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:

$$\begin{pmatrix} x(1) \\ x(2) \\ \vdots \\ x(m) \end{pmatrix} \approx \begin{pmatrix} r(1) \\ r(2) \\ \vdots \\ r(m) \end{pmatrix}$$

$$= \begin{pmatrix} x(0) & x(-1) & x(-2) & \dots & x(1-n) \\ x(1) & x(0) & x(-1) & \dots & x(2-n) \\ \vdots & \vdots & \vdots & & \vdots \\ x(m-1) & x(m-2) & x(m-3) & \dots & x(m-n) \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix}$$

Or more compactly as:

$$\mathbf{x} \approx \mathbf{r} = \mathbf{V}\mathbf{y}$$

where bold lower-case letters correspond to vectors, and bold upper-case letters are matrices.

If the values of the coefficients ($\mathbf{y}$) are unknown, then (as mentioned in Section 2) appropriate values can be determined by finding those coefficients that minimise the sum of the squared error between the actual sample values ($\mathbf{x}$) and the predicted values ($\mathbf{V}\mathbf{y}$). One way to achieve this least-squares minimisation is to perform gradient descent on the residual error, by initialising $\mathbf{y}$ to zero and then recursively applying the following equations (Achler, 2014; Harpur, 1997):

$$\mathbf{e} = \mathbf{x} - \mathbf{V}\mathbf{y} \tag{3}$$

$$\mathbf{y} \leftarrow \mathbf{y} + \mu\mathbf{W}\mathbf{e} \tag{4}$$

where $\mathbf{W}$ is the transpose of $\mathbf{V}$ (*i.e.*, $\mathbf{W} = \mathbf{V}^T$) and μ is a parameter controlling the rate of the gradient descent. These equations can be implemented by a neural network, like that shown in Fig. 1a (Harpur, 1997). This neural network consists of two populations of neurons, configured like an autoencoder. One population of neurons receives the signal values ($\mathbf{x}$) and the predictions of these values ($\mathbf{V}\mathbf{y}$) and combines these using Eq. (3) to output the residual error. These will be called the error neurons. The other population of neurons receives inputs from the error neurons and implements Eq. (4) to update the estimates of the coefficients ($\mathbf{y}$). These will be called the prediction neurons. The two populations of neurons are linked by weighted connections with values defined by the matrices $\mathbf{W}$ and $\mathbf{V}$.

Up to this point it has been assumed that the inputs are a sequence of values from a time series and that the weight matrices contain preceding samples taken from the time series, and hence, that the neural network defined by Eqs. (3) and (4) implements LPC exactly. However, the algorithm is not constrained to this one application (indeed Harpur (1997) did not consider this case). The inputs need not be a temporal sequences of samples, they could be values recorded simultaneously from multiple sensors, they could be pixel intensity values from an image, they could be the firing rates of neurons, or anything else. Furthermore, the weight matrices are not restricted to contain input values recorded in the recent past, instead they are free parameters that define a generative model. One way to define the parameters of the generative model (*i.e.*, the weights of the neural network) is through learning. Harpur and Prager (1996) proposed a learning rule that adjusts the weights so as to reduce the residual error. Hence, learning (finding the parameters of the generative model, $\mathbf{V}$) and inference (calculating the coefficients, $\mathbf{y}$) can both be implemented by minimising, on different time-scales, the same objective function of reconstruction error minimisation.

The prediction of the input to the network, $\mathbf{V}\mathbf{y}$, is calculated as a linear combination of columns of $\mathbf{V}$ weighted by the coefficients $\mathbf{y}$. Each column of $\mathbf{V}$ (each row of $\mathbf{W}$) can be considered to be an "elementary component" or "basis vector". The inference process, implemented by Eqs. (3) and (4), finds appropriate values for the coefficients so that the basis vectors are added together in the correct proportions to reconstruct the input with minimal residual error. The basis vectors encode the possible causes of sensory inputs that are known to the neural network (*i.e.*, they constitute the network's internal model of the external environment). The coefficients thus represent the estimates of the causes of the current sensory-driven input.

