LATENT VARIABLE MODELS

ABHISHEK SARKAR

1. Introduction

The essential aspect of a *latent variable model* (LVM) is that each observation $\mathbf{x}_i$ (which is p-dimensional, say), has an associated latent variable $\mathbf{z}_i$ (which is k-dimensional). Usually, one assumes k is much less than p, corresponding to an assumption that the differences in the high-dimensional observations $\mathbf{x}_i$ are (mostly) explained by differences in the lower-dimensional latent variables $\mathbf{z}_i$.

2. A simple example: Probabilistic PCA

One of the simplest examples of an LVM is *probabilistic PCA* (PPCA) [1]

$$\mathbf{x}_i \mid \mathbf{z}_i, \mathbf{W}, \sigma^2 \sim \mathcal{N}(\mathbf{W}\mathbf{z}_i, \sigma^2\mathbf{I}) \tag{1}$$

$$\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{2}$$

In this model, observations $\mathbf{x}_i$ are generated by first applying the linear transform $\mathbf{W}$ to the low-dimensional latent variable $\mathbf{z}_i$, and then adding multivariate Gaussian noise with covariance $\sigma^2\mathbf{I}$. There are two inference goals for PPCA:

(1) Estimate $\mathbf{W}, \sigma^2$ given observed data $\mathbf{x}_1, \ldots, \mathbf{x}_n$. This involves maximizing the likelihood, that is, solving the optimization problem

$$\hat{\mathbf{W}}, \hat{\sigma}^2 = \arg\max_{\mathbf{W}, \sigma^2} \sum_{i=1}^n \ln p(\mathbf{x}_i \mid \mathbf{W}, \sigma^2) \tag{3}$$

$$= \arg\max_{\mathbf{W}, \sigma^2} \sum_{i=1}^n \ln \left(\int_{\mathcal{Z}} p(\mathbf{x}_i \mid \mathbf{z}_i, \mathbf{W}, \sigma^2)\, dp(\mathbf{z}_i) \right), \tag{4}$$

where $\mathcal{Z}$ denotes the set of values $\mathbf{z}_i$ can take.

(2) Estimate $\mathbf{z}_i$ given $\mathbf{x}_i$. This involves computing the posterior mean $\mathrm{E}[\mathbf{z}_i \mid \mathbf{x}_i, \hat{\mathbf{W}}, \hat{\sigma}^2]$, using the result $\hat{\mathbf{W}}, \hat{\sigma}^2$ from the previous step.

How does one estimate $\boldsymbol{\theta}$ given observed data $\mathbf{x}_1, \ldots, \mathbf{x}_n$? In the case of PPCA, the integrals in equation (4) have closed forms (following from properties of the Gaussian distribution); thus,

$$\hat{\mathbf{W}}, \hat{\sigma}^2 = \arg\max_{\mathbf{W}, \sigma^2} \sum_{i=1}^n -\frac{1}{2} \ln\det(\mathbf{W}\mathbf{W}' + \sigma^2\mathbf{I}) - \frac{1}{2}\mathbf{x}_i'(\mathbf{W}\mathbf{W}' + \sigma^2\mathbf{I})^{-1}\mathbf{x}_i. \tag{5}$$

One can further show that the solution to the optimization problem (5) is unique and has closed forms related to the eigendecomposition of $\mathbf{X}'\mathbf{X}$, where $\mathbf{X}$ denotes the $n \times p$ matrix $(\mathbf{x}_1, \ldots, \mathbf{x}_n)'$. Specifically, $\hat{\mathbf{W}}$ is the matrix of the first k eigenvectors (having largest eigenvalues), and $\hat{\sigma}^2$ is the ratio of the sum of the remaining $p - k$ eigenvalues to the sum of all the eigenvalues. (Equivalently, the closed forms are related to the singular value decomposition of $\mathbf{X}$.)

How does one estimate $\mathbf{z}_i$ given $\mathbf{x}_i$? One can think of $p(\mathbf{x}_i \mid \mathbf{z}_i, \mathbf{W}, \sigma^2)$ as a likelihood and $p(\mathbf{z}_i)$ as the prior in a Bayesian setting. Both are Gaussian, and therefore conjugate; therefore, the posterior is also Gaussian:

$$\mathbf{z}_i \mid \mathbf{x}_i, \mathbf{W}, \sigma^2 \sim \mathcal{N}((\mathbf{W}'\mathbf{W} + \sigma^2\mathbf{I})^{-1}\mathbf{W}'\mathbf{x}_i, \sigma^2(\mathbf{W}'\mathbf{W} + \sigma^2\mathbf{I})^{-1}), \tag{6}$$

yielding the closed-form posterior mean $(\hat{\mathbf{W}}'\hat{\mathbf{W}} + \hat{\sigma}^2\mathbf{I})^{-1}\hat{\mathbf{W}}'\mathbf{x}_i$.

