1 The KDE Model

One way of treating population codes as distributions is in terms of kernel density estimates (KDEs) [Anderson, 1994]. Here, activities $\mathbf{r}$ represent distribution $\hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x})$

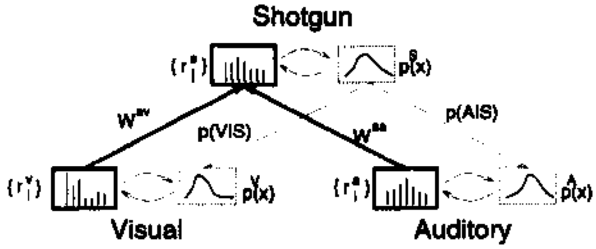

Figure 1: An example of a network formulated to combine population code representations of PDFs. The boldface elements depict explicit network components.

through a linear combination of basis functions $\psi_i(\mathbf{x})$, *ie* $\hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x}) = \sum_i r_i'\psi_i(\mathbf{x})$ where the $\{r_i'\}$ are normalized such that $\hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x})$ is a probability distribution. The kernel functions $\psi_i(\mathbf{x})$ are *not* what are called tuning functions of the cells (the tuning function $f_i(\mathbf{x})$ describes the relationship between the cell's firing rate r_i and the underlying variable $\mathbf{x}$, where $f_i(\mathbf{x})$ could be Gaussian about some preferred value $\mathbf{x}_i$). The kernel functions need have *no* neural instantiation; instead, they form part of the interpretive structure for the population code. If the $\psi_i(\mathbf{x})$ are probability distributions, and so are positive, then the range of spatial frequencies in $\mathcal{P}[\mathbf{x}|\omega]$ that they can reproduce in $\hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x})$ is likely to be severely limited.

Evaluating the KDE model requires some method of representing $\mathcal{P}[\mathbf{x}|\omega]$ in a firing rate vector $\mathbf{r}$, which we term *encoding*. One way to encode is to use the Kullback-Leibler divergence as a measure of the discrepancy between $\mathcal{P}[\mathbf{x}|\omega]$ and $\sum_i r_i'\psi_i(\mathbf{x})$ and use the expectation-maximization (EM) algorithm to fit the $\{r_i'\}$, treating them as mixing proportions in a mixture model [Dempster *et al.*, 1977]. This relies on $\{\psi_i(\mathbf{x})\}$ being probability distributions themselves. The *projection method* [Anderson, 1994] is a one-shot linear filtering based alternative using the $\mathcal{L}_2$ distance. Here the r_i are computed as a projection of $\mathcal{P}[\mathbf{x}|\omega]$ onto tuning functions $f_i(\mathbf{x})$ that are calculated from $\psi_j(\mathbf{x})$:

$$r_i = \int_{\mathbf{x}} \mathcal{P}[\mathbf{x}|\omega] f_i(\mathbf{x}) d\mathbf{x} \tag{4}$$

$$f_i(\mathbf{x}) = \sum_j A_{ij}^{-1}\psi_j(\mathbf{x}) \quad A_{ij} = \int_{\mathbf{x}} \psi_i(\mathbf{x})\psi_j(\mathbf{x}) d\mathbf{x}$$

An extremely attractive property of the KDE model is that it makes combination of population codes very simple [Anderson, 1994]. If $\hat{\mathcal{P}}^{\mathbf{r}^v}(\mathbf{v}) = \sum_i r_i^v\psi_i(\mathbf{v})$ is an adequate model of $\mathcal{P}[\mathbf{v}|\omega]$ and $\hat{\mathcal{P}}^{\mathbf{r}^a}(\mathbf{a}) = \sum_j r_j^a\psi_j(\mathbf{a})$ is an adequate model of $\mathcal{P}[\mathbf{a}|\omega]$, then

$$\begin{aligned} &\int_{\mathbf{v},\mathbf{a}} \mathcal{P}[\mathbf{v}|\omega]\mathcal{P}[\mathbf{a}|\omega]\mathcal{P}[\mathbf{v},\mathbf{a}|\mathbf{s}]\mathcal{P}[\mathbf{s}]d\mathbf{v}d\mathbf{a} \\ &\quad\sim \int_{\mathbf{v},\mathbf{a}} \hat{\mathcal{P}}^{\mathbf{r}^v}(\mathbf{v})\hat{\mathcal{P}}^{\mathbf{r}^a}(\mathbf{a})\mathcal{P}[\mathbf{v},\mathbf{a}|\mathbf{s}]\mathcal{P}[\mathbf{s}]d\mathbf{v}d\mathbf{a} \\ &\quad= \sum_{ij} r_i^v r_j^a w_{ij}(\mathbf{s}) \end{aligned} \tag{5}$$

where

$$w_{ij}(\mathbf{s}) = \int_{\mathbf{v},\mathbf{a}} \psi_i(\mathbf{v})\psi_j(\mathbf{a})\mathcal{P}[\mathbf{v},\mathbf{a}|\mathbf{s}]\mathcal{P}[\mathbf{s}]d\mathbf{v}d\mathbf{a}.$$

This makes Equation 3 into a simple bilinear sum, even when $\mathbf{v}$ and $\mathbf{a}$ are not independent given $\mathbf{s}$.

Probabilistically correct combination is not so simple in this model if the underlying multiplication in Equation 5 is not allowed. In the combination network described below, we restricted the combination to the standard mechanism of linear summation followed by a nonlinear activation function.

2.2 The Extended Poisson Model

Unfortunately, the KDE model has significant difficulties representing probability distributions that involve higher spatial frequencies than the kernel functions. This is a natural, but significant limitation of this method, since the situation of nearly complete certainty in $\mathbf{v}$ or $\mathbf{a}$ or $\mathbf{s}$ is particularly important. An alternative method has been suggested that is based on the same probabilistic analysis that underlies most standard applications of population coding [Zemel *et al.*, 1997].

For these applications the starting point is usually the neurophysiological finding, largely ignored by the KDE model, that the relevant cells have unimodal tuning functions $f_i(\mathbf{x})$. Standard accounts also assume that firing rates vary, even for a fixed input, *eg* the Poisson model [Seung and Sompolinsky, 1993] for which

$$\mathcal{P}[r_i|\mathbf{x}] = e^{-f_i(\mathbf{x})}(f_i(\mathbf{x}))^{r_i}/r_i! \tag{6}$$

This is an *encoding* model, since it relates how $\mathbf{x}$ is coded in $\mathbf{r}$. The form of Equation 1 is then specified (in an operation referred to as *decoding*) as the statistical inverse to this encoding model $\hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x}) \equiv \mathcal{P}[\mathbf{x}|\{r_i\}] \propto \mathcal{P}[\mathbf{x}]\prod_i \mathcal{P}[r_i|\mathbf{x}]$, where $\mathcal{P}[\mathbf{x}]$ is the prior probability distribution over $\mathbf{x}$. In these terms, the KDE model is specified by its method of decoding—Equation 4 automatically follows as the method of encoding. Note that a single value of $\mathbf{x}$ gives rise to an estimate $\hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x})$ that is a distribution over $\mathbf{x}$ due to the assumed variability.

Whereas the KDE model has problems representing peaked distributions, the conventional Poisson model has problems representing broad distributions, since Equation 6 assumes that there is a single underlying value of $\mathbf{x}$. If the information provided to the hunter is somewhat unspecific with respect to the visual information, then it does not make sense to assume that $\mathbf{r}^v$ are all determined (stochastically) according to some single particular value. In the extended Poisson model, the recorded activities $\mathbf{r}$ are allowed to depend on general $\mathcal{P}[\mathbf{x}|\omega]$, having Poisson statistics with mean:

$$\langle r_i \rangle = \int_{\mathbf{x}} \mathcal{P}[\mathbf{x}|\omega] f_i(\mathbf{x}) d\mathbf{x}. \tag{7}$$

This equation is identical to that for the KDE model (Equation 4), except that variability is built into the Poisson statistics, and decoding is now required to be the Bayesian inverse of encoding. Note that since r_i depends stochastically on $\mathcal{P}[\mathbf{x}|\omega]$, the full Bayesian inverse will specify a distribution $\mathcal{P}[\mathcal{P}[\mathbf{x}|\omega]|\mathbf{r}]$ over possible distributions. The distribution in Equation 1 can then be generated as the maximum likelihood (or rather maximum *a posteriori* with respect to a smoothness prior) distribution that $\mathcal{P}[\mathcal{P}[\mathbf{x}|\omega]|\mathbf{r}]$ implies.

Decoding in this model may be performed by approximating $\mathcal{P}[\mathbf{x}|\omega]$ as a piece-wise constant histogram which takes the value ϕ_j in $(\mathbf{x}_j, \mathbf{x}_{j+1}]$, and $f_i(\mathbf{x})$ by a piece-wise constant histogram that take the values f_{ij} in $(\mathbf{x}_j, \mathbf{x}_{j+1}]$. The maximum *a posteriori* estimate for $\{\phi_j\}$ can be derived by maximizing:

$$L(\{\hat{\phi}_j\}) = \sum_i r_i \log \left[\sum_j \hat{\phi}_j f_{ij} \right] - \epsilon \sum_j \left(\hat{\phi}_j - \hat{\phi}_{j+1} \right)^2$$

where ϵ is the variance of an assumed smoothness prior. A form of EM may be used to maximize the likelihood $L(\{\hat{\phi}_j\})$. By comparison with the linear decoding of the KDE method, the extended Poisson model offers a *non-linear* way of combining a set of activities $\{r_i\}$ to give a probability distribution $\hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x})$ over the underlying variable $\mathbf{x}$. Figure 2 describes the full extended Poisson model and illustrates the underlying probabilistic framework for population codes.

Unfortunately, there are no short cuts like Equation 5 for combining these extended Poisson population codes. Instead, we have adopted some particular functional form for the combination and optimized its parameters in order to satisfy Equation 3 as best as possible.

3 Experiments

The central question for both coding methods is whether the combination network can form a set of activities $\{r_k^s\}$ that makes $\hat{\mathcal{P}}^{\mathbf{r}^s}(\mathbf{s})$ a close approximation to the implicit distribution $\mathcal{P}[\mathbf{s}|\omega]$. For the purpose of evaluating network performance, and adapting its weights, several aspects need to be determined: 1). the implicit distribution $\mathcal{P}[\mathbf{s}|\omega]$. In the hunting example, this distribution describes the likely shotgun motions based on all information available to the hunter, so multiple peaks correspond to different possible motions and entropy corresponds to uncertainty about these motions. 2). the generative model $\mathcal{P}[\mathbf{v}, \mathbf{a}|\mathbf{s}]$. The implicit distributions for the network inputs are produced by applying $\mathcal{P}[\mathbf{v}, \mathbf{a}|\mathbf{s}]$ to $\mathcal{P}[\mathbf{s}|\omega]$. In these simulations, we made the simplifying assumption that the visual and auditory signals are independent given s. 3). the encoding model. The inputs $\mathbf{r}^v$ and $\mathbf{r}^a$ are obtained from the input implicit distributions via appropriate encoding model (Equation 4 for KDE; Equation 7 for the extended Poisson method). 4). a *combination function.* The network inputs produce an output $\mathbf{r}^s$ based on a weighted combination of $\mathbf{r}^v$ and $\mathbf{r}^a$. In these simulations we had both excitatory W and inhibitory weights U between each input and output unit, and the combination function was:

$$r_k^s = \frac{b_k + \sum_i r_i^v W_{ki}^{sv} + \sum_j r_j^a W_{kj}^{sa}}{c_k + \sum_i r_i^v U_{ki}^{sv} + \sum_j r_j^a U_{kj}^{sa}} \tag{8}$$

Note that this is not quite general enough to implement Equation 5 exactly.

We evaluate the networks' performances by comparing the $\hat{\mathcal{P}}^{\mathbf{r}^s}$ (s) obtained by decoding the explicit representation s in the network to the true implicit distribution

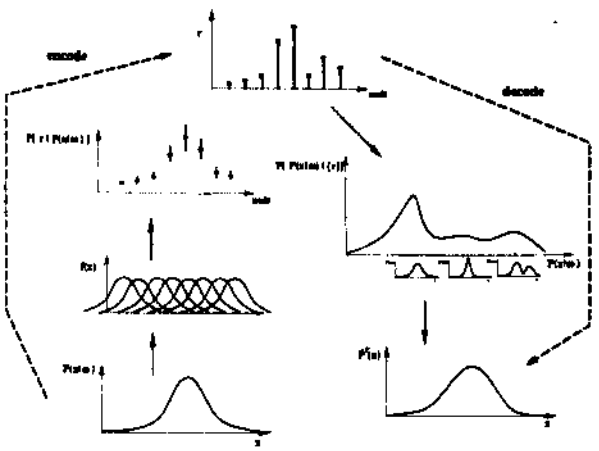

Figure 2: The encoding-decoding framework in the extended Poisson model. *Left:* Activities r may be interpreted as encoding a PDF in implicit space. *Top:* The output of the encoding process is the explicit activities of units, assumed to have been generated by the independent application of each cell's tuning function and additive noise to the implicit representation *{Bottom:* an implicit distribution $\mathcal{P}[\mathbf{x}|\omega]$). *Right:* Decoding the rates into a distribution $\hat{\mathcal{P}}^{\mathbf{r}}(\mathbf{x})$ involves an approximate form of maximum likelihood in distributions over x.

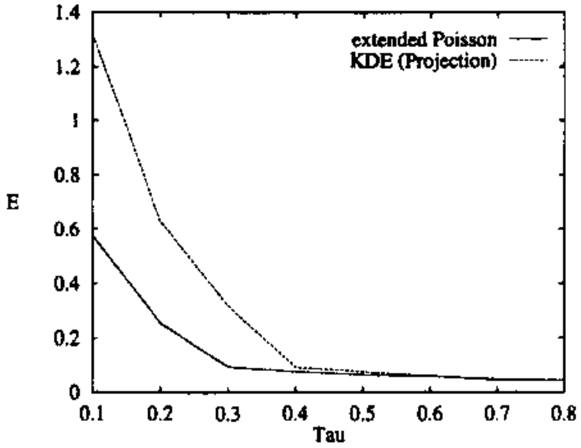

Figure 3: Combination network performance for a bimodal target distribution of width τ. The error measure is E from Equation 9.

$\mathcal{P}[s|\omega]$. The error function is the Kullback-Leibler divergence:

$$E = \int_s \mathcal{P}[s|\omega] \log \frac{\mathcal{P}[s|\omega]}{\hat{\mathcal{P}}^{r'}(s)} ds \qquad (9)$$

The network weights (excitatory and inhibitory) are adapted to minimize this error function using the delta-rule.

We examined the combination network performance under several target distributions and generative models. Here we focus on a difficult, important target distribution, a bimodal Gaussian distribution $1/2\mathcal{N}(2,\tau) + 1/2\mathcal{N}(-2,\tau)$ that contains both uncertainty in a given position (τ) and multiple possible values ($\mathbf{x} = \{2,-2\}$). The generative model is simple: $\mathcal{P}[\mathbf{v}|\mathbf{s}]$ and $\mathcal{P}[\mathbf{a}|\mathbf{s}]$ are Gaussians with variance 0.5 and means at 1.0 and -1.0 respectively.

We compared the two methods of probabilistic interpretation on this problem. For the kernel functions (KDE) and the tuning functions (extended Poisson model), we used Gaussians $\mathcal{N}(\mathbf{x}_i, 0.3)$. To reduce the number of units in the network, we considered a simplified situation where all signals are constrained to one dimension. This reduction is only a matter of convenience; the network can readily be extended to include higher-dimensional implicit spaces. We used 50 units in each population code, and the $\mathbf{x}_i$ were spaced evenly in the range $\mathbf{x} = [-10, 10]$.

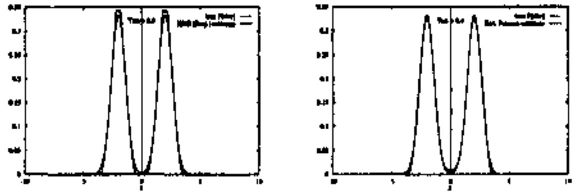

Figure 4: Examples of a combination network's reconstruction of a bimodal target distribution $\mathcal{P}[s|\omega]$ of width $\tau = 0.6$. Results using the KDE coding method is shown on the left, extended Poisson on the right.

Results are shown in Figures 3 and 4. At low values of τ, the target distribution approximates two spikes, at $\mathbf{x} = 2$ and $\mathbf{x} = -2$. The KDE method is not as accurate here, since it is unable to retain the high frequency information required for precise recovery of the these two location. As the target distribution contains more uncertainty, both methods are able to recover the implicit distribution with high fidelity. Note that an error rate of 0.7 bits for this target distribution would be obtained if $\hat{\mathcal{P}}^{r'}$ (s) has the correct peaks and is off by a factor of 2 in r (see Equation 9).

We have also conducted a number of other experiments with this combination method. In one set of experiments, we modeled the task of combining monocular and stereo cues to estimate depth in a particular visual illusion. In the double-nail illusion, the task is to estimate the depth of a nail aligned directly behind another nail in the observer's line of sight. Here computational vision systems based on binocular stereo produce a PDF for depth estimates with two peaks, one at the correct value and another at the illusory frontoparallel interpretation (both nails side-by-side). A PDF based on monocular cues will not have the same ambiguity, but it is typically a much broader distribution [Yuille and Biilthoff, 1994]. These two PDFs must be combined multiplicatively to produce the correct peak.

We simulated this problem by training a combination network identical to the network described above except in the generative model. Here P[b|t] is a multimodal Gaussian l/3N[t, 1/2] + 2/3N[t + 2,1/2] (with a frontoparallel bias) and P[m|t] is a broader unimodal Gaussian N[t, 1], where b, m and t are the binocular, monocular and true depth estimates, respectively. After training on 300 cases in which the target distribution was a narrow Gaussian N[t, .01], the network produced output distributions on novel inputs that were within .1 bits of the true distributions.

Other experiments have examined the combination network's ability to recover PDFs in which the certainty as to the presence of the output (ie the integral under the PDF) is < 1. Good performance on this task suggests that the method can be useful for recognition *(eg* recognizing an instance of an object based on the spatial locations of its features).

4 Discussion

We have presented a general framework for mapping between population codes that approximates statistically correct inference. The framework applies and extends two recent methods for the probabilistic interpretation of population codes to the problem of combining these codes. This framework has a wide variety of applications, including any context in which probabilistic information from several sources, each represented in a distributed manner, must be combined. The simulation results demonstrate that a feedforward network can capture the appropriate probabilistic relationships between some simple population-coded PDFs. Generally, several population-coded inputs should be multiplied (to compute a full joint PDF), but we found empirically that they can be combined reasonably using a non-linearity.

A straightforward alternative to the proposed framework would extract single values from the input population codes, combine these values, and then form a new population code at the output. Aside from biological realism, the computational advantage of constructing direct mappings between population codes without requiring an intermediate step of extracting single values is that information about whole distributions can be brought to bear—including the ambiguity and uncertainty in the underlying variables.

Integral to the framework is an interpretation of a population code as encoding a probability distribution over the underlying quantity. The framework can thus be seen as a generalization of [Salinas and Abbott, 1995], in which a network is trained to map one population code to another, where each code is interpreted as representing a single value. Our method extends this mapping to probabilistic interpretations while maintaining the biologically realistic representations.

There are many open issues, particularly understanding the nature of encoding and decoding. Both operations are only implicit in the system so some freedom exists in choosing ones appropriate for particular tasks. Based on neurobiological and engineering considerations, one expects a consistent interpretation across levels; maintaining this interpretation should lead to a simple learning rule. Noise is a second key issue. If constructing one population code from others introduces substantial extra noise, the system will be unable to convey information accurately. Here the restriction of the network to feedforward connections might be relaxed in order to allow lateral connections between units within a population, which may be useful in cleaning up the codes.

References

[Anderson, 1994] C. H. Anderson. Basic elements of biological computational systems. *International Journal of Modern Physics C,* 5(2):135-137, 1994.

[Dempster *et al,* 1977] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B,* 39:1-38, 1977.

[Georgopoulos *et al,* 1986] A. P. Georgopoulos, A. B. Schwartz, and R. E. Kettner. Neuronal population coding of movement direction. *Science,* 243:1416-1419, September 1986.

[Hinton *et ai,* 1995] G. E. Hinton, P. Dayan, B. J. Frey, and R. M. Neal. The wake-sleep algorithm for unsupervised neural networks. *Science,* 268(5214): 1158—1161, 1995.

[Knudsen and Konishi, 1978] E. I. Knudsen and M. Konishi. A neural map of auditory space in the owl. *Science,* 200:795-797, 1978.

[Kohonen, 1982] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics,* 43:59-69, 1982.

[Salinas and Abbott, 1994] E. Salinas and L. F. Abbott. Vector reconstruction from firing rates. *Journal of Computational Neuroscience,* 1:89-107, 1994.

[Salinas and Abbott, 1995] E. Salinas and L. F. Abbott. Transfer of coded information from sensory to motor networks. *Journal of Neuroscience,* 15(10):6461-6474, 1995.

[Seung and Sompolinsky, 1993] H. S. Seung and H. Sompolinsky. Simple models for reading neuronal population codes. *Proceedings of the National Academy of Sciences, USA,* 90:10749-10753,1993.

[Snippe, 1996] H. P. Snippe. Parameter extraction from population codes: a critical assessment. *Neural Computation,* 8(3):511-530, 1996.

[Tanaka, 1996] K. Tanaka. Inferotemporal cortex and object vision. *Annual Review of Neuroscience,* 19:109—139, 1996.

[Yuille and Bulthoff, 1994] A. L. Yuille and H. H. Bulthoff. Bayesian decision theory and psychophysics. In *Perception as Bayesian Inference.* Cambridge University Press, 1994.

[Zemel *et al,* 1997] R. S. Zemel, P. Dayan, and A. Pouget. Probabilistic interpretation of population codes. In M. C. Mozer, M. I. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems 9,* Cambridge, MA, 1997. MIT Press.