The Rao and Ballard (1999) predictive coding model of cortical function proposes that the cortex is built from a hierarchy of networks like that described above (see Fig. 1b). In common with the predictive coding model of the retina (see Section 3), it is proposed that the output of each cortical region is the residual error (observe that Eq. (3) is equivalent to Eq. (2) for multiple samples). The feedforward connections between cortical regions are thus believed to transmit residual errors while the cortical feedback connections convey the predicted causes. Hence, this version of predictive coding, in common with several previous theories (*e.g.*, Barlow, 1994, chap. 1; Mumford, 1992), hypothesises that cortical feedback connections act to suppress information which is predicted by higher-level cortical regions. Rao and Ballard (1999) also included additional connections within the hierarchical architecture between error neurons and the preceding population of prediction neurons. These connections allow the predictions generated at one level of the hierarchy to influence (via the intervening error neurons) the predictions generated at the preceding stage of the hierarchy so that the predictions at different levels will be mutually consistent. If superscripts of the form *Si* indicate processing stage i of the hierarchical neural network, then (the linear version of) the Rao and Ballard (1999) algorithm is as follows:

$$\mathbf{y}^{Si} \leftarrow \nu\mathbf{y}^{Si} + \mu\mathbf{W}^{Si}\mathbf{e}^{Si-1} - \eta\mathbf{e}^{Si}$$

where

$$\mathbf{e}^{Si-1} = \mathbf{y}^{Si-1} - \mathbf{V}^{Si}\mathbf{y}^{Si} \quad \text{and (equivalently)} \quad \mathbf{e}^{Si} = \mathbf{y}^{Si} - \mathbf{V}^{Si+1}\mathbf{y}^{Si+1}$$

and ν, μ, and η are non-negative parameters.

5. Predictive coding in cortex: PC/BC-DIM

PC/BC-DIM is a version of Predictive Coding (PC; Rao & Ballard, 1999) reformulated to make it compatible with Biased Competition (BC) theories of cortical function (Spratling, 2008a, 2008b), and that is implemented using Divisive Input Modulation (DIM; Spratling, De Meyer, & Kompass, 2009) as the method for updating error and prediction neuron activations. DIM calculates the residual errors using division rather than subtraction. As a result the equations for calculating the prediction coefficients and the residual errors underlying the Rao and Ballard (1999) algorithm (*i.e.*, Eqs. (3)), are replaced with the following equations:

$$\mathbf{e} = \mathbf{x} \oslash (\epsilon_2 + \mathbf{V}\mathbf{y}) \tag{5}$$

$$\mathbf{y} \leftarrow (\epsilon_1 + \mathbf{y}) \otimes \mathbf{W}\mathbf{e} \tag{6}$$

where $\oslash$ and $\otimes$ indicate element-wise division and multiplication respectively, and ϵ_1 and ϵ_2 are non-negative parameters that, respectively, prevent prediction neurons becoming permanently non-responsive and prevent division-by-zero errors.

The motivations behind these changes are as follows. Firstly, the Rao and Ballard (1999) algorithm requires neurons to be able to produce both positive and negative firing rates, which is biologically-implausible. While it is possible to re-implement the algorithm using only non-negative firing rates (Ballard & Jehee, 2012), this results in a model that is extremely complex and requires a degree of coordination between the actions of different

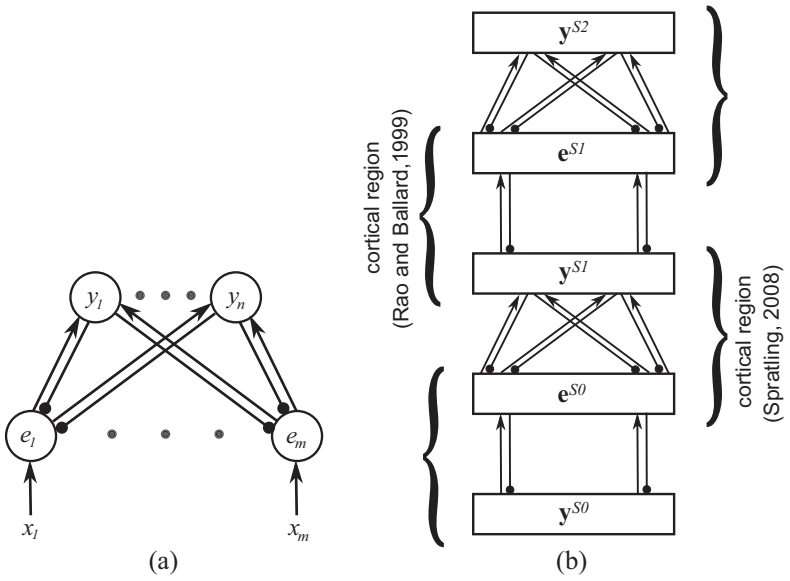

Fig. 1. (a) A neural network implementation of predictive coding. Neurons are shown as large white circles and the connections between neurons are shown as lines with arrowheads for excitatory connections and circular heads for inhibitory connections. (b) The hierarchical predictive coding architecture of Rao and Ballard (1999), which consists of a stack of networks like that shown in (a). Populations of neurons are shown as rectangles. The assignment of neural populations to cortical regions proposed by Rao and Ballard (1999) is indicated on the left, while that proposed by Spratling (2008a, 2008b) is indicated on the right. There is a one-to-one correspondence between neurons in a prediction population and those in the subsequent error neuron population (*e.g.*, between $\mathbf{y}^{S1}$ and $\mathbf{e}^{S1}$) and these neurons are connected in a one-to-one manner (indicated by the vertical connections in the diagram). In the PC/BC-DIM algorithm, the one-to-one inhibitory feedback connections are replaced by many-to-many excitatory connections (see Spratling, 2008a, 2008b, for details).

connections that is unlikely to be feasible in a biological system. In contrast, the neural activations in the PC/BC-DIM algorithm are inherently bounded to be non-negative. Secondly, as discussed in Section 4, Eqs. (3) and (4) perform gradient descent to find the coefficient values (*i.e.*, the estimates of the causes of the sensory input) that minimise the sum squared residual error. However, this method is slow if μ is too small and unstable if μ is too large (Harpur, 1997). In contrast, PC/BC-DIM is closely related to the particular method of performing non-negative matrix factorisation (NMF) proposed by Lee and Seung (2001) (Solbakken & Junge, 2011; Spratling et al., 2009). This form of NMF minimises the Kullback–Leibler (KL) divergence between the input ($\mathbf{x}$) and the reconstruction of the input ($\mathbf{Vy}$). As a result PC/BC-DIM is stable and converges quickly to an estimate of the causes ($\mathbf{y}$). The stability and speed of the PC/BC-DIM algorithm has enabled the development of very large scale simulations containing 10 s of millions of neurons and 100 s of billions[1] of connections (Spratling, 2013, 2014b). PC/BC-DIM typically finds a sparse set of causes, meaning that the input is reconstructed from a small number of basis functions. This is important for finding a unique solution when the set of basis functions is overcomplete (Olshausen & Field, 1997).[2] Thirdly, the Rao and Ballard (1999) algorithm proposes that inter-regional feedback connections carry predictions that need to be subtracted from the input signal in order to calculate the residual. Cortical feedback connections are the axon projections of a subpopulation of pyramidal cells which are excitatory. While a small proportion (10–20%) of these connections terminate on inhibitory neurons (Gonchar & Burkhalter, 2003), the primary targets are other pyramidal cells (Anderson & Martin, 2006; Budd, 1998; Cauller, 1995; Salin & Bullier, 1995), where cortical feedback has an excitatory (modulatory) effect on response (Hupé et al., 1998; Johnson & Burkhalter, 1997; Larkum, Senn, & Lüscher, 2004; Phillips, 2017; Shao & Burkhalter, 1996) at least in the short term (Shlosberg, Amitai, & Azouz, 2006). PC/BC-DIM proposes a different grouping of neural populations (see Fig. 1b) that requires both inter-cortical feedforward and feedback connections to be excitatory. Furthermore, if inter-cortical feedforward connections convey errors, as proposed by Rao and Ballard (1999), then it would be expected that a sub-population of cortical pyramidal cells, whose axon projections form the feedforward connections, should have response properties (as measured with single-cell electro-physiology) consistent with calculating error. While Rao and Ballard (1999) showed that error neuron responses could explain the phenomena of end-stopping observed in primary visual cortex pyramidal cells, Spratling (2010) showed that prediction neuron behaviour could explain end-stopping and a very wide range of other response properties observed in primary visual cortex pyramidal cells. These additional pyramidal cell response properties were further shown to be inconsistent with error neurons (Spratling, 2010, supplementary material). A final difference between PC/BC-DIM and the Rao and Ballard (1999) algorithm is that in PC/BC-DIM the requirement that $\mathbf{W} = \mathbf{V}^T$ is relaxed, although typically, the two sets of weights are still equal up to a scaling factor.

6. Predictive coding in cortex: free energy

The free energy principle is outwardly very similar to the Rao and Ballard (1999) algorithm, in that it proposes a hierarchy of layers that alternate between error detection and prediction and in which the prediction errors are conveyed by inter-cortical feedforward connections (Friston, 2009, 2010). However, unlike the Rao and Ballard (1999) algorithm, or any of the other algorithms described above, the variables in the free energy model do not represent the values of signals, instead they represent the statistics of these signals. Rather than represent individual samples (like the $x(i)$ values in Eq. (1)), the free energy model reconstructs the probability

[1] Where 1 billion equals 10^9.

[2] A version of the Rao and Ballard (1999) algorithm that finds a sparse set of coefficients has also been proposed (Jehee & Ballard, 2009; Jehee, Rothkopf, Beck, & Ballard, 2006).

distribution from which these samples are believed to come: it estimates a posterior probability density.

A particular, simplified, version of the free energy minimisation scheme has been described as "predictive coding" (Friston, 2009, 2010; Friston & Kiebel, 2009). This predictive coding algorithm assumes that the sensory data can be described by a Gaussian distribution (the "Laplace assumption"), and hence, that the posterior density is Gaussian (Friston, 2009, 2010; Friston & Kiebel, 2009). The parameterized generative model, which can have a deep hierarchical structure, acts to estimate the mean and variance of this Gaussian. The interaction between error and prediction layers in the model implements an inference process that serves to update the estimate the mean and variance, and this version of predictive coding is therefore equivalent to Kalman filtering (Bastos et al., 2012; Kiebel & Friston, 2011; Perrinet, Adams, & Friston, 2014). By precluding multimodal beliefs this model is consistent with our inability to perceive two things in the same place at the same time, as demonstrated by binocular rivalry and bistable perception (Hohwy, Roepstorff, & Friston, 2008).

7. Discussion

All the versions of predictive coding that have been reviewed above share the common computational goal of fitting a model to data. However, there are considerable differences in the specific mechanisms that they apply in order to achieve this goal. The form of the model varies between algorithms. It is a Gaussian in the free energy version of predictive coding, it is a sequence of previous samples from a time series in LPC, and it is a set of basis functions in the PC/BC-DIM and the Rao and Ballard (1999) algorithms. The criteria used to fit the model to the data also varies between algorithms. Fitting is achieved by minimising the sum of squared error in LPC and the Rao and Ballard (1999) algorithm, by minimising free energy in the method proposed by Friston (2010, 2009), and by minimising the KL divergence in PC/BC-DIM. Another difference is that some predictive coding algorithms are concerned with finding the coefficients which encode the underlying causes of the sensory data (see Sections 2 and 5), while others are concerned with finding these coefficients only for the purpose of calculating, and transmitting, the residual error (see Sections 3, 4, and 6). Hence, in some algorithms (*i.e.*, the predictive coding model of the retina and the Rao & Ballard (1999) model of cortex) redundant information is removed from the signal that is transmitted for further processing, while in others (*i.e.*, the PC/BC-DIM algorithm) it is the estimates of the causes of the input that are transmitted for further processing. The original LPC algorithm was used in some applications to do the former, and in other applications to do the latter.

Moving down to the implementation level of analysis (Marr, 1982), the algorithms also vary in how they propose that the prediction coefficients (the expected causes of the sensory data) are represented in biological neural circuits. The coefficients are believed to correspond to the synaptic weights of the lateral connections in the retina (Srinivasan et al., 1982), to the firing rates of feedback-projecting cortical pyramidal cells (Bastos et al., 2012; Friston, 2009; Kiebel & Friston, 2011), or the responses of feedforward-projecting pyramidal cells (Spratling, 2008b, 2012). Finally, the role envisaged for learning differs between the algorithms. Learning plays no role in LPC because the parameters of the generative model are defined to be previous samples taken from the time series. In contrast, learning, through minimising variational free energy, is used as a method of finding the parameters of the generative model in the free energy version of predictive coding. Similarly, learning is proposed as a way of defining the basis functions that form the generative model in the PC/BC-DIM (Spratling, 2012) and the Rao and Ballard (1999) algorithms. In both these algorithms, these parameters of the generative model are conceived as being stored in synaptic weights. This is in opposition to the predictive coding model of the retina (Srinivasan et al., 1982) in which synaptic weights are believed to define the prediction coefficients, not the parameters of the generative model.

