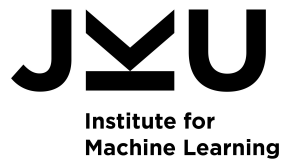
UNIT 2

Bayes Techniques, pPCA

Johannes Brandstetter
Institute for Machine Learning



Copyright statement:

Likelihood

- Likelihood is defined as:

$$\mathcal{L}(\{\boldsymbol{z}\};\boldsymbol{w}) = p(\{\boldsymbol{z}\};\boldsymbol{w}) = \prod_{i=1}^{l} p(\boldsymbol{z}_i;\boldsymbol{w})$$

- I.e., probability of the (data-distribution) model $p(\boldsymbol{z};\boldsymbol{w})$ to (re-)produce the dataset
- In supervised learning we can write:

$$p(\boldsymbol{z};\boldsymbol{w}) = p(\boldsymbol{x})p(y|\boldsymbol{x};\boldsymbol{w})$$

- and

$$p(\{\boldsymbol{z}\};\boldsymbol{w}) = \prod_{i=1}^{l} p(\boldsymbol{x}_i) \prod_{i=1}^{l} p(y_i|\boldsymbol{x}_i;\boldsymbol{w})$$

Likelihood

- Because $\prod_{i=1}^{l} p(\boldsymbol{x}_i)$ is independent of the parameters, it is sufficient to maximize the conditional likelihood:

$$p(y|\{\boldsymbol{x}\};\boldsymbol{w}) = \prod_{i=1}^{l} p(y_i|\boldsymbol{x}_i;\boldsymbol{w})$$

- 2 views: (i) parameter vector $\boldsymbol{w}$ is used to parameterize the likelihood, and (ii) **likelihood is conditioned on w**!
- **Supervised machine learning**: We choose parameters $\boldsymbol{w}$ such that they maximize the probability of the observed dataset $\{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_l\}$ given our choice of model $\mathcal{M}$:

$$p(\{\boldsymbol{z}\}|\boldsymbol{w}, \mathcal{M}) = \prod_{i=1}^{l} p(\boldsymbol{z}_i|\boldsymbol{w}, \mathcal{M}) \ .$$

- Usually $\mathcal{M}$ is left out for simplicity

Frequentist approach

- From a probabilistic perspective, frequentists are trying to maximize the likelihood $p(\{\boldsymbol{z}\}|\boldsymbol{w}, \mathcal{M})$
- A (statistical) model is simply a probability distribution over data $\Rightarrow$ **Maximum likelihood estimation** (MLE) to obtain, or "train", predictive models
- If the model is complex enough, only maximizing the likelihood **leads to overfitting**
- In the most extreme case the model produces the training examples with equal probability and other data with probability zero:
  $\Rightarrow p(\{\boldsymbol{z}\}|\boldsymbol{w}, \mathcal{M})$ is the sum of Dirac delta-distributions

From Frequentist to the Bayesian approach

- E.g., to avoid overfitting we can assume that **certain w are more probable to be observed** in the real world than others
- That means some models are more likely
- The fact that some models are more likely can be expressed by a distribution $p(\boldsymbol{w})$, the **prior distribution**
- The information in $p(\boldsymbol{w})$ stems from prior knowledge about the problem

Bayesianists

- Crucial property of the Bayesian approach is to realistically **quantify uncertainty**
- Instead of a parameter point estimate, a Bayesian approach defines a full **probability distribution** over parameters $\Rightarrow$ **posterior distribution**
- Posterior represents our **belief/hypothesis/uncertainty** about the value of each parameter (setting)

Bayes' theorem

- Bayes' theorem:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

- Bayes' theorem for data modeling:

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Evidence}}, \text{ or } p(\boldsymbol{w}|\{\boldsymbol{z}\}) = \frac{p(\{\boldsymbol{z}\}|\boldsymbol{w}) \times p(\boldsymbol{w})}{p_{\boldsymbol{w}}(\{\boldsymbol{z}\})}$$