3. The general scheme

An LVM is typically given as a joint distribution

$$p(\mathbf{x}_i, \mathbf{z}_i \mid \boldsymbol{\theta}) = p(\mathbf{x}_i \mid \mathbf{z}_i, \boldsymbol{\theta}) p(\mathbf{z}_i \mid \boldsymbol{\theta}), \tag{7}$$

where $\boldsymbol{\theta}$ denotes a collection of additional model parameters. One can think of $p(\mathbf{x}_i \mid \mathbf{z}_i, \boldsymbol{\theta})$ as the likelihood, and $p(\mathbf{z}_i \mid \boldsymbol{\theta})$ as the prior in a Bayesian setting. There are two inference goals for LVMs:

(1) Estimate $\boldsymbol{\theta}$ given observed data $\mathbf{x}_1, \ldots, \mathbf{x}_n$. This involves solving an optimization problem

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \ln \left(\int_{\mathcal{Z}} p(\mathbf{x}_i \mid \mathbf{z}_i, \boldsymbol{\theta}) \, dp(\mathbf{z}_i \mid \boldsymbol{\theta}) \right), \tag{8}$$

where $\mathcal{Z}$ denotes the set of values $\mathbf{z}_i$ can take. This problem may not be tractable, and may require approximate inference methods. (In the case of PPCA, the integrals had closed forms.) In a Bayesian setting, this task corresponds to estimating the prior from the data, which is termed *empirical Bayes*. (In the case of PPCA, the prior did not have any free parameters to be estimated; however, in other models, the prior does have such parameters.)

(2) Estimate $\mathbf{z}_i$ given $\mathbf{x}_i, \hat{\boldsymbol{\theta}}$. This involves computing the posterior mean, by combining the likelihood with the prior estimated in the previous step. As above, this may not be tractable, and may require approximate inference methods. (In the case of PPCA, the prior was conjugate to the likelihood; therefore, the posterior mean had a closed form.)

One may also want to generate new observations $\mathbf{x}$. However, this is not "inference" as commonly understood in statistics; it is merely a case of random sampling from the distribution $p(\mathbf{x} \mid \hat{\boldsymbol{\theta}})$, which is achieved by first sampling $\mathbf{z} \sim p(\mathbf{z} \mid \hat{\boldsymbol{\theta}})$, then sampling $\mathbf{x} \sim p(\mathbf{x} \mid \mathbf{z}, \hat{\boldsymbol{\theta}})$.

4. Variational autoencoder

A more complex LVM is the *variational autoencoder* (VAE) [2]

$$\mathbf{x}_i \mid \mathbf{z}_i, \boldsymbol{\theta} \sim \mathcal{N}(f(\mathbf{z}_i), \sigma^2 \mathbf{I}) \tag{9}$$

$$\mathbf{z}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{10}$$

where f denotes a neural network with k-dimensional input and p-dimensional output, and $\boldsymbol{\theta}$ denotes σ^2 and the parameters of f. As a simple example, suppose f is a fully-connected feed-forward network with one hidden layer and a linear output layer, which can be written

$$f(\mathbf{z}_i) = \mathbf{W}_1 h(\mathbf{W}_0 \mathbf{z}_i + \mathbf{b}_0) + \mathbf{b}_1, \tag{11}$$

where h is a non-linearity (say, ReLU) applied element-wise. In this case, $\boldsymbol{\theta} = (\sigma^2, \mathbf{W}_0, \mathbf{W}_1, \mathbf{b}_0, \mathbf{b}_1)$, and the VAE is a non-linear version of PPCA. Specifically, one generates observations $\mathbf{x}_i$ by applying the non-linear transform f to the low-dimensional latent variable $\mathbf{z}_i$, and then adding multivariate Gaussian noise with covariance $\sigma^2\mathbf{I}$. As was the case for PPCA, the transform f will be learned from the data (in PPCA, the transform was $\mathbf{W}$); however, unlike PPCA, in general it does not have a closed form.

How does one estimate $\boldsymbol{\theta}$ given observed data $\mathbf{x}_1, \ldots, \mathbf{x}_n$? The integrals in equation (4) do not have closed forms for the VAE, so one needs to take a different strategy. The strategy taken for VAEs is *variational inference* [3]. Recall the objective function,

$$\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \ln \left(\int_{\mathcal{Z}} p(\mathbf{x}_i \mid \mathbf{z}_i, \boldsymbol{\theta}) p(\mathbf{z}_i \mid \boldsymbol{\theta}) \, d\mathbf{z}_i \right) \tag{12}$$