Acknowledgements

Thanks to the organisers of, and the participants at, the Lorentz Centre Workshop on Perspectives on Human Probabilistic Inference (May 2014) for discussions that inspired this work. Additional thanks to Bill Phillips, and the two anonymous referees, for very helpful feedback on earlier drafts of this paper.

References

Achler, T. (2014). Symbolic neural networks for cognitive capacities. *Biologically Inspired Cognitive Architectures, 9*(0), 71–81.
Anderson, J. C., & Martin, K. A. C. (2006). Synaptic connection from cortical area v4 to v2 in macaque monkey. *The Journal of Comparative Neurology, 495*, 709–721.
Attneave, F. (1954). Informational aspects of visual perception. *Psychological Review, 61*, 183–193.
Ballard, D. H., & Jehee, J. (2012). Dynamic coding of signed quantities in cortical feedback circuits. *Frontiers in Psychology, 3*(254).
Barlow, H. B. (1960). The coding of sensory messages. In W. H. Thorpe & O. L. Zangwill (Eds.), *Current problems in animal behaviour* (pp. 331–360). Cambridge, UK: Cambridge University Press.
Barlow, H. B. (1994). What is the computational goal of the neocortex? In C. Koch & J. L. Davis (Eds.), *Large-scale neuronal theories of the brain* (pp. 1–22). Cambridge, MA: MIT Press.
Barlow, H. (2001). Redundancy reduction revisited. *Network: Computation in Neural Systems, 12*, 241–253.
Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron, 76*(4), 695–711.
Bubic, A., von Cramon, D. Y., & Schubotz, R. I. (2010). Prediction, cognition and the brain. *Frontiers in Human Neuroscience, 4*(25), 1–15.
Budd, J. M. L. (1998). Extrastriate feedback to primary visual cortex in primates: A quantitative analysis of connectivity. *Proceedings of the Royal Society of London. Series B, Biological Sciences, 265*(1400), 1037–1044.
Cauller, L. J. (1995). Layer I of primary sensory neocortex: Where top-down converges upon bottom-up. *Behavioural Brain Research, 71*(1–2), 163–170.
Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(03), 181–204.
Friston, K. J. (2009). The free-energy principle: A rough guide to the brain? *Trends in Cognitive Sciences, 13*(7), 293–301.
Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11*, 127–138.
Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 364*, 1211–1221.
Gonchar, Y., & Burkhalter, A. (2003). Distinct GABAergic targets of feedforward and feedback connections between lower and higher areas of rat visual cortex. *The Journal of Neuroscience, 23*(34), 10904–10912.
Harpur, G. F. (1997). *Low entropy coding with unsupervised neural networks*. PhD thesis. Department of Engineering, University of Cambridge.
Harpur, G., & Prager, R. (1996). Development of low entropy coding in a recurrent network. *Network: Computation in Neural Systems, 7*(2), 277–284.
Harrison, C. W. (1952). Experiments with linear prediction in television. *Bell System Technical Journal, 31*(4), 764–783.
Hohwy, J., Roepstorff, A., & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition, 108*(3), 687–701.
Hosoya, T., Baccus, S. A., & Meister, M. (2005). Dynamic predictive coding by the retina. *Nature, 436*(7047), 71–77.
Huang, Y., & Rao, R. P. N. (2011). Predictive coding. *WIREs Cognitive Science, 2*, 580–593.
Hupé, J. M., James, A. C., Payne, B. R., Lomber, S. G., Girard, P., & Bullier, J. (1998). Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons. *Nature, 394*(6695), 784–787.
Jehee, J. F. M., & Ballard, D. H. (2009). Predictive feedback can account for biphasic responses in the lateral geniculate nucleus. *PLoS Computational Biology, 5*(5), e1000373.
Jehee, J. F. M., Rothkopf, C., Beck, J. M., & Ballard, D. H. (2006). Learning receptive fields using predictive feedback. *Journal of Physiology – Paris, 100*, 125–132.
Johnson, R. R., & Burkhalter, A. (1997). A polysynaptic feedback circuit in rat visual cortex. *The Journal of Neuroscience, 17*(18), 7129–7140.
Kiebel, S. J., & Friston, K. J. (2011). Free energy and dendritic self-organization. *Frontiers in Systems Neuroscience, 5*(80).
Larkum, M. E., Senn, W., & Lüscher, H.-R. (2004). Top-down dendritic input increases the gain of layer 5 pyramidal neurons. *Cerebral Cortex, 14*(10), 1059–1070.
Laughlin, S. (1990). Coding efficiency and visual processing. In C. Blakemore (Ed.), *Vision: Coding and efficiency* (pp. 25–31). Cambridge University Press.