- **Prior distribution** $p(w)$ over the parameters to capture our belief about what our model parameters should look like prior to observing any data
- Using our dataset, we can **update** (multiply) our prior belief with the **likelihood** $p(\{\boldsymbol{z}\}|w)$

Bayes' theorem cont'd

- Likelihood $p(\{\boldsymbol{z}\}|\boldsymbol{w})$ is the same quantity we saw in the frequentist approach. It tells us how well the observed data is explained by a specific parameter setting $\boldsymbol{w}$
- The product $p(\{\boldsymbol{z}\}|\boldsymbol{w})p(\boldsymbol{w})$ must be evaluated for **each parameter** setting, and normalized:
  $\Rightarrow$ **Marginalizing** (summing or integrating) over all parameter settings
- The normalization constant
  $$p_{\boldsymbol{w}}(\{\boldsymbol{z}\}) = \int_W p(\{\boldsymbol{z}\}|\boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w}$$
  is called **evidence** for a class of models
- $p_{\boldsymbol{w}}(\{\boldsymbol{z}\})$ provides evidence for how good our model as a whole, i.e., how likely the data is, taking into account all **parameter settings**

Example: Conjugate prior

- **Conjugate distribution** (conjugate pair): pair of (i) sampling distribution (likehood) and (ii) prior distribution, for which the resulting posterior distribution belongs into the same parametric family of distributions than the prior distribution
- We say that the **prior distribution** is a conjugate prior for this sampling distribution
- Example: Gamma distribution is a conjugate prior of the Poisson distribution
- Conjugate priors give convenience: we can do statistical inference, e.g., hypothesis testing, with the posterior distribution

Conjugate pair: Gamma & Poisson distr

- Gamma distribution:

$$f(\boldsymbol{x};\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}\boldsymbol{x}^{\alpha-1}e^{-\beta\boldsymbol{x}} \;, \quad \Gamma(n) = (n-1)!$$

- Likelihood (Poisson distribution):

$$p(\{\boldsymbol{x}\}|\lambda) = \prod_{i=1}^{l} p(\boldsymbol{x}_i|\lambda) = \prod_{i=1}^{l} \lambda^{\boldsymbol{x}_i}\frac{e^{-\lambda}}{\boldsymbol{x}_i!} \propto \lambda^{\sum_{i=1}^{l}\boldsymbol{x}_i}e^{-l\lambda} \;, \quad \boldsymbol{x}_i = \{0,1,\ldots\}$$

- We treat the term $\int \propto p(\{\boldsymbol{x}\}|\lambda)p(\lambda)$ as constant since it does not depend on λ
- **Unnormalized posterior distribution** for λ:

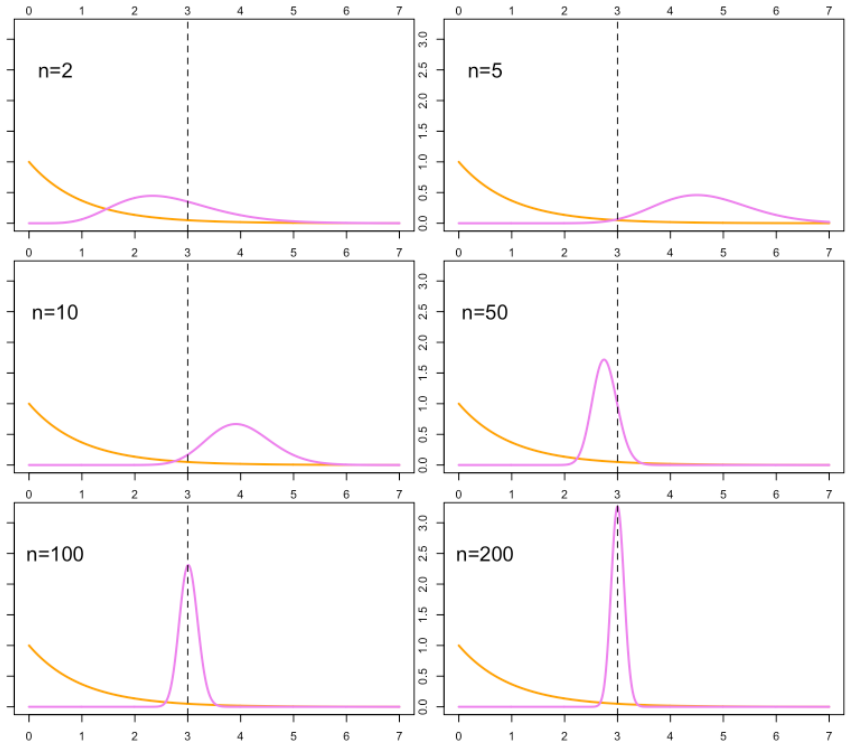
$$\begin{aligned} p(\lambda|\{\boldsymbol{x}\}) &\propto p(\{\boldsymbol{x}\}|\lambda)p(\lambda) \\ &\propto \lambda^{l\bar{\boldsymbol{x}}}e^{-l\lambda}\lambda^{\alpha-1}e^{-\beta\alpha} \\ &\propto \lambda^{\alpha+l\bar{\boldsymbol{x}}-1}e^{-(\beta+l)\lambda} \quad \text{with} \quad \bar{\boldsymbol{x}} = \sum_{i=1}^{l}\boldsymbol{x}_i \end{aligned}$$

