Homework 3 COS 513

All figures, equations, and descriptions are adopted directly from Dr. Bishop's article: Variational Principal Components

Introduction

Here we show one key part of homework 3 (shown with an asterisk in the instructions) — the description of the graphical model for the method you have chosen.

The method we explore here is Bayesian probabilistic principal components analysis. This is a Bayesian analogue of probabilistic principal components analysis (PCA), where there is a graphical (probabilistic) model defined over the low-dimensional representation of the data.

One of the primary considerations in applying PCA in a real data setting is determining the latent dimensionality of the dataset. That is, "how many principal components should we 'keep' after performing PCA." To be concrete, PCA biplots are usually in 2-dimensions, which means retaining and plotting the first 2 PC's. However, we can visualize up to 3-dimensions, and perform a clustering method of choice on higher-dimensional data (typically after defining a kernel).

Standard PCA

Data D of d-dimensional vectors $D = t_n$ where $n \in 1, .., N$.

Then compute the variance-covariance matrix:

$S = \frac{1}{N} \sum_{n=1}^{N} (t_n - \bar{t})(t_n - \bar{t})^T$

Where $\bar{t} = \frac{1}{N} \sum_n t_n$

The eigenvectors u_i are the solutions to $Su_i = \lambda_i \mu_i$

The eigenvectors corresponding to the q largest eigenvalues, where $q < d$ are kept for further dimensional reduction, where, for PCA, the samples are projected onto the top q eigenvectors: $U_q^T(t_n - \bar{t})$

Probabilistic PCA

Define a latent variable variable x that is q-dimensional, such that $P(x) \sim N(x|0, I_q)$.

The observed variable, t follows:

$P(t|x) \sim N(t|Wx + \mu, \sigma^2 I_d)$.

t is simply a linear transformation of the latent variable x with Gaussian noise.

The marginal distribution is then defined as:

$P(t) = \int P(t|x)P(x)dx$
$P(t) = N(\mu, C)$

with $C = WW^T + \sigma^2 * I_d$.

The maximum likelihood estimate of W can be shown to be:

$W_{ML} = U_q(\Lambda - \sigma^2 I_q)^{1/2}$ where U_q and Λ_q are the q greatest eigenvectors and eigenvalues of the sample covariance matrix, respectively.

Now that we have a probabilistic framework, it is possible (more feasible in some circumstances) to optimize over values of q; however, a more 'automated' and compuatationally efficient approach is desireable.

Bayesian PCA

To further automate our choice of q we must define a prior distribution over the parameters of our model.

The first prior is a hierarchical prior over the matrix W. The prior over W has one hyperparameter, α, which is a q-dimensional vector, each element of which controls a column of the matrix W.

The completely specification of the prior follows a Gaussian distribution:

$P(W|\alpha) = \prod_{i=1}^{q} (\frac{\alpha_i}{2\pi})^{d/2} e^{-\frac{1}{2}\alpha_i ||w_i||^2}$

Each α_i is a prior over the precision of each w_i, such that if the posterior distribution for a given α_i is concentrated on a small variance (large precision), then the w_i direction will be small and "effectively 'switched off"'.

The graphical model and full specification of priors follows:

$P(\mu) = N(\mu|0, \beta^{-1}I)$
$P(\alpha) = \prod_{i=1}^{q} \Gamma(\alpha_i|a_\alpha, b_\alpha)$
$P(\tau) = \Gamma(\tau|c_\tau, d_\tau)$

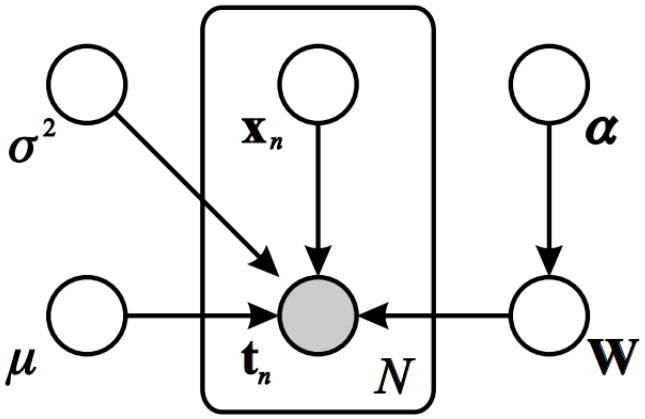

Figure 1: Representation of Bayesian PCA as a probabilistic graphical model showing the hierarchical prior over $\mathbf{W}$ governed by the vector of hyperparameters $\boldsymbol{\alpha}$. The box denotes a 'plate' comprising a data set of N independent observations of the visible vector $\mathbf{t}_n$ (shown shaded) together with the corresponding latent variables $\mathbf{x}_n$.