Lee, D. D., & Seung, H. S. (2001). Algorithms for non-negative matrix factorization. In T. K. Leen, T. G. Dietterich, & V. Tresp (Eds.). *Advances in neural information processing systems* (Vol. 13). Cambridge, MA: MIT Press.
Makhoul, J. (1975). Linear prediction: A tutorial review. *Proceedings of the IEEE, 63*(4), 561–580.
Marr, D. (1982). *Vision*. San Francisco, CA: Freeman.
Mumford, D. (1992). On the computational architecture of the neocortex II: The role of cortico-cortical loops. *Biological Cybernetics, 66*, 241–251.
Oliver, B. M. (1952). Efficient coding. *Bell System Technical Journal, 31*(4), 724–750.
Olshausen, B. A., & Field, D. J. (1996). Natural image statistics and efficient coding. *Network: Computation in Neural Systems, 7*(2), 333–339.
Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research, 37*(23), 3311–3325.
O'Shaughnessy, D. (1988). Linear predictive coding. *IEEE Potentials, 7*(1), 29–32.
Perrinet, L. U., Adams, R. A., & Friston, K. J. (2014). Active inference, eye movements and oculomotor delays. *Biological Cybernetics, 108*(6), 777–801.
Phillips, W. A. (2017). On the cognitive functions of intracellular mechanisms for contextual amplification. *Brain and Cognition, 112*, 39–53.
Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2*(1), 79–87.
Salin, P. A., & Bullier, J. (1995). Corticocortical connections in the visual system: Structure and function. *Physiological Reviews, 75*, 107–154.
Shao, Z., & Burkhalter, A. (1996). Different balance of excitation and inhibition in forward and feedback circuits of rat visual cortex. *The Journal of Neuroscience, 16*(22), 7353–7365.
Shlosberg, D., Amitai, Y., & Azouz, R. (2006). Time-dependent, layer-specific modulation of sensory responses mediated by neocortical layer 1. *Journal of Neurophysiology, 96*, 3170–3182.
Solbakken, L. L., & Junge, S. (2011). Online parts-based feature discovery using competitive activation neural networks. In *Proceedings of the international joint conference on neural networks* (pp. 1466–1473).
Spratling, M. W. (2008a). Predictive coding as a model of biased competition in visual selective attention. *Vision Research, 48*(12), 1391–1408.
Spratling, M. W. (2008b). Reconciling predictive coding and biased competition models of cortical function. *Frontiers in Computational Neuroscience, 2*(4), 1–8.
Spratling, M. W. (2010). Predictive coding as a model of response properties in cortical area V1. *The Journal of Neuroscience, 30*(9), 3531–3543.
Spratling, M. W. (2012). Unsupervised learning of generative and discriminative weights encoding elementary image components in a predictive coding model of cortical function. *Neural Computation, 24*(1), 60–103.
Spratling, M. W. (2013). Image segmentation using a sparse coding model of cortical area V1. *IEEE Transactions on Image Processing, 22*(4), 1631–1643.
Spratling, M. W. (2014a). Predictive coding. In D. Jaeger & R. Jung (Eds.), *Encyclopedia of computational neuroscience* (pp. 1–5). New York, NY: Springer.
Spratling, M. W. (2014b). A single functional model of drivers and modulators in cortex. *Journal of Computational Neuroscience, 36*(1), 97–118.
Spratling, M. W., De Meyer, K., & Kompass, R. (2009). Unsupervised learning of overlapping image components using divisive input modulation. *Computational Intelligence and Neuroscience, 2009*(381457), 1–19.
Srinivasan, M. V., Laughlin, S. B., & Dubs, A. (1982). Predictive coding: A fresh view of inhibition in the retina. *Proceedings of the Royal Society of London. Series B, Biological Sciences, 216*(1205), 427–459.
Vaseghi, S. V. (2000). *Advanced digital signal processing and noise reduction* (2nd ed.). John Wiley and Sons Ltd..