- The posterior distribution is thus:

$$p(\lambda|\{\boldsymbol{x}\}) \propto \Gamma(\alpha + l\bar{\boldsymbol{x}}, \beta + l)$$

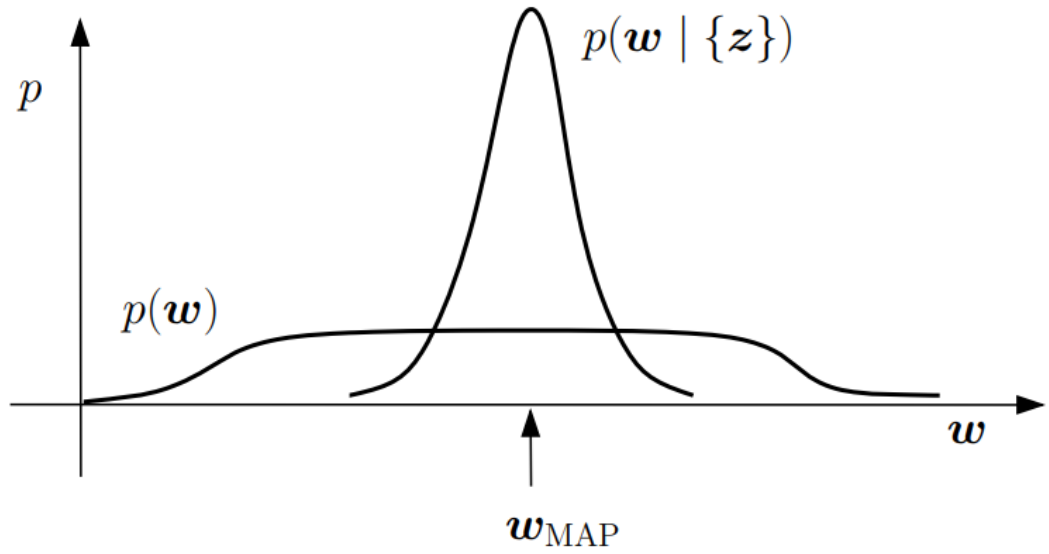
Conjugate pair: Gamma & Poisson distribution

- Example: Generate observations from Poisson distribution with a known parameter $\lambda = 3$
- Prior distribution: $\lambda \sim \text{Gamma}(\alpha = 1, \beta = 1)$
- Prior distribution: $\lambda \sim \text{Gamma}(\alpha = 1, \beta = 1)$



Example: Maximum A posteriori (MAP)

- The Maximum A Posteriori Approach (MAP) searches for the maximal posterior $p(\boldsymbol{w}|\{\boldsymbol{z}\})$



- MAP estimate picks the parameter setting that has the **highest probability** assigned to it (the distribution's mode)
- MAP provides a point estimate (not fully Bayesian)

MAP for Gaussian weight prior

- Gaussian weight prior:

$$p(\boldsymbol{w}) = \frac{1}{Z_{\boldsymbol{w}}(\alpha)} \exp\left(-\frac{1}{2}\alpha\|\boldsymbol{w}\|^2\right)$$

$$Z_{\boldsymbol{w}}(\alpha) = \int_W \exp\left(-\frac{1}{2}\alpha\|\boldsymbol{w}\|^2\right) d\boldsymbol{w} = \left(\frac{2\pi}{\alpha}\right)^{\frac{W}{2}}$$

- The negative log-posterior is:

$$-\ln p(\boldsymbol{w}|\{\boldsymbol{z}\}) = -\ln(\{\boldsymbol{z}\}|\boldsymbol{w}) - \ln p(\boldsymbol{w}) + \ln p_{\boldsymbol{w}}(\{\boldsymbol{z}\})$$

- For MAP estimation only $-\ln(\{\boldsymbol{z}\}|\boldsymbol{w}) - \ln p(\boldsymbol{w})$ must be minimized which results in the terms:

$$R(\boldsymbol{w}) = \underbrace{\beta R_{\text{emp}}}_{\text{empirical error}} + \underbrace{\alpha\Omega(\boldsymbol{w})}_{\text{complexity term}}$$

- The complexity term $\Omega(\boldsymbol{w}) = \|\boldsymbol{w}\|^2$ can be viewed in most cases as MAP estimation

MAP for Gaussian weight prior

- For Gaussian noise models, the likelihood is the exponential function with empirical error as an argument:

$$p(\{\boldsymbol{z}|\boldsymbol{w}\}) = \frac{1}{Z_R(\beta)} \exp\left(-\frac{1}{2}\beta R_{\text{emp}}\right) ,$$

- the prior is an exponential function of the complexity:

$$p(\boldsymbol{w}) = \frac{1}{Z_W(\alpha)} \exp\left(-\frac{1}{2}\alpha\Omega(\boldsymbol{w})\right) ,$$

- and the posterior is:

$$p(\boldsymbol{w}|\{\boldsymbol{z}\}) = \frac{1}{Z(\alpha,\beta)} \exp\left(-\frac{1}{2}(\alpha\Omega(\boldsymbol{w}) + \beta R_{\text{emp}})\right)$$

$$Z(\alpha,\beta) = \int_W \exp\left(-\frac{1}{2}(\alpha\Omega(\boldsymbol{w}) + \beta R_{\text{emp}})\right) d\boldsymbol{w}$$

Evidence

- The full fledged Bayesian approach is to specify a **predictive distribution**, i.e., the evidence:

$$p(y|\{\boldsymbol{z}\}, \boldsymbol{x}) = \int p(y|\boldsymbol{w}, \boldsymbol{x}) p(\boldsymbol{w}|\{\boldsymbol{z}\}) d\boldsymbol{w}$$

- This defines the probability for class label y given new input x and dataset $\{\boldsymbol{z}\}$
- To compute the evidence we need to **marginalize over all parameter settings**
- We multiply the posterior probability of each setting w with the probability of label y given input x using parameter setting $w \Rightarrow$ **Bayesian Model Averaging (BMA)**

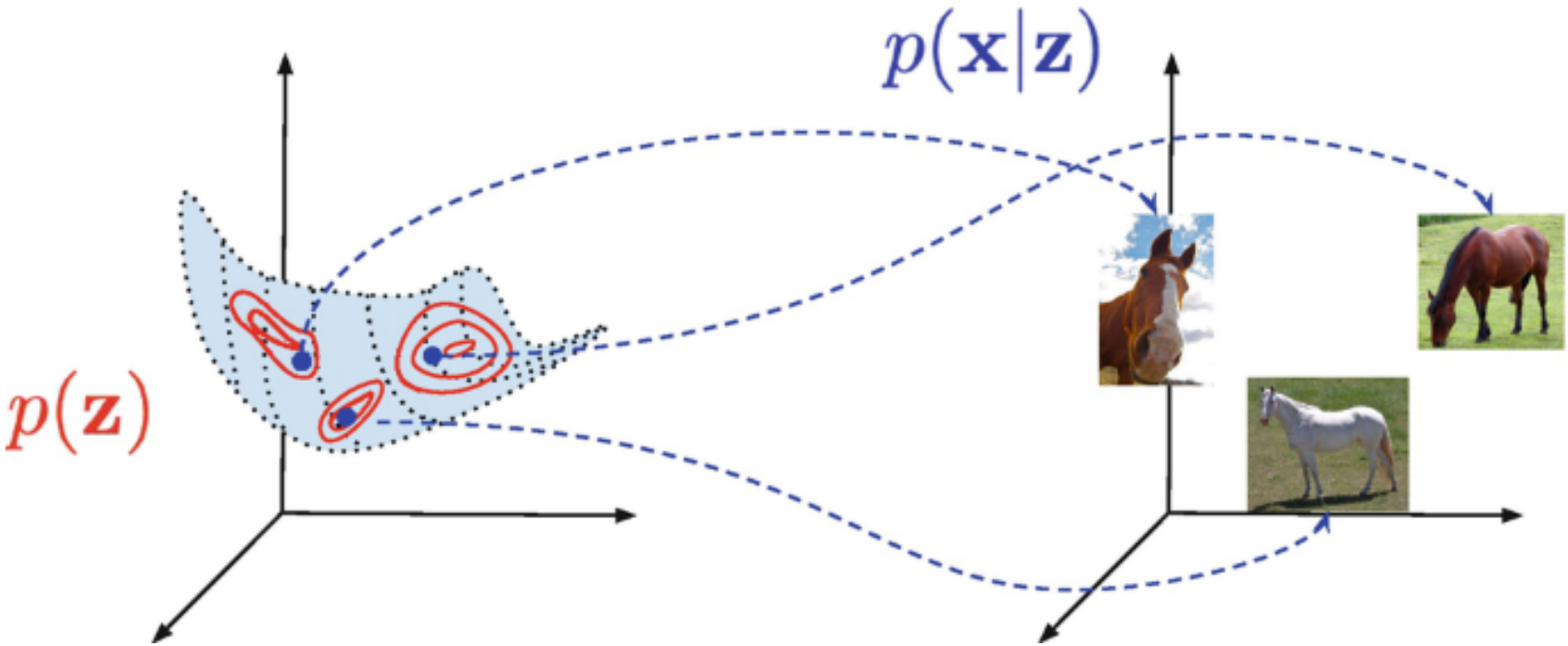
Marginalization

- In practice marginalizing over the whole parameter space is often impossible (**intractable**) as we can have infinitely many such settings $\Rightarrow$ Bayesian approach is fundamentally about marginalization
- Different family of methods to learn parameter values:
  - **Sampling methods** such as Markov Chain Monte Carlo (MCMC)
  - **Variational inference** techniques use a simpler, tractable family of distributions. Prominent examples are **Variational Autoencoders**

Probabilistic Principal Component Analysis

Latent variable models

Deep Generative Modeling, Tomczak



Latent variable models

- Describe generative process as:

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{z} = \int p(\boldsymbol{x}|\boldsymbol{z}) p(\boldsymbol{z}) d\boldsymbol{z}$$

- How to calculate this integral?
  - Tractable: Integrate
  - Intractable: Variational inference
- !! Unfortunately, literature uses $\boldsymbol{z}$ for latent variables. This coincides with $\{\boldsymbol{z}\}$ which we used so for to describe dataset !!

Principal Component Analysis

- Mean and covariance of dataset $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_l\}$, where $\boldsymbol{x}_i = (x_i^{(1)}, \ldots, x_i^{(d)})^T \in \mathbb{R}^d$:

$$\bar{\boldsymbol{x}} = \frac{1}{l}\sum_{i=1}^{l} \boldsymbol{x}_i \ , \quad \mathbf{S} = \frac{1}{l}\sum_{i=1}^{l} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T$$

- We consider a unit vector $\boldsymbol{u}_1 \in \mathbb{R}^d, i.e., \boldsymbol{u}_1^T \boldsymbol{u}_1 = 1$
- $\boldsymbol{u}_1^T \boldsymbol{x}_i$ projects each data point onto a scalar value
- The variance of the projected data is given by:

$$\frac{1}{l}\sum_{i=1}^{l} \left(\boldsymbol{u}_1^T \boldsymbol{x}_i - \boldsymbol{u}_1^T \bar{\boldsymbol{x}}\right)^2 = \boldsymbol{u}_1^T \mathbf{S} \boldsymbol{u}_1$$

Principal Component Analysis

- We **maximize the projected variance** $u_1^T \mathbf{S} u_1$ wrt $\boldsymbol{u}_1$
- Normlization condition $\boldsymbol{u}_1^T \boldsymbol{u}_1 = 1 \Rightarrow$ we enforce normalization constraint with Lagrangian multiplier α_1
- Unconstrained maximization of

$$\boldsymbol{u}_1^T \mathbf{S} \boldsymbol{u}_1 + \alpha_1 \left(1 - \boldsymbol{u}_1^T \boldsymbol{u}_1\right)$$

- Stationary point when:

$$\mathbf{S} \boldsymbol{u}_1 = \alpha_1 \boldsymbol{u}_1$$

- $\Rightarrow \boldsymbol{u}_1$ must be an eigenvector of $\mathbf{S}$
- Using $\boldsymbol{u}_1^T \mathbf{S} \boldsymbol{u}_1 = \alpha_1$: projected variance is maximum when $\boldsymbol{u}_1$ equals the eigenvector of $\mathbf{S}$ with the largest eigenvalue $\lambda_1 \Rightarrow \boldsymbol{u}_1$ is the first principal component of $\mathbf{S}$
- PCA: **linear projection** of the data onto a subspace of lower dimensionality than the original data space

Probabilistic Principal Component Analysis

- All of the marginal and conditional distributions are Gaussian:
  - **Latent variable** $\boldsymbol{z} \in \mathbb{R}^m$ corresponding to the $m-$**dimensional principal-component** subspace
  - Gaussian prior distribution $p(\boldsymbol{z})$ over the latent variable
  - Gaussian conditional distribution $p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\mathbf{W}\boldsymbol{z} + \boldsymbol{\mu}, \sigma^2\mathbf{I})$ for the observed data $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_l\} \in \mathbb{R}^d$ conditioned on $\boldsymbol{z}$:

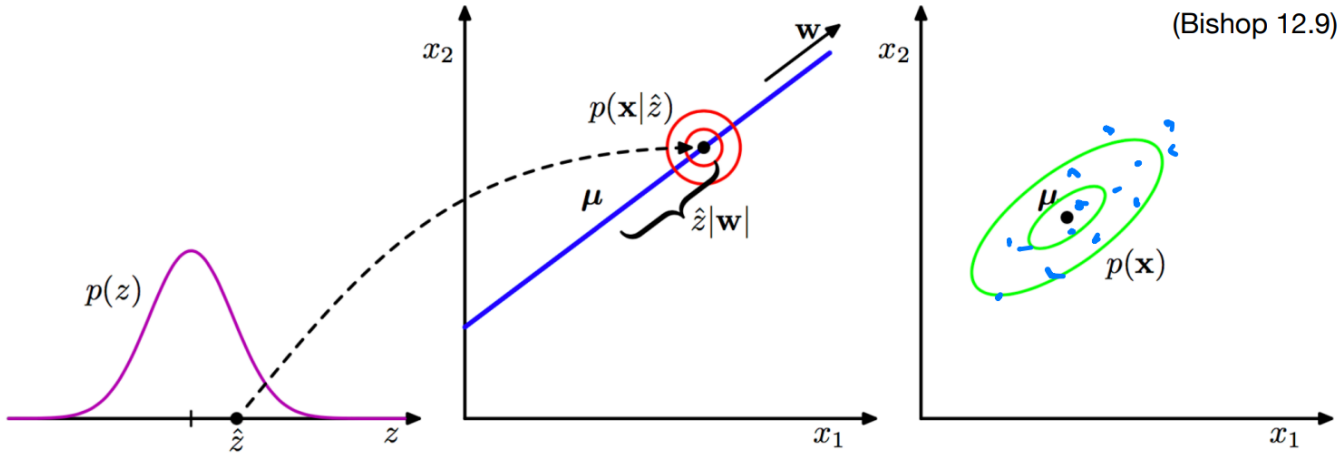
$$p(\boldsymbol{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$$
$$p(\boldsymbol{x}|\boldsymbol{z}) = \mathcal{N}(\mathbf{W}\boldsymbol{z} + \boldsymbol{\mu}, \sigma^2\mathbf{I}) \,, \quad \mathbf{W} \in \mathbb{R}^{d\times m}, \; \boldsymbol{\mu} \in \mathbb{R}^d, \; \mathbf{I} \in \mathbb{R}^{d\times d}$$

  - ... assuming a linear dependency between $\boldsymbol{z}$ and $\boldsymbol{x}$:

$$\boldsymbol{x} = \mathbf{W}\boldsymbol{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon} \,, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}|\mathbf{0}, \sigma^2\mathbf{I})$$

- The columns of $\mathbf{W}$ span a **linear subspace** within the data space that corresponds to the principal subspace

Probabilistic Principal Component Analysis



- We choose a value for the latent z
- ... and then sample x from an isotropic Gaussian distribution with mean $\mathbf{W}z + \boldsymbol{\mu}$ and covariance $\sigma\mathbf{I}$

Probabilistic Principal Component Analysis

- The marginal distribution $p(\boldsymbol{x}) = \int p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})$ is also Gaussian (sum of Gaussians is Gaussian, product of Gaussians is Gaussian), i.e., $\mathcal{N}(\boldsymbol{x}|?,?)$

$$
\begin{aligned}
\mathrm{E}_{p(\boldsymbol{x}|\boldsymbol{z})}(\boldsymbol{x}) &= \mathrm{E}_{p(\boldsymbol{x}|\boldsymbol{z})}(\mathbf{W}\boldsymbol{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon}) \\
&= \mathbf{W}\underbrace{\mathrm{E}_{p(\boldsymbol{x}|\boldsymbol{z})}(\boldsymbol{z})}_{=0} + \mathrm{E}_{p(\boldsymbol{x}|\boldsymbol{z})}(\boldsymbol{\mu}) + \underbrace{\mathrm{E}_{p(\boldsymbol{x}|\boldsymbol{z})}(\boldsymbol{\epsilon})}_{=0} \\
&= \boldsymbol{\mu} \\
\mathrm{Cov}_{p(\boldsymbol{x}|\boldsymbol{z})}(\boldsymbol{x}) &= \mathrm{E}_{p(\boldsymbol{x}|\boldsymbol{z})}\left((\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T\right) \\
&= \mathrm{E}_{p(\boldsymbol{x}|\boldsymbol{z})}\left((\mathbf{W}\boldsymbol{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon} - \boldsymbol{\mu})(\mathbf{W}\boldsymbol{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon} - \boldsymbol{\mu})^T\right) \\
&= \mathrm{E}_{p(\boldsymbol{x}|\boldsymbol{z})}\left(\mathbf{W}\boldsymbol{z}\boldsymbol{z}^T\mathbf{W}^T + \mathbf{W}\boldsymbol{z}\boldsymbol{\epsilon}^T + \boldsymbol{\epsilon}\boldsymbol{z}^T\mathbf{W}^T + \boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\right) \\
&= \mathbf{W}\underbrace{\mathrm{E}(\boldsymbol{z}\boldsymbol{z}^T)}_{=\mathbf{I}}\mathbf{W}^T + \underbrace{\mathbf{W}\mathrm{E}(\boldsymbol{z}\boldsymbol{\epsilon}^T)}_{=\mathbf{W}\mathrm{E}(\boldsymbol{z})\mathrm{E}(\boldsymbol{\epsilon}^T)} + \underbrace{\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{z}^T)\mathbf{W}^T}_{=\mathrm{E}(\boldsymbol{\epsilon})\mathrm{E}(\boldsymbol{z}^T)\mathbf{W}^T} + \underbrace{\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)}_{=\sigma^2\mathbf{I}} \\
p(\boldsymbol{x}) &= \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})
\end{aligned}
$$

