STA 414/2104: Statistical Methods of Machine Learning II

Week 10: Probabilistic PCA/Bayesian Regression

Murat A. Erdoğdu and **Piotr Zwiernik**

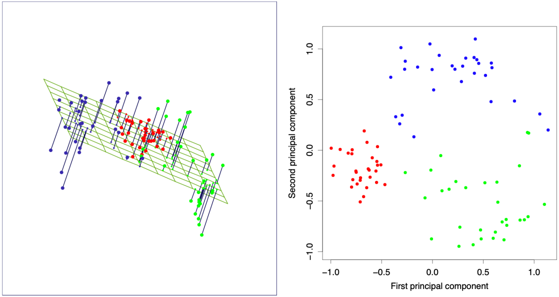
University of Toronto

Overview

- A probabilistic model for continuous latent variables.
  - Probabilistic interpretation of the PCA
- Earlier formulation of PCA was motivated geometrically.
- We will show that it can be expressed as the maximum likelihood estimate of a certain probabilistic model.

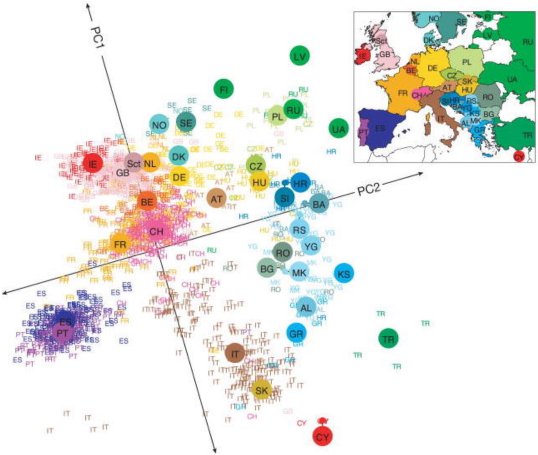
Low dimensional representation

- In practice, even though data is very high dimensional, its important features can be accurately captured in a low dimensional subspace.



- Find a low dimensional representation of your data.
  - Computational benefits
  - Interpretability, visualization
  - Generalization

Nice example



Source: Novembre et al, Genes mirror geography within Europe, Nature, 2009.

Recall: Principal Component Analysis (PCA)

- Data set $\{\mathbf{x}^{(i)}\}_{i=1}^{N}$
- Each input vector $\mathbf{x}^{(i)} \in \mathbb{R}^D$ is approximated as $\overline{\mathbf{x}} + \mathbf{U}\mathbf{z}^{(i)}$,

$$\mathbf{x}^{(i)} \approx \tilde{\mathbf{x}}^{(i)} = \overline{\mathbf{x}} + \mathbf{U}\mathbf{z}^{(i)}$$

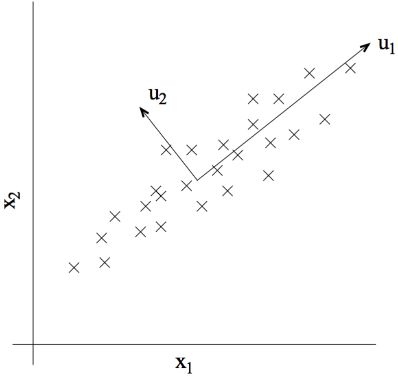
where $\overline{\mathbf{x}} = \frac{1}{n}\sum_i \mathbf{x}^{(i)}$ is the data mean, $\mathbf{U} \in \mathbb{R}^{D\times K}$ is the orthogonal basis for the principal subspace, and $\mathbf{z}^{(i)} \in \mathbb{R}^K$ is the code vector

$$\mathbf{z}^{(i)} = \mathbf{U}^\top(\mathbf{x}^{(i)} - \overline{\mathbf{x}})$$

- $\mathbf{U}$ is chosen to minimize the reconstruction error

$$\mathbf{U}^* = \arg\min_{\mathbf{U}} \sum_i \|\mathbf{x}^{(i)} - \overline{\mathbf{x}} - \mathbf{U}\mathbf{U}^\top(\mathbf{x}^{(i)} - \overline{\mathbf{x}})\|^2$$

We are looking for directions



- For example, in a 2-dimensional problem, we are looking for the direction $\boldsymbol{u}_1$ along which the data is well represented: (?)
  - e.g. direction of higher variance
  - e.g. direction of minimum reconstruction error
  - Recall: they are the same!

