Brain and Cognition 112 (2017) 92–97

Contents lists available at ScienceDirect

Brain and Cognition

journal homepage: www.elsevier.com/locate/b&c

Special Invited Review

A review of predictive coding algorithms

M.W. Spratling *

King's College London, Department of Informatics, London, UK

ARTICLE INFO

Article history:
Received 21 May 2015
Revised 9 November 2015
Accepted 13 November 2015
Available online 19 January 2016

Keywords:
Predictive coding
Signal processing
Retina
Cortex
Free energy
Neural networks

ABSTRACT

Predictive coding is a leading theory of how the brain performs probabilistic inference. However, there are a number of distinct algorithms which are described by the term "predictive coding". This article provides a concise review of these different predictive coding algorithms, highlighting their similarities and differences. Five algorithms are covered: linear predictive coding which has a long and influential history in the signal processing literature; the first neuroscience-related application of predictive coding to explaining the function of the retina; and three versions of predictive coding that have been proposed to model cortical function. While all these algorithms aim to fit a generative model to sensory data, they differ in the type of generative model they employ, in the process used to optimise the fit between the model and sensory data, and in the way that they are related to neurobiology.

Contents

1. Introduction

To correctly interpret sensory data the brain is faced with solving an inverse problem: one where the causes need to be inferred from the perceived outcomes. For example, during visual perception the brain has access to information, measured by the eyes, about the spatial distribution of the intensity and wavelength of the incident light. From this information the brain needs to infer the arrangement of objects (the causes) that gave rise to the perceived image (the outcome of the image formation process). Inverse problems are typically ill-posed, meaning that they have multiple solutions (or none at all). For example, different sets of objects arranged in different configurations and viewed under different lighting conditions could potentially give rise to the same image. Solving such an ill-posed problem requires additional constraints to be imposed in order to narrow down the number of possible solutions to the single, most likely, one. In other words, constraints are required to infer the most likely causes of the sensory data. Constraints on perceptual inference might come from many sources, including knowledge learnt from prior experience (such as typical lighting conditions, the shapes and sizes of common objects, *etc.*), the recent past (knowledge about recently perceived causes, and expectations about how these might change or stay the same), and the present (such as information from elsewhere in the image or from another sensory modality).

Predictive coding suggests one way in which the brain might apply constraints in order to solve the inverse problem of perception (Bubic, von Cramon, & Schubotz, 2010; Clark, 2013; Huang & Rao, 2011; Rao & Ballard, 1999; Spratling, 2014a). Specifically, predictive coding suggests that the brain is equipped with an internal

* Address: Department of Informatics, King's College London, Strand, London WC2R 2LS, UK.
E-mail address: michael.spratling@kcl.ac.uk

http://dx.doi.org/10.1016/j.bandc.2015.11.003
0278-2626/

model of the world, or multiple models of specific aspects of the world embedded in different brain regions. This internal model encodes possible causes of sensory inputs as parameters of a generative model. New sensory inputs are then represented in terms of these known causes. Determining which combination of the many possible causes best fits the current sensory data is achieved through a process of minimising the error between the sensory data and the sensory inputs predicted by the expected causes.

Predictive coding sets out a process theory of information processing. One defined at the computational level in terms of Marr's levels of analysis (Marr, 1982). There are many possible ways in which this scheme could be realised at the algorithmic level, and several different algorithms have been proposed to implement predictive coding. This article sets out to describe each of these algorithms in order to provide a concise summary of their similarities and differences. The algorithms are reviewed in roughly the chronological order in which they were developed, starting with linear predictive coding (LPC) which was developed for signal processing not as a model of brain function. These ideas were then applied to explain efficient encoding in the retina and then subsequently to model approximate Bayesian inference in the cortical visual system (as described in the preceding paragraph). To aid comparison between algorithms a consistent mathematical notation is used throughout: x is used to denote sensory input (or the "signal"); y is used to denote the inferred causes of the sensory input (or the "coefficients"); V denotes the parameters of the generative model (or the "weights"); r denotes the sensory input predicted by the current estimate of the causes (or the "recon struction"); and e is used to denote the error between the reconstruction and the actual sensory input (or the "residual"). The same letters in bold are used to denote vectors and matrices containing multiple values of these parameters and variables.