True posterior

- Due to properties of Gaussians (Bishop 2.116), we can also calculate the true posterior $p(\boldsymbol{z}|\boldsymbol{x})$

$$p(\boldsymbol{z}|\boldsymbol{x}) = \mathcal{N}\left((\mathbf{W}^T\mathbf{W} + \sigma^2\mathbf{I})^{-1}\mathbf{W}^T(\boldsymbol{x} - \boldsymbol{\mu}), \sigma^{-2}(\mathbf{W}^T\mathbf{W} + \sigma^2\mathbf{I})\right)$$

- If we find $\mathbf{W}$, we can calculate $p(\boldsymbol{z}|\boldsymbol{x}) \Rightarrow$ maximum-likelihood estimation
- For a given observation $\boldsymbol{x}$, we can calculate the **distribution over the latent factors**

Maximum likelihood PCA

- Log-likelihood of $p(\boldsymbol{x})$:

$$\begin{aligned}\ln\left(p(\boldsymbol{x})\right) &= \sum_{i=1}^{l} \ln\left(p(\boldsymbol{x}_i|\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I})\right) \\ &= -\frac{ld}{2}\ln(2\pi) - \frac{l}{2}\ln\left(|\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}|\right) \\ &\quad -\frac{1}{2}\sum_{i=1}^{l}(\boldsymbol{x}-\boldsymbol{\mu})^T\left(\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}\right)^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\end{aligned}$$

- Once again, we take the derivative w.r.t. the parameter of interest, set it to zero, and solve it. For the mean:

$$\sum_{i=1}^{l}\left(\mathbf{W}\mathbf{W}^T + \sigma^2\mathbf{I}\right)^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = 0$$

$$\Rightarrow \quad \boldsymbol{\mu} = \frac{1}{l}\sum_{i=1}^{l}\boldsymbol{x}_i = \bar{\boldsymbol{x}}$$

Maximum likelihood PCA

- Maximization with respect to $\mathbf{W}$ and σ^2 is more complex, but nonetheless has an **exact closed-form solution** (Tipping & Bishop 1999)

$$\mathbf{W}_{mk} = \mathbf{U}_m \left(\mathbf{L}_m - \sigma^2 \mathbf{I}\right)^{\frac{1}{2}} \mathbf{R}$$

- $\mathbf{U}_m$ is a $d \times m$ matrix whose columns are given by any subset (of size M) of the eigenvectors of the data covariance matrix $\mathbf{S}$
- $\mathbf{L}_m$ is an $m \times m$ diagonal matrix with elements corresponding to the eigenvalues
- $\mathbf{R}$ is an arbitrary $m \times m$ orthogonal matrix
- **Maximum of the likelihood function is obtained when the m eigenvectors are chosen to be those whose eigenvalues are the m largest**

Probabilistic PCA has closed-form solutions

- Probabilistic PCA is the probabilistic generative version of PCA: we can also draw samples from it
- Probabilistic PCA is a form of Gaussian distribution with number of parameters restricted (by the latent space)
- Limitations:
  - Latent space $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\mathbf{0}, \mathbf{I})$ is very simple
  - For generation of $p(\boldsymbol{x}|\boldsymbol{z})$ only **linear transformations** possible