Probabilistic PCA

Consider the following latent variable model.

- Similar to the Gaussian mixture model but with Gaussian latents:

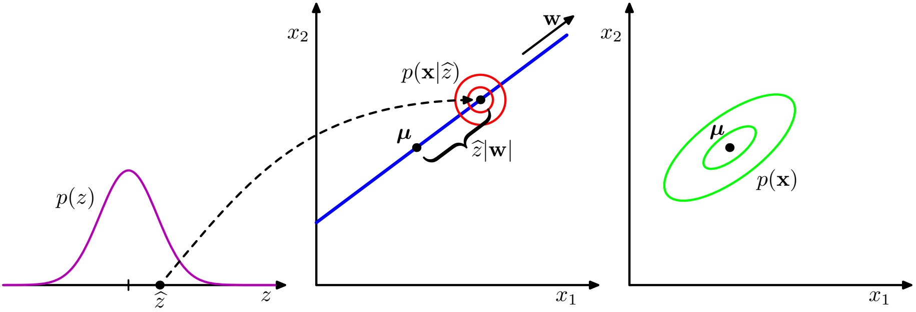
$$\mathbf{z} \sim \mathcal{N}_K(\mathbf{0}, \mathbf{I}_K)$$
$$\mathbf{x} \mid \mathbf{z} \sim \mathcal{N}_D(\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \sigma^2 \mathbf{I}_D)$$

- This is similar to naive Bayes graphical model, because $p(\mathbf{x} \mid \mathbf{z})$ factorizes with respect to the dimensions of $\mathbf{x}$.
- What sort of data does this model produce?

Matrix-vector multiplication: $\mathbf{Wz}$ is a linear combination of the columns of $\mathbf{W}$ with coefficients $\mathbf{z} = (z_1, \dots, z_K)$.

Probabilistic PCA

- $\mathbf{Wz}$ is a random linear combination of the columns of $\mathbf{W}$
- To get the random variable $\mathbf{x}$, we sample a standard normal $\mathbf{z}$ and then add a small amount of isotropic noise to $\mathbf{Wz} + \boldsymbol{\mu}$.



The column span of $\mathbf{W}$ refers to the principal subspace in PCA.

Probabilistic PCA : The Likelihood function

- To perform maximum likelihood in this model, we need to maximize the following:

$$\max_{\mathbf{W},\boldsymbol{\mu},\sigma^2} \log p(\mathbf{x} \,|\, \mathbf{W}, \boldsymbol{\mu}, \sigma^2) = \max_{\mathbf{W},\boldsymbol{\mu},\sigma^2} \log \int p(\mathbf{x} \,|\, \mathbf{z}, \mathbf{W}, \boldsymbol{\mu}, \sigma^2) p(\mathbf{z}) \, d\mathbf{z}$$

- This is easier than for the Gaussian mixture model.
- $\mathbf{x} = \mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \epsilon$ ($\mathbf{x}$ is an affine transformations of Gaussian vars)
- $p(\mathbf{x} \,|\, \mathbf{W}, \boldsymbol{\mu}, \sigma^2)$ is Gaussian
  - Only need to compute $\mathbb{E}[\mathbf{x}]$ and $\mathrm{Cov}[\mathbf{x}]$.