2. Linear predictive coding in digital signal processing

Digital signal processing concerns the manipulation and analysis of a continuous signal, x, sampled at discrete time points (indexed by i) so that the signal is represented as a sequence of numbers, $x(i)$, called a "time series". The basic idea of linear predictive coding (Makhoul, 1975; O'Shaughnessy, 1988; Vaseghi, 2000) is that each sample of a time series can be approximated as a linear combination of preceding samples, such that:

$$x(i) \approx r(i) = y_1 x(i-1) + y_2 x(i-2) + \cdots + y_n x(i-n)$$

Or more compactly:

$$x(i) \approx r(i) = \sum_{j=1}^{n} y_j x(i-j) \tag{1}$$

where $r(i)$ is the estimate of $x(i)$ and n is a parameter, called the order of the model, that determines how many previous samples are used in the estimation. For the predictor coefficients, $y_1 \ldots y_n$, to be appropriate for estimating every sample, Eq. (1) needs to be true for all values of i. The coefficients are therefore determined by minimising the error (the squared difference) between the actual value of the signal and the linearly predicted one, summed over every sample in the time series:

$$\sum_i [x(i) - r(i)]^2$$

Several different methods (such as the autocorrelation method and the covariance method) have been developed for finding the parameters that minimise the sum of the squared error. For signals that vary over time (such as continuous speech) it is necessary to split the time series into shorter sequences (or "frames") and calculate the coefficients, separately, for each frame. Alternatively, it is possible to continuously update the coefficients as each new sample is received.

Having found the coefficients it is possible to use them to predict future samples of the signal. It is also possible to use the coefficients to estimate samples of the signal that are missing or have been corrupted. Hence, LPC has applications in signal interpolation, signal restoration, and noise reduction. The original signal is characterised by relatively few coefficients values. This can be used for signal compression, where only the coefficients and the first n samples need to be stored or transmitted and then the remaining signal is approximated (or synthesised) from these values by the recursive application of Eq. (1). Finally, the coefficients are a (compact) representation of the original signal. Similar signals should have similar coefficients which can be exploited to recognise similar signals or to identify the content of a signal by comparing its coefficients to those of known signals.

3. Predictive coding in retina

When LPC is applied to signal restoration, interpolation, compression or recognition (as described in the preceding paragraph), it is assumed that the coefficients, $y_1 \ldots y_n$, or the resulting reconstruction of the signal, $r(i)$, are informative and worth preserving, while the residual error between the prediction and the actual signal is uninformative and can be discarded. However, in other applications the opposite is true: the predictable component of the signal is removed to reduce the signal amplitude in order to allow more efficient transmission (Harrison, 1952; Oliver, 1952). In this case, the estimated value of the signal, as calculated by Eq. (1), is subtracted from the true value, $x(i)$, to determine the residual error, $e(i)$, for transmission:

$$e(i) = x(i) - \sum_{j=1}^{n} y_j x(i-j) \tag{2}$$

This residual has a smaller dynamic range than the original signal, and hence, can be transmitted with greater accuracy using the same bandwidth.

This form of predictive coding has been used to explain the function of the retina (Laughlin, 1990, Srinivasan, Laughlin, & Dubs, 1982). Specifically, it has been proposed that, at each location on the retinal surface, the coefficients act to calculate a moving average of the intensity of incident light, and that this average intensity is subtracted from the instantaneous value, $x(i)$. Srinivasan et al. (1982) extended this concept to the spatial domain, proposing that the predicted local intensity value is calculated from intensity values measured at nearby locations as well as from those measured at preceding times. To obtain the optimal estimate of the predicted intensity the coefficient values should change with the luminance (Srinivasan et al., 1982). More generally, experimental evidence suggests that the retina dynamically adjusts the coefficients (and hence the predicted intensity of the input) to the statistics of the current visual environment (Hosoya, Baccus, & Meister, 2005). By removing predictable information from the transmitted signal the retina can be considered to perform efficient coding or redundancy reduction (Attneave, 1954; Barlow, 1960, 2001; Laughlin, 1990, chap. 2; Olshausen & Field, 1996). However, it should be noted that if only the residual error is transmitted, then the receiver (in the case of the retina the receiver is the lateral geniculate nucleus and subsequently the cortex) cannot recover the components of the signal that have been removed, so rather than redundancy being reduced, redundant information is being removed.

4. Predictive coding in cortex: Rao and Ballard's algorithm

Consider applying Eq. (1) to predict a sequence of samples. Rather than writing a separate version of Eq. (1) for each sample, the calculation can be written in matrix form: