Lecture 5

Gaussian Models - Part 2

Luigi Freda

ALCOR Lab
DIAG
University of Rome "La Sapienza"

December 20, 2016

Outline

Outline

Intro

- once we are given a Gaussian joint distribution $p(\mathbf{x}_1, \mathbf{x}_2)$, it is useful to be able to compute the marginals $p(\mathbf{x}_1)$ and conditionals $p(\mathbf{x}_1|\mathbf{x}_2)$
- in the following slides we see how to compute these probability densities

Marginals and Conditionals

Theorem 1

(Marginals and conditionals for an MVN)
Suppose $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2) \sim \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, *i.e.* $\mathbf{x}$ *is jointly Gaussian with parameters*

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}, \quad \boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} \end{bmatrix}$$

then the **marginals** *are given by*

$$p(\mathbf{x}_1) = \mathcal{N}(\mathbf{x}_1|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$$
$$p(\mathbf{x}_2) = \mathcal{N}(\mathbf{x}_2|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$$

and the **posterior conditional** *is given by*

$$\begin{aligned} p(\mathbf{x}_1|\mathbf{x}_2) &= \mathcal{N}(\mathbf{x}_1|\boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2}) \\ \boldsymbol{\mu}_{1|2} &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \\ &= \boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\Lambda}_{12}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \\ \boldsymbol{\Sigma}_{1|2} &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} = \boldsymbol{\Lambda}_{11}^{-1} \end{aligned}$$

Marginals and Conditionals

from the previous theorem we have

$$p(\mathbf{x}_1) = \mathcal{N}(\mathbf{x}_1|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$$
$$p(\mathbf{x}_2) = \mathcal{N}(\mathbf{x}_2|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$$
$$p(\mathbf{x}_1|\mathbf{x}_2) = \mathcal{N}(\mathbf{x}_1|\boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2})$$

- the marginal and the conditional distributions are Gaussian
- for the marginals, we just extract the rows and columns corresponding to $\mathbf{x}_1$ and $\mathbf{x}_2$

Marginals and Conditionals

Example with a 2D Gaussian

- consider a 2D example with

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

  where $\rho = \frac{\text{cov}[X_1, X_2]}{\sigma_1\sigma_2}$ is the correlation coefficient

- the marginal $p(x_1)$ is 1D Gaussian, obtained by projecting the joint distribution onto the x_1 line

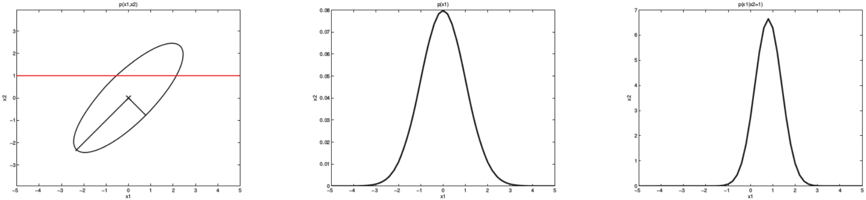
$$p(x_1) = \mathcal{N}(x_1|\mu_1, \sigma_1)$$

Marginals and Conditionals

Example with a 2D Gaussian

- suppose we observe $X_2 = x_2$, the conditional $p(x_1|x_2)$ is obtained by slicing $p(x_1, x_2)$ through the $X_2 = x_2$ line

$$p(x_1|x_2) = \mathcal{N}\left(x_1 \middle| \mu_1 + \frac{\rho\sigma_1\sigma_2}{\sigma_2^2}(x_2 - \mu_2), \sigma_1^2 - \frac{(\rho\sigma_1\sigma_2)^2}{\sigma_2^2}\right)$$

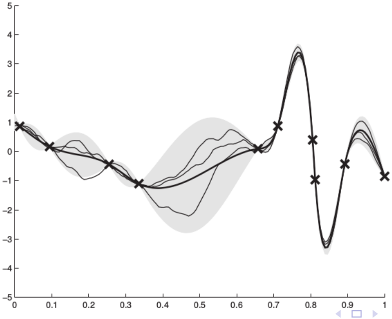


- *left*: joint Gaussian distribution $p(x_1, x_2)$ with a correlation coefficient of 0.8; we plot the 95% contour and the principal axes.
- *center*: the unconditional marginal $p(x_1)$
- *right*: the conditional $p(x_1|x_2) = \mathcal{N}(x_1|0.8, 0.36)$, obtained by slicing $p(x_1, x_2)$ at height $x_2 = 1$

Outline

Interpolation of Noise-free Data

- suppose we want to estimate a 1D function $y = f(t)$, defined on the interval $[0, T]$, starting from N observed points $y_i = f(t_i)$
- we assume for now the data is no **noise-free**
- as a matter of fact, we want to **interpolate the data**, i.e. fit a function that goes exactly though the data
- question: how does the function behave in between observed points?
- the first thing is to assume that the unknown function is **smooth**
- we'll encode the smoothness in a **prior**



Interpolation of Noise-free Data

- in order to encode the prior we start by **discretizing** the problem
- we discretize the interval $[0, T]$ in D equal subintervals such that

$$x_j = f(t_j), \quad t_j = j\Delta, \quad \Delta = \frac{T}{D}, \quad j \in \{1, ..., D\}$$

- we can encode the **smoothness prior** by assuming

$$x_j = \frac{1}{2}(x_{j-1} + x_{j+1}) + \epsilon_j \qquad\qquad j \in \{2, ..., D-1\}$$

  where ϵ_j is a Gaussian noise
- we assume $\boldsymbol{\epsilon} = [\epsilon_2, ..., \epsilon_{D-1}] \sim \mathcal{N}(\mathbf{0}, \frac{1}{\lambda}\mathbf{I})$ where the precision λ controls the smoothness degree
- the above equation can be restated in vector form as

$$\mathbf{L}\mathbf{x} = \boldsymbol{\epsilon}$$

  where

$$\mathbf{L} = \frac{1}{2}\begin{bmatrix} -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & & & \\ & & & -1 & 2 & -1 \end{bmatrix} \in \mathbb{R}^{(D-2)\times D}$$

  is a second order **finite difference matrix**

Interpolation of Noise-free Data

- given a vector $\mathbf{x}$ the degree of smoothness can be represented by the norm $\|\boldsymbol{\epsilon}\|$
- a smoothness prior should give higher probabilities to vectors $\mathbf{x}$ which correspond to smaller $\|\boldsymbol{\epsilon}\|$, hence

$$p(\mathbf{x}) \propto \exp(-\frac{\lambda}{2}\|\mathbf{L}\mathbf{x}\|_2^2)$$

  where a factor λ can be used to weigh the overall smoothness
- the **smoothness prior** can be expressed by using a Gaussian distribution as

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x) = \mathcal{N}(\mathbf{x}|\mathbf{0}, (\lambda\mathbf{L}^T\mathbf{L})^{-1}) \propto \exp(-\frac{\lambda}{2}\|\mathbf{L}\mathbf{x}\|_2^2)$$

Interpolation of Noise-free Data

- smoothness prior

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x) = \mathcal{N}(\mathbf{x}|\mathbf{0}, (\lambda \mathbf{L}^T \mathbf{L})^{-1})$$

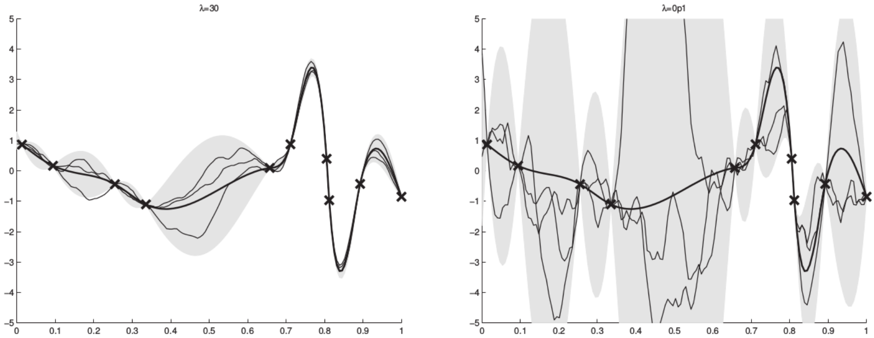
- let's assume that we have used λ to scale $\mathbf{L}$ so that we can ignore it
- note that $\boldsymbol{\Lambda}_x = \mathbf{L}^T \mathbf{L} \in \mathbb{R}^{D \times D}$ and, since $\mathbf{L} \in \mathbb{R}^{(D-2) \times D}$, one has[1] $\text{rank}(\boldsymbol{\Lambda}_x) = D - 2$
- hence $\boldsymbol{\Lambda}_x = \mathbf{L}^T \mathbf{L}$ defines an improper prior known as **intrinsic Gaussian random field**
- however it's possible to show that if we observe $N \geq 2$ points, the posterior will be proper

[1] recall that $\text{rank}(AB) = \min(\text{rank}(A), \text{rank}(B))$

Interpolation of Noise-free Data

- now suppose that in our D discretized intervals we have N noise-free observations gathered in $\mathbf{x}_2 \in \mathbb{R}^N$ and we want to compute the remaining $N - D$ function values $\mathbf{x}_1 \in \mathbb{R}^{D-N}$
- we know that
$$p(\mathbf{x}_1, \mathbf{x}_2) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x) = \mathcal{N}(\mathbf{x}|\mathbf{0}, (\mathbf{L}^T\mathbf{L})^{-1})$$
- we can partition $\mathbf{L} = [\mathbf{L}_1, \mathbf{L}_2]$ where $\mathbf{L}_1 \in \mathbb{R}^{(D-2)\times(D-N)}$ and $\mathbf{L}_2 \in \mathbb{R}^{(D-2)\times N}$
- one has
$$\boldsymbol{\Lambda} = \mathbf{L}^T\mathbf{L} = \begin{bmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{L}_1^T\mathbf{L}_1 & \mathbf{L}_1^T\mathbf{L}_2 \\ \mathbf{L}_2^T\mathbf{L}_1 & \mathbf{L}_2^T\mathbf{L}_2 \end{bmatrix}$$
- by using theorem 1 one has
$$p(\mathbf{x}_1|\mathbf{x}_2) = \mathcal{N}(\mathbf{x}_1|\boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2})$$
$$\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\Lambda}_{12}(\mathbf{x} - \boldsymbol{\mu}_2) = -(\mathbf{L}_1^T\mathbf{L}_1)^{-1}\mathbf{L}_1^T\mathbf{L}_2\mathbf{x}$$