Introducing a distribution $q(\mathbf{z}_i \mid \boldsymbol{\phi})$ (sometimes termed the *variational surrogate* or *variational approximation*) by multiplying and dividing by its density, and optimizing over its parameters $\boldsymbol{\phi}$ also,

$$= \max_{\boldsymbol{\phi}, \boldsymbol{\theta}} \sum_{i=1}^{n} \ln \left(\int_{\mathcal{Z}} p(\mathbf{x}_i \mid \mathbf{z}_i, \boldsymbol{\theta}) \frac{q(\mathbf{z}_i \mid \boldsymbol{\phi})}{q(\mathbf{z}_i \mid \boldsymbol{\phi})} p(\mathbf{z}_i \mid \boldsymbol{\theta}) \, d\mathbf{z}_i \right) \tag{13}$$

By the definition of expectation with respect to $q(\mathbf{z}_i \mid \boldsymbol{\phi})$,

$$= \max_{\boldsymbol{\phi}, \boldsymbol{\theta}} \sum_{i=1}^{n} \ln \left(\mathrm{E}_q \left[\frac{p(\mathbf{x}_i \mid \mathbf{z}_i, \boldsymbol{\theta}) p(\mathbf{z}_i \mid \boldsymbol{\theta})}{q(\mathbf{z}_i \mid \boldsymbol{\phi})} \right] \right) \tag{14}$$

Finally, one can exchange log and expectation using Jensen's inequality,

$$\geq \max_{\boldsymbol{\phi}, \boldsymbol{\theta}} \sum_{i} \mathrm{E}_q [\ln p(\mathbf{x}_i \mid \mathbf{z}_i, \boldsymbol{\theta}) + \ln p(\mathbf{z}_i \mid \boldsymbol{\theta}) - \ln q(\mathbf{z}_i \mid \boldsymbol{\phi})]. \tag{15}$$

Equation (15) is the objective function for variational inference in LVMs, which can be optimized by a number of different algorithms, for example, gradient descent. (Variational inference in other models involves an objective function with similar form, involving the expectation of the joint probability density $\mathrm{E}_q[\ln p]$ and the expectation of the variational density $\mathrm{E}_q[\ln q]$.)

Before detailing how this objective function is optimized for VAEs specifically, first note that equation (13) holds for any distribution q (since we multiplied and divided by its density); however, one can show that the optimal q (equivalently, optimal $\boldsymbol{\phi}$) is the true posterior $p(\mathbf{z}_i \mid \mathbf{x}_i, \boldsymbol{\theta})$ (equivalently, the parameters of this distribution) [4]. Adding and subtracting $\ln p(\mathbf{x}_i \mid \boldsymbol{\theta})$,

$$\begin{aligned} = \max_{\boldsymbol{\phi},\boldsymbol{\theta}} \sum_{i=1}^{n} \mathrm{E}_q[&\ln p(\mathbf{x}_i \mid \boldsymbol{\theta}) + \ln p(\mathbf{x}_i \mid \mathbf{z}_i, \boldsymbol{\theta}) + \ln p(\mathbf{z}_i \mid \boldsymbol{\theta}) \\ &- \ln p(\mathbf{x}_i \mid \boldsymbol{\theta}) - \ln q(\mathbf{z}_i \mid \boldsymbol{\phi})] \end{aligned} \tag{16}$$

Using the definition of conditional probability,

$$= \max_{\boldsymbol{\phi},\boldsymbol{\theta}} \sum_{i=1}^{n} \mathrm{E}_q[\ln p(\mathbf{x}_i \mid \boldsymbol{\theta}) - \ln p(\mathbf{z}_i \mid \mathbf{x}_i, \boldsymbol{\theta}) - \ln q(\mathbf{z}_i \mid \boldsymbol{\phi})] \tag{17}$$

Using linearity of expectation,

$$= \max_{\boldsymbol{\phi},\boldsymbol{\theta}} \sum_{i=1}^{n} \mathrm{E}_q[\ln p(\mathbf{x}_i \mid \boldsymbol{\theta})] - \mathrm{E}_q[\ln p(\mathbf{z}_i \mid \mathbf{x}_i, \boldsymbol{\theta}) - \ln q(\mathbf{z}_i \mid \boldsymbol{\phi})] \tag{18}$$

Since $p(\mathbf{x}_i \mid \boldsymbol{\theta})$ does not depend on $\mathbf{z}_i$, $\mathrm{E}_q[\ln p(\mathbf{x}_i \mid \boldsymbol{\theta})] = \ln p(\mathbf{x}_i \mid \boldsymbol{\theta})$, and

$$= \max_{\boldsymbol{\phi},\boldsymbol{\theta}} \sum_{i=1}^{n} \ln p(\mathbf{x}_i \mid \boldsymbol{\theta}) - \mathrm{E}_q[\ln p(\mathbf{z}_i \mid \mathbf{x}_i, \boldsymbol{\theta}) - \ln q(\mathbf{z}_i \mid \boldsymbol{\phi})] \tag{19}$$

