1 Partial least squares model

The partial least squares model is presented as a latent variable model in Kevin Murphy's 2012 book Machine Learning: A Probabilistic Perspective.

$$\begin{aligned} p(z) &= \mathcal{N}(z|0, I) & (1) \\ p(v|z, W, \mu, \sigma) &= \mathcal{N}(v|Wz + \mu, \sigma^2 I) & (2) \end{aligned}$$

where

$$\begin{aligned} W &= \begin{pmatrix} W_y & 0 \\ W_x & B_x \end{pmatrix} & (3) \\ z &= (z^s; z^x) & (4) \\ v &= (y; x) & (5) \\ \mu &= (\mu_y; \mu_x). & (6) \end{aligned}$$

Marginalising z gives

$$\begin{aligned} p(v|W, \mu, \sigma) &= \int \mathcal{N}(v|Wz + \mu, \sigma^2 I)\mathcal{N}(z|0, I)\, dz & (7) \\ &= \mathcal{N}(v|\mu, WW^T + \sigma^2 I) & (8) \end{aligned}$$

Conditioning on x gives

$$p(y|x) = \mathcal{N}(y|m_{y|x}, S_{y|x}) \qquad (9)$$

where

$$\begin{aligned} C &= \left(B_x B_x^T + W_x W_x^T + \sigma^2 I\right)^{-1} & (10) \\ m_{y|x} &= \mu_y + W_y W_x^T C(x - \mu_x) & (11) \\ S_{y|x} &= \sigma^2 I + W_y W_y^T - W_y W_x^T C W_x W_y^T & (12) \end{aligned}$$

Suppose we now obtain N independent observations from the model

$$v_n = (y_n; x_n), \quad 1 \leq n \leq N. \qquad (13)$$

We wish to estimate W, μ and σ. We can do this by maximising the likelihood of the data $v = x, y$ (8) using stochastic gradient descent.

One way to validate that our implementation works is to sample data from the model and compare the estimated parameters to those that we used to sample with. Suppose we sample v_n from our model parameterised by $\sigma^*, \mu_x^*, \mu_y^*, W_y^*, W_x^*, B_x^*$. We can estimate $\hat{\sigma}, \hat{\mu}_x, \hat{\mu}_y, \hat{W}_y, \hat{W}_x, \hat{B}_x$ that maximise

$$\sum_n \log p(v|\sigma, \mu, W_y, W_x, B_x) \qquad (14)$$

We would love to compare the estimated parameters to the underlying parameters but unfortunately (14) is invariant to orthonormal transformations of the W_y, W_x, B_x. That is

$$\sum_n \log p(v_n|\sigma, \mu, W_y U_s, W_x U_s, B_x U_x) = \sum_n \log p(v_n|\sigma, \mu, W_y, W_x, B_x) \tag{15}$$

for any orthonormal square matrices (of suitable dimension) U_s, U_x. We can see this as W only appears as WW^T in the likelihood and $WW^T = WUU^TW^T$ for any orthonormal U.