Probabilistic PCA : Maximum Likelihood

$$\mathbb{E}[\mathbf{x}] = \mathbb{E}[\mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \epsilon] = \boldsymbol{\mu}$$

$$\begin{aligned}\mathrm{Cov}[\mathbf{x}] &= \mathbb{E}[(\mathbf{W}\mathbf{z} + \epsilon)(\mathbf{W}\mathbf{z} + \epsilon)^\top] \\ &= \mathbb{E}[(\mathbf{W}\mathbf{z}\mathbf{z}^\top\mathbf{W}^\top] + \mathrm{Cov}[\epsilon] \\ &= \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I}_D\end{aligned}$$

Recall: $\mathbf{R}$ orthogonal if $\mathbf{R}\mathbf{R}^\top = \mathbf{I}$.

This model is not identifiable because $\mathbf{W}\mathbf{W}^\top = (\mathbf{W}\mathbf{R})(\mathbf{W}\mathbf{R})^\top$.

Probabilistic PCA : Maximum Likelihood

Thus, the log-likelihood of the data under this model is given by

$$-\frac{ND}{2}\log(2\pi) - \frac{N}{2}\log|\mathbf{C}| - \frac{1}{2}\sum_{i=1}^{N}(\mathbf{x}^{(i)} - \boldsymbol{\mu})^\top \mathbf{C}^{-1}(\mathbf{x}^{(i)} - \boldsymbol{\mu})$$

where $\mathbf{C} = \mathbf{W}\mathbf{W}^\top + \sigma^2 \mathbf{I}_D$.

Here the MLE $(\widehat{\boldsymbol{\mu}}, \widehat{\mathbf{W}}, \widehat{\sigma}^2)$ is given in a closed-form!

Check Tipping and Bishop (Probabilistic PCA, 1999) for details.

The maximum likelihood estimates

The maximum likelihood estimator is:

$$\hat{\boldsymbol{\mu}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}^{(i)}$$

$$\widehat{\mathbf{W}} = \widehat{\mathbf{U}}(\hat{\mathbf{L}} - \hat{\sigma}^2 \mathbf{I}_K)^{\frac{1}{2}} \mathbf{R}$$

$$\hat{\sigma}^2 = \frac{1}{D - K} \sum_{i=K+1}^{D} \lambda_i$$

- The columns of $\widehat{\mathbf{U}} \in \mathbb{R}^{D \times K}$ are the K unit eigenvectors of the empirical covariance matrix $\widehat{\boldsymbol{\Sigma}}$ that have the largest eigenvalues,
- $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D$ are the eigenvalues of $\widehat{\boldsymbol{\Sigma}}$.
- $\hat{\mathbf{L}} = \mathrm{diag}(\lambda_1, \ldots, \lambda_K)$ is the diagonal matrix whose elements are the corresponding eigenvalues, and $\mathbf{R}$ is any orthogonal matrix.

Probabilistic PCA : Maximum Likelihood

- That seems complex, to get an intuition about how this model behaves when it is fit to data, lets consider the MLE density.
- Recall that the marginal distribution on $\mathbf{x}$ in our fitted model is a Gaussian with mean

$$\widehat{\boldsymbol{\mu}} = \overline{\mathbf{x}}$$

  and covariance

$$\widehat{C} = \widehat{\mathbf{W}}\widehat{\mathbf{W}}^\top + \widehat{\sigma}^2\mathbf{I} = \widehat{\mathbf{U}}(\widehat{\mathbf{L}} - \widehat{\sigma}^2\mathbf{I})\widehat{\mathbf{U}}^\top + \widehat{\sigma}^2\mathbf{I}$$

- The covariance gives us a nice intuition about the model.

Probabilistic PCA : Maximum Likelihood

- Center the data and check the variance along one of the unit eigenvectors $\mathbf{u}_i$, which are the vectors forming the columns of $\widehat{\mathbf{U}}$:

$$\begin{aligned}\mathrm{Var}(\mathbf{u}_i^\top(\mathbf{x}-\overline{\mathbf{x}})) &= \mathbf{u}_i^\top \mathrm{Cov}[\mathbf{x}]\mathbf{u}_i \\ &= \mathbf{u}_i^\top \widehat{\mathbf{U}}(\hat{\mathbf{L}} - \hat{\sigma}^2\mathbf{I})\widehat{\mathbf{U}}^\top \mathbf{u}_i + \hat{\sigma}^2 \\ &= \lambda_i - \hat{\sigma}^2 + \hat{\sigma}^2 = \lambda_i\end{aligned}$$

