In [?] Lawrence noticed that, instead of integrating all latent variables, it is possible to integrate out all parameters.

We assume now that W has a Gaussian prior:

$$P(W) = \prod_{i=1}^{d} \mathcal{N}(\boldsymbol{w}_i|0, \boldsymbol{I})$$

Here d is the dimensionality of the observed space. Then all parameters are integrated out, leading to:

$$p(Y|X) = \prod_{i=1}^{d} \mathcal{N}(\boldsymbol{y}_i|0, XX^T + \sigma^2 I)$$

This is stated in the paper as the dual representation of the previous approach and is called the *Dual Probabilistic PCA*.

Lawrence also noted that the covariance matrix can be interpreted as a linear kernel $XX^T + \sigma^2 I = K$.

$$P(Y|X) = \prod_{i=1}^{d} \mathcal{N}(\boldsymbol{y}_i|0, K)$$

Thus the *Dual Probabilistic PCA* can be interpreted as product of *Gaussian Processes* with a linear kernel.

By exchanging the linear kernel with a non-linear one, we automatically have a technique for non-linear dimensionality reduction, which has a probabilistic interpretation [?]. The general non-linear model is called the *Gaussian Process - Latent Variable Model*. The *Gaussian Processes* learn a mapping from latent space to observed space - $\mathcal{N}(\boldsymbol{y}_i|0, K)$.

Using the trace properties

$$tr(a) = a \text{ , } a \text{ is a scalar}$$

$$tr(AB) = tr(BA)$$

we can change the mahalanobis distance term $\boldsymbol{y}_i^T K^{-1} \boldsymbol{y}_i = tr(\boldsymbol{y}_i^T K^{-1} \boldsymbol{y}_i) = tr(K^{-1} \boldsymbol{y}_i \boldsymbol{y}_i^T)$

The log likelihood is thus [?]:

$$log\, p(Y|X) \propto -\frac{p}{2} \log\left(|K|\right) - \frac{1}{2} tr(K^{-1} YY^T)$$

In the general case, when there is no linear kernel any more, this equation cannot be solved in closed form. Therefore we have to do gradient based optimization. But marginalizing over all latent space samples means that we have to include these in the optimization. This fact makes the optimization problem very hard as the dimensionality is high – number of samples n times number of latent dimensions l + hyperparameters – and, for general problems, has many local minima.

As initially proposed the standard way of initializing the latent space is using *Principal Components Analysis*.