

Data Augmentation, Internal Representation, and Unsupervised Learning

Ying Nian Wu
UCLA Department of Statistics

I am grateful to the editor for inviting me to contribute this discussion. I have learned a great deal from this exceedingly clever paper by Yu and Meng. Ever since the ground breaking work of Meng and van Dyk (1997), there have been many interesting developments in the art of data augmentation for both EM and MCMC. This paper is yet another significant contribution to this line of research. While I feel I can contribute little to the discussion of the proposed method, I would like to mention a different perspective of data augmentation, in the hope of broadening the scope of the discussion. Data augmentation is not only a useful tool for MCMC, but it is also an essential ingredient in the so-called unsupervised learning, which involves augmenting latent variables or hidden units to explain the observed or visible data. In the context of neural science, the observed data are collected by the sensors in the form of images or sounds, and the latent variables or hidden units form the internal representations of the sensory data. The learning of such internal representations often does not require class labels or detailed annotations of the training examples, thus the learning is said to be unsupervised.

Latent variable models are abound in statistical literature, such as factor analysis, mixture model, t-model, random effects model, probit regression, hidden Markov model, just to name a few. In what follows, I shall briefly review two popular latent variable models in neural science and unsupervised learning, as well as their hierarchical extensions.

The first model is the sparse coding model of Olshausen and Field (1996). Let $Y = (y_1, \dots, y_M)$ be the M -dimensional vector, such as an image (where M is the number of pixels). Let $Z = (z_1, \dots, z_K)$ be the K -dimensional vector of hidden units for representing Y . The model is of the following form:

$$z_k \sim p(z) \text{ independently,} \quad (1)$$

$$Y = \sum_{k=1}^K z_k B_k + \epsilon, \quad (2)$$

where B_k 's are unknown M -dimensional basis vectors, and ϵ is the residual. The model appears to be very similar to factor analysis, except that K is often assumed to be greater than M , so that the representation is said to be "overcomplete." Moreover, $p(z)$ is assumed to be a heavy tailed distribution, such as Laplacian distribution, t-distribution, or a mixture of a point mass at 0 and a normal distribution with a large variance. Such $p(z)$ captures the sparsity of Z in the sense that most of the K components of Z are small or 0. ϵ is often assumed to be white noise although this is quite unrealistic. The goal is to learn the dictionary of the basis elements $\mathbf{B} = (B_k, k = 1, \dots, K)$ from training data $\{Y_i, i = 1, \dots, n\}$, such as n image patches randomly cropped from some images of natural scenes.

The second model is the restricted Boltzmann machine (Hinton, Osindero and Teh, 2006):

$$p(Y, Z \mid W) \propto \exp\left\{\sum_{k,m} w_{km} z_k y_m\right\}, \quad (3)$$

where both y_m and z_k are assumed to be binary, and $W = (w_{km}, k = 1, \dots, K, m = 1, \dots, M)$ are the unknown parameters or the connection weights between hidden units z_k and the visible units y_m . This model looks rather unusual to statisticians, in the sense that it is not in the form of $p(Z|W)$ and $p(Y|Z, W)$. In fact, the prior distribution $p(Z|W)$ is implicit, and only the joint distribution $p(Y, Z|W)$ is specified. However, this model has the advantage that both $p(Y|Z, W)$ and $p(Z|Y, W)$ are simple. Given Z and W , y_m are independent, and given Y and W , z_k are independent. The model can be extended to the situation where y_k are continuous, so that $p(Y|Z, W)$ is in a comparable form as in equation (2).

Both the sparse coding model (1) and (2) and the restricted Boltzmann machine (3) can be extended by introducing a higher layer of hidden variables on top of the layer of Z . The extension of (3) leads to the so-called deep belief network (Hinton, Osindero and Teh, 2006). The key observation in this endeavor is that the undirected graphical model $p(Y, Z|W)$ is equivalent to an infinite layer directed graphical model where each layer is a step of Gibbs sampler with $p(Y, Z|W)$ being the target distribution. The extension of (1) and (2) is quite different because of the sparsity of Z . Recently we propose an active basis model (Wu, Si, Gong and Zhu, 2010), where we assume that on top of Z is a layer of templates or partial templates, each being a composition of a small number of B_k 's selected from the dictionary $\mathbf{B} = (B_k, k = 1, \dots, K)$. Each selected B_k can be considered a "stroke" for sketching the template. See Figure 1 for three templates learned from natural images, where each B_k is illustrated by a small line segment, and the compositions of different sets of selected B_k 's form different templates.



Figure 1: Templates formed by different sets of selected B_k 's, where each B_k is depicted by a small line segment. Numbers of the selected B_k 's are, respectively, 80, 30, 50.

While statisticians may not be at home with discriminative supervised learning such as max-margin classification, the hierarchical latent variable models and the associated likelihood or Bayesian learning should be very familiar and natural to statisticians, who are well equipped to make useful contributions.

Acknowledgment

I would like to acknowledge the support of NSF DMS 1007889.

References

- [1] Hinton, G. E., Osindero, S. and Teh, Y. (2006) A fast learning algorithm for deep belief nets. *Neural Computation*, 18, 1527-1554.

- [2] Meng, X.-L. and van Dyk, D. (1997) The EM algorithm - an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society, B*, 59 511-567.
- [3] Olshausen, B. A. and Field, D. J. (1996) Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381 607-609.
- [4] Wu, Y. N., Si, Z., Gong, H., and Zhu, S. C. (2010) Learning active basis model for object detection and recognition. *International Journal of Computer Vision*, 90, 198-235.