- Now, center the data and check the variance along any unit vector orthogonal to the subspace spanned by $\widehat{\mathbf{U}}$:

$$\begin{aligned}\mathrm{Var}(\mathbf{u}_i^\top(\mathbf{x}-\overline{\mathbf{x}})) &= \mathbf{u}_i^\top \widehat{\mathbf{U}}(\hat{\mathbf{L}} - \hat{\sigma}^2\mathbf{I})\widehat{\mathbf{U}}^\top \mathbf{u}_i + \hat{\sigma}^2 \\ &= \hat{\sigma}^2\end{aligned}$$

- The model captures the variance along the principle axes and approximates it in all remaining directions with a single variance.

How does it relate to PCA?

- The posterior mean is given by

$$\mathbb{E}[\mathbf{z} \,|\, \mathbf{x}] = \left(\mathbf{W}^\top \mathbf{W} + \sigma^2 \mathbf{I}\right)^{-1} \mathbf{W}^\top (\mathbf{x} - \boldsymbol{\mu})$$

- Posterior variance:

$$\text{Cov}[\mathbf{z}|\mathbf{x}] = \sigma^2 (\mathbf{W}^\top \mathbf{W} + \sigma^2 \mathbf{I})^{-1}$$

- In the limit $\sigma^2 \to 0$, we get

$$\mathbb{E}[\mathbf{z} \,|\, \mathbf{x}] \overset{\sigma^2 \to 0}{\to} \left(\mathbf{W}^\top \mathbf{W}\right)^{-1} \mathbf{W}^\top (\mathbf{x} - \boldsymbol{\mu})$$

- Plugging in the MLEs, this limit recovers the standard PCA.

Why Probabilistic PCA (PPCA)?

- Fitting a full-covariance Gaussian model of data requires $D(D+1)/2 + D$ parameters. With PPCA we model only the K most significant correlations and this only requires $\mathcal{O}(KD)$ parameters as long as K is small.
- Bayesian PCA gives us a Bayesian method for determining the low dimensional principal subspace.
- Existence of likelihood functions allows direct comparison with other probabilistic models.
- Instead of solving directly, we can also use EM. The EM can be scaled to very large high- dimensional datasets.

Summary: Some Gaussian models

- Gaussian mixture model.
  - Gaussian latent variable model $p(\mathbf{x}) = \sum_z p(\mathbf{x}, z)$ used for clustering.
- Probabilistic PCA.
  - Gaussian latent variable model $p(\mathbf{x}) = \int_z p(\mathbf{x}, z)$ used for dimensionality reduction.
- Bayesian linear regression (next hour).
  - Gaussian discriminative model $p(y \mid \mathbf{x})$ used for regression with a Bayesian analysis for the weights.

Overview of the next hour

- Continuing in our theme of probabilistic models for continuous variables.
- We give a probabilistic interpretation of linear regression.
- Chapter 3.3 in Bishop's book.

Completing the Square for Gaussians

Useful technique to find moments of Gaussian random variables.

- It is a multivariate generalization of completing the square.
- The density of $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ satifies:

$$\begin{aligned} \log p(\mathbf{x}) &= -\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) + \text{const} \\ &= -\tfrac{1}{2}\mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{x}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \text{const} \end{aligned}$$

- Thus, if we know $\mathbf{w}$ is Gaussian with *unknown* mean and covariance, and we also know that

$$\log p(\mathbf{w}) = -\tfrac{1}{2}\mathbf{w}^\top \mathbf{A}\mathbf{w} + \mathbf{w}^\top \mathbf{b} + \text{const}$$

  for $\mathbf{A}$ positive definite, then we know that

$$\mathbf{w} \sim \mathcal{N}(\mathbf{A}^{-1}\mathbf{b}, \mathbf{A}^{-1}).$$

Bayesian Linear Regression

- We take the Bayesian approach to linear regression.
  - This is in contrast with the standard regression.
  - By inferring a posterior distribution over the *parameters*, the model can know what it doesn't know.
- How can uncertainty in the predictions help us?
  - Smooth out the predictions by averaging over lots of plausible explanations
  - Assign confidences to predictions
  - Make more robust decisions

Recap: Linear Regression

- Given a training set of inputs and targets $\{(\mathbf{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$
- Linear model:

$$y = \mathbf{w}^\top \boldsymbol{\psi}(\mathbf{x}) + \epsilon$$

- Vectorized, we have the design matrix $\mathbf{X}$ in input space and

$$\boldsymbol{\Psi} = \begin{bmatrix} - & \boldsymbol{\psi}(\mathbf{x}^{(1)}) & - \\ - & \boldsymbol{\psi}(\mathbf{x}^{(2)}) & - \\ & \vdots & \\ - & \boldsymbol{\psi}(\mathbf{x}^{(N)}) & - \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(N)} \end{bmatrix}$$

and predictions

$$\hat{\mathbf{y}} = \boldsymbol{\Psi}\mathbf{w}$$

Recap: Ridge Regression

- Penalized sum of squares (ridge regression):

$$\text{minimize} \quad \frac{1}{2}\|\mathbf{y} - \mathbf{\Psi}\mathbf{w}\|^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

- The gradient: $(\mathbf{\Psi}^\top\mathbf{\Psi} + \lambda\mathbf{I})\mathbf{w} - \mathbf{\Psi}^\top\mathbf{y}$.
- Solution 1: solve analytically by setting the gradient to 0

$$\mathbf{w} = (\mathbf{\Psi}^\top\mathbf{\Psi} + \lambda\mathbf{I})^{-1}\mathbf{\Psi}^\top\mathbf{y}$$

- Solution 2: solve approximately using gradient descent

$$\mathbf{w} \leftarrow (1 - \alpha\lambda)\mathbf{w} - \alpha\mathbf{\Psi}^\top(\mathbf{\Psi}\mathbf{w} - \mathbf{y})$$

Linear Regression as Maximum Likelihood

- We can give linear regression a probabilistic interpretation by assuming a Gaussian noise model:

$$y \mid \mathbf{x} \sim \mathcal{N}(\mathbf{w}^\top \boldsymbol{\psi}(\mathbf{x}),\ \sigma^2)$$

- Linear regression is just maximum log-likelihood under this model:

$$\begin{aligned}
\sum_{i=1}^{N} \log p(y^{(i)} \mid \mathbf{x}^{(i)}; \mathbf{w}, b) &= \sum_{i=1}^{N} \log \mathcal{N}(y^{(i)}; \mathbf{w}^\top \boldsymbol{\psi}(\mathbf{x}^{(i)}), \sigma^2) \\
&= \sum_{i=1}^{N} \log \left[\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \mathbf{w}^\top \boldsymbol{\psi}(\mathbf{x}^{(i)}))^2}{2\sigma^2} \right) \right] \\
&= \text{const} - \frac{1}{2\sigma^2} \sum_{i=1}^{N} (y^{(i)} - \mathbf{w}^\top \boldsymbol{\psi}(\mathbf{x}^{(i)}))^2 \\
&= \text{const} - \frac{1}{2\sigma^2} \|\mathbf{y} - \boldsymbol{\Psi}\mathbf{w}\|^2
\end{aligned}$$

Regularized Linear Regression as MAP Estimation

- View an L_2 regularizer as MAP inference with a Gaussian prior.

$$\arg\max_{\mathbf{w}} \log p(\mathbf{w} \mid \mathcal{D}) = \arg\max_{\mathbf{w}} [\log p(\mathbf{w}) + \log p(\mathcal{D} \mid \mathbf{w})]$$

- We just derived the likelihood term $\log p(\mathcal{D} \mid \mathbf{w})$:

$$\log p(\mathcal{D} \mid \mathbf{w}) = \text{const} - \frac{1}{2\sigma^2} \|\mathbf{y} - \boldsymbol{\Psi}\mathbf{w}\|^2$$

- Assume a Gaussian prior, $\mathbf{w} \sim \mathcal{N}(\mathbf{m}, \mathbf{S})$:

$$\begin{aligned} \log p(\mathbf{w}) &= \log \left[\frac{1}{(2\pi)^{D/2} |\mathbf{S}|^{1/2}} \exp\left(-\tfrac{1}{2} (\mathbf{w} - \mathbf{m})^\top \mathbf{S}^{-1} (\mathbf{w} - \mathbf{m}) \right) \right] \\ &= -\tfrac{1}{2} (\mathbf{w} - \mathbf{m})^\top \mathbf{S}^{-1} (\mathbf{w} - \mathbf{m}) + \text{const} \end{aligned}$$

- Commonly, $\mathbf{m} = \mathbf{0}$ and $\mathbf{S} = \eta \mathbf{I}$, so

$$\log p(\mathbf{w}) = -\frac{1}{2\eta} \|\mathbf{w}\|^2 + \text{const}.$$

This is just L_2 regularization!

Full Bayesian Inference

- Full Bayesian inference makes predictions by averaging over all likely explanations under the posterior distribution.
- Compute posterior using Bayes' Rule:

$$p(\mathbf{w} \mid \mathcal{D}) \propto p(\mathbf{w})p(\mathcal{D} \mid \mathbf{w})$$

- Make predictions using the posterior predictive distribution:

$$p(y \mid \mathbf{x}, \mathcal{D}) = \int p(\mathbf{w} \mid \mathcal{D})\, p(y \mid \mathbf{x}, \mathbf{w})\, \mathrm{d}\mathbf{w}$$

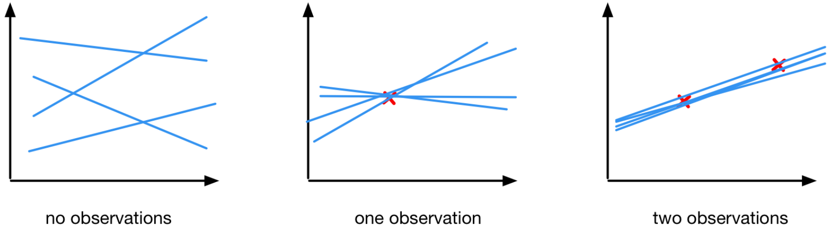
- Doing this lets us quantify our uncertainty.

Bayesian Linear Regression

- **Prior distribution:** $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{S})$
- **Likelihood:** $y \,|\, \mathbf{x}, \mathbf{w} \sim \mathcal{N}(\mathbf{w}^\top \boldsymbol{\psi}(\mathbf{x}),\ \sigma^2)$
- Assuming fixed/known $\mathbf{S}$ and σ^2 is a big assumption. More on this later.

Bayesian Linear Regression

- Bayesian linear regression considers various plausible explanations for how the data were generated.
- It makes predictions using all possible regression weights, weighted by their posterior probability.
- Here are samples from the prior $p(\mathbf{w})$ and posteriors $p(\mathbf{w} \mid \mathcal{D})$



Bayesian Linear Regression: Posterior

- **Deriving the posterior distribution:**

$$
\begin{aligned}
\log p(\mathbf{w} \mid \mathcal{D}) &= \log p(\mathbf{w}) + \log p(\mathcal{D} \mid \mathbf{w}) + \text{const} \\
&= -\tfrac{1}{2}\mathbf{w}^\top \mathbf{S}^{-1}\mathbf{w} - \frac{1}{2\sigma^2}\|\boldsymbol{\Psi}\mathbf{w} - \mathbf{y}\|^2 + \text{const} \\
&= -\tfrac{1}{2}\mathbf{w}^\top \mathbf{S}^{-1}\mathbf{w} - \frac{1}{2\sigma^2}\left(\mathbf{w}^\top \boldsymbol{\Psi}^\top \boldsymbol{\Psi}\mathbf{w} - 2\mathbf{y}^\top \boldsymbol{\Psi}\mathbf{w} + \mathbf{y}^\top \mathbf{y}\right) + \text{const} \\
&= -\tfrac{1}{2}\mathbf{w}^\top \left(\sigma^{-2}\boldsymbol{\Psi}^\top \boldsymbol{\Psi} + \mathbf{S}^{-1}\right)\mathbf{w} + \frac{1}{\sigma^2}\mathbf{y}^\top \boldsymbol{\Psi}\mathbf{w} + \text{const} \quad \text{(complete the square!)}
\end{aligned}
$$

Thus $\mathbf{w} \mid \mathcal{D} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where

$$
\begin{aligned}
\boldsymbol{\mu} &= \left(\boldsymbol{\Psi}^\top \boldsymbol{\Psi} + \sigma^2 \mathbf{S}^{-1}\right)^{-1} \boldsymbol{\Psi}^\top \mathbf{y} \\
\boldsymbol{\Sigma} &= \sigma^2 \left(\boldsymbol{\Psi}^\top \boldsymbol{\Psi} + \sigma^2 \mathbf{S}^{-1}\right)^{-1}
\end{aligned}
$$

Bayesian Linear Regression: Posterior

- Gaussian prior leads to a Gaussian posterior, and so the Gaussian distribution is the conjugate prior for linear regression model.
- Compare $\boldsymbol{\mu}$ to the closed-form solution for linear regression:

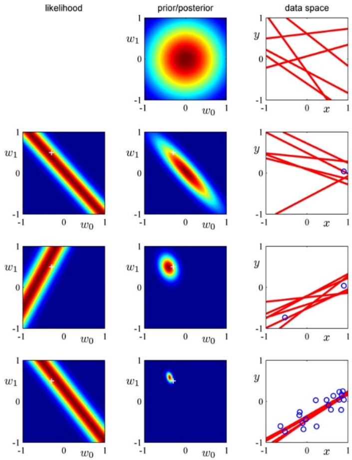
$$\mathbf{w} = \left(\boldsymbol{\Psi}^\top \boldsymbol{\Psi} + \lambda \mathbf{I}\right)^{-1} \boldsymbol{\Psi}^\top \mathbf{y}$$

  This is the mean of the posterior for $\mathbf{S} = \frac{\sigma^2}{\lambda}\mathbf{I}$.

- As $\lambda \to 0$, the standard deviation of the prior goes to ∞, and the mean of the posterior converges to the MLE.

Bayesian Linear Regression

Illustration of sequential Bayesian learning for $y = w_0 + w_1 x$, $w_0 = -0.3$, $w_1 = 0.5$.



Left column:

- Likelihood of a single data point.
- Single point does not identify a line.
- Fix (x, y) then $w_0 = y - w_1 x$.

Middle column:

- Prior/posterior.

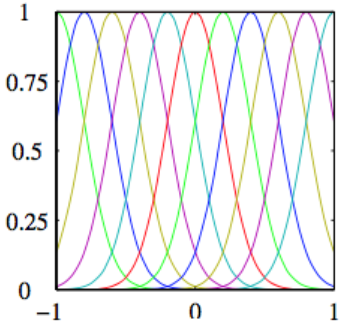
Right column:

- Lines: samples from the posterior.
- Dots: data points.

Radial bases example

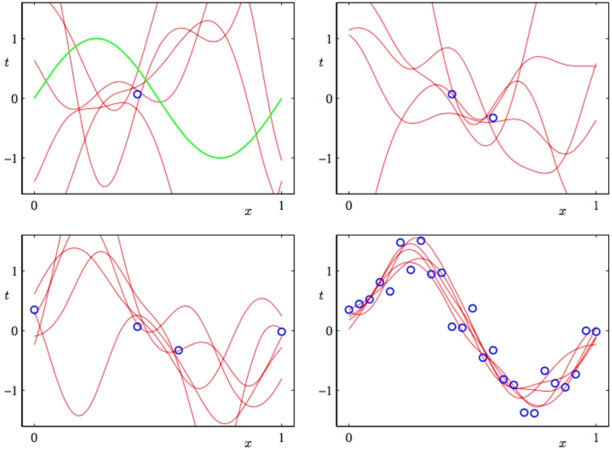
- Example with radial basis function (RBF) features

$$\psi_j(x) = \exp\left(-\frac{(x-\mu_j)^2}{2s^2}\right)$$



Radial bases example

Functions sampled from the posterior:



Posterior predictive distribution

- The posterior just gives us distribution over the parameter space, but if we want to make predictions, the natural choice is to use the posterior predictive distribution.
- Posterior predictive distribution:

$$p(y \mid \mathbf{x}, \mathcal{D}) = \int \underbrace{p(y \mid \mathbf{x}, \mathbf{w})}_{\mathcal{N}(y\,;\,\mathbf{w}^\top \boldsymbol{\psi}(\mathbf{x}),\sigma)} \underbrace{p(\mathbf{w} \mid \mathcal{D})}_{\mathcal{N}(\mathbf{w}\,;\,\boldsymbol{\mu},\boldsymbol{\Sigma})} \,\mathrm{d}\mathbf{w}$$

- Another interpretation: $y = \mathbf{w}^\top \boldsymbol{\psi}(\mathbf{x}) + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma)$ is independent of $\mathbf{w} \mid \mathcal{D} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Recall

$$\boldsymbol{\mu} = \left(\boldsymbol{\Psi}^\top \boldsymbol{\Psi} + \sigma^2 \mathbf{S}^{-1}\right)^{-1} \boldsymbol{\Psi}^\top \mathbf{y}$$

$$\boldsymbol{\Sigma} = \sigma^2 \left(\boldsymbol{\Psi}^\top \boldsymbol{\Psi} + \sigma^2 \mathbf{S}^{-1}\right)^{-1}$$

Bayesian Linear Regression

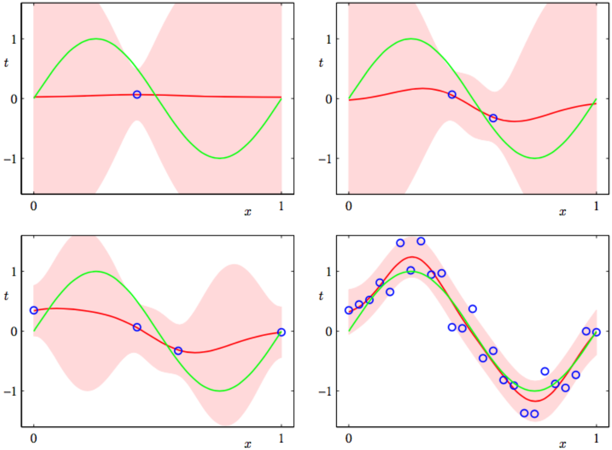
- Another interpretation: $y = \mathbf{w}^\top \boldsymbol{\psi}(\mathbf{x}) + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma)$ is independent of $\mathbf{w} \mid \mathcal{D} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
- Again by the fact that affine transformations of Gaussian vectors are Gaussian, y is a Gaussian distribution with parameters

$$
\begin{aligned}
\mu_{\text{pred}} &= \boldsymbol{\mu}^\top \boldsymbol{\psi}(\mathbf{x}) \\
\sigma^2_{\text{pred}} &= \boldsymbol{\psi}(\mathbf{x})^\top \boldsymbol{\Sigma} \boldsymbol{\psi}(\mathbf{x}) + \sigma^2
\end{aligned}
$$

- Hence, the posterior predictive distribution is $\mathcal{N}(y \mid \mu_{\text{pred}}, \sigma^2_{\text{pred}})$.

Bayesian Linear Regression

Here we visualize confidence intervals based on the posterior predictive mean and variance at each point:



Summary

- This lecture covered the basics of Bayesian regression.

What's remaining:

- Week 11: Neural networks.
- Week 12: Kernel methods, Gaussian processes.
- Week 13: Diffusions.