The latter term is the definition of the KL divergence between $q(\mathbf{z}_i \mid \boldsymbol{\phi})$ and $p(\mathbf{z}_i \mid \mathbf{x}_i, \boldsymbol{\theta})$; it equals zero when the two distributions are the same. Thus, the optimal q (equivalently, optimal ϕ), holding $\boldsymbol{\theta}$ fixed, is the true posterior given $\boldsymbol{\theta}$ (equivalently, its parameters), and the inequality (15) becomes an equality

$$= \max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \ln p(\mathbf{x}_i \mid \boldsymbol{\theta}). \tag{20}$$

However, for many models (including VAEs), the true posterior does not have a closed form. In this case, optimizing the objective function (15) with respect to q, where q can be any distribution, is difficult (since one needs to find parameters $\boldsymbol{\phi}$ that can represent any distribution). So, instead one optimizes over some restricted family of distributions $q \in \mathcal{Q}$ (equivalently, some specific choice of parameter space for the values $\boldsymbol{\phi}$), which may not contain the true posterior.

For example, one may assume that q is a Gaussian distribution with unknown mean $\boldsymbol{\mu}$ and unknown covariance matrix $\boldsymbol{\Sigma}$ (since then the posterior mean is easy to compute; not since this is necessarily the true posterior), so $\boldsymbol{\phi} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$. If the true posterior is not Gaussian, the inequality (15) is strict, and is termed the *evidence lower bound* (ELBO). And, from equation (19), maximizing the ELBO with respect to $\boldsymbol{\phi}$ holding $\boldsymbol{\theta}$ fixed is equivalent to minimizing the KL divergence between $q(\mathbf{z}_i \mid \boldsymbol{\phi})$ and $p(\mathbf{z}_i \mid \mathbf{x}_i, \boldsymbol{\theta})$ holding $\boldsymbol{\theta}$ fixed. Thus, the result will be the "best" approximating distribution q, in the sense that it has minimum KL divergence from the true posterior (given $\boldsymbol{\theta}$), despite not being able to compute the true posterior! Further, because one assumed q was Gaussian, computing the approximate posterior mean is trivial – it is just $\boldsymbol{\mu}$.

In a VAE, the model is parameterized by a neural network f, termed the *decoder network* (since it "decodes" the low-dimensional latent variable $\mathbf{z}_i$ to produce the observed $\mathbf{x}_i$, minus the noise). Therefore, it is natural to choose q to also be parameterized by a neural network, termed the *encoder network* (since it "encodes" $\mathbf{x}_i$ to produce $\mathbf{z}_i$). Specifically,

$$q(\mathbf{z}_i \mid \mathbf{x}_i, \boldsymbol{\phi}) = \mathcal{N}(m(\mathbf{x}_i), \operatorname{diag}(s^2(\mathbf{x}_i))), \tag{21}$$

where m, s^2 are each k-dimensional outputs of a neural network taking p-dimensional input, denoting the mean and diagonal of the covariance matrix, respectively. In essence, one assumes the posterior is multivariate Gaussian with diagonal covariance, which is a special case of the example we gave above. (The mean and covariance matrix depend on the data $\mathbf{x}_i$ for computational reasons [5]).

Thus, one estimates $\boldsymbol{\theta}$ given observed data $\mathbf{x}_1, \ldots, \mathbf{x}_n$ by optimizing the ELBO (15) with respect to $\boldsymbol{\phi}, \boldsymbol{\theta}$ (i.e., σ^2 and the neural network parameters of f, m, s^2). This procedure also yields an estimate $\hat{\boldsymbol{\phi}}$, corresponding to the approximate posterior distribution q (equivalently, the neural network parameters of m, s^2). For VAEs, this is achieved by stochastic optimization via gradient descent [2]. Briefly, the fundamental difficulty is that (15) does not have a closed form, and so it is replaced by a stochastic objective function that does have a closed form, whose expected value is the true objective function value, and whose gradient can be computed using backpropagation.

How does one estimate $\mathbf{z}_i$ given $\mathbf{x}_i$? The approximate posterior mean is $m(\mathbf{x}_i)$, that is, the output of the encoder network.

References

1. Tipping, M. E. & Bishop, C. M. Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61,** 611–622.
2. Kingma, D. P. & Welling, M. *Auto-Encoding Variational Bayes* in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings* (eds Bengio, Y. & LeCun, Y.) (2014).
3. Blei, D. M., Kucukelbir, A. & McAuliffe, J. D. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association* **112,** 859–877 (2017).

4. Neal, R. M. & Hinton, G. E. in *Learning in Graphical Models* (ed Jordan, M. I.) 355–368 (Springer Netherlands, Dordrecht, 1998).
5. Gershman, S. & Goodman, N. *Amortized inference in probabilistic reasoning* in *Proceedings of the annual meeting of the cognitive science society* **36** (2014).