Interpolation of Noise-free Data



- *left*: Gaussian with prior precision $\lambda = 30$
- *right*: prior with $\lambda = 0.01$
- the posterior mean $\boldsymbol{\mu}_{1|2}$ equals the observed data at the specified points and smoothly interpolates in between
- the plots show the 95% pointwise marginals credibility intervals $\mu_j \pm 2\sqrt{\Sigma_{1|2,jj}}$
- N.B.: the variance goes up as we move aways from the the data

Outline

Linear Gaussian System

Problem and Assumptions

problem

- suppose we have two variables $\mathbf{x} \in \mathbb{R}^{D_x}$ and $\mathbf{y} \in \mathbb{R}^{D_y}$
- $\mathbf{y}$ is a noisy observation of $\mathbf{x}$
- $\mathbf{x}$ is an hidden variable we want to estimate

assumptions

- the **prior** is

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$$

- the **likelihood** is

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \boldsymbol{\Sigma}_{y|x})$$

  where $\mathbf{A} \in \mathbb{R}^{D_y \times D_x}$ and $\mathbf{b} \in \mathbb{R}^{D_y}$ are known

N.B.: the above model is equivalent to assume $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon}$ is a noise characterized by the Gaussian distribution $\mathcal{N}(0, \boldsymbol{\Sigma}_{y|x})$

Theorem 2

(Bayes rule for linear Gaussian systems)
Given a linear Gaussian system, as the one described in the previous slide, the **posterior** $p(\mathbf{y}|\mathbf{x})$ *is given by*

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{x|y}, \boldsymbol{\Sigma}_{x|y})$$
$$\boldsymbol{\Sigma}_{x|y}^{-1} = \boldsymbol{\Sigma}_x^{-1} + \mathbf{A}^T \boldsymbol{\Sigma}_y^{-1} \mathbf{A}$$
$$\boldsymbol{\mu}_{x|y} = \boldsymbol{\Sigma}_{x|y}[\mathbf{A}^T \boldsymbol{\Sigma}_y^{-1}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Sigma}_x^{-1} \boldsymbol{\mu}_x]$$

In addition the normalization constant $p(\mathbf{y})$ *is given by*

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu}_x + \mathbf{b}, \boldsymbol{\Sigma}_{y|x} + \mathbf{A}\boldsymbol{\Sigma}_x\mathbf{A}^T)$$

Outline

- suppose we make N **noisy measurements** $y_i \in \mathbb{R}$ of some underlying quantity $x \in \mathbb{R}$, i.e.

$$y_i = x_i + \epsilon_i$$

where $\epsilon_i \sim \mathcal{N}(0, \lambda_y^{-1})$ and $\lambda_y = 1/\sigma^2$

- the **likelihood** is

$$p(y_i|x) = \mathcal{N}(y_i|x, \lambda_y^{-1})$$

- we assume a **Gaussian prior**

$$p(x) = \mathcal{N}(x|\mu_0, \lambda_0^{-1})$$

- given $\mathcal{D} = \{y_1, ..., y_N\}$ we want then to compute the posterior $p(x|\mathcal{D})$ by using a Bayesian approach

- in order to use the theorem 2, we can introduce a variable $\mathbf{y} \triangleq [y_1, ..., y_N]^T \in \mathbb{R}^N$, a matrix $\mathbf{A} = \mathbf{1}_N^T \in \mathbb{R}^{1\times N}$ and $\mathbf{\Sigma}_{y|x} = \lambda_y \mathbf{I}$
- then we get the posterior

$$p(x|\mathbf{y}) = \mathcal{N}(x|\mu_N, \lambda_N^{-1})$$
$$\lambda_N = \lambda_0 + N\lambda_y$$
$$\mu_N = \frac{\lambda_y \sum_i y_i + \lambda_0\mu_0}{\lambda_N} = \frac{N\lambda_y\overline{y} + \lambda_0\mu_0}{N\lambda_y + \lambda_0} = \frac{N\lambda_y}{N\lambda_y + \lambda_0}\overline{y} + \frac{\lambda_0}{N\lambda_y + \lambda_0}\mu_0$$

where $\overline{y} \triangleq \frac{1}{N}\sum_i y_i$

- in this case the MLE estimate of x is exactly $x_{MLE} = \overline{y}$ since

$$x_{MLE} = \underset{\mathbf{x}}{\text{argmax}}\ p(\mathcal{D}|\theta) = \underset{\mathbf{x}}{\text{argmax}} \prod_i p(y_i|x) = \underset{\mathbf{x}}{\text{argmax}} \prod_i \mathcal{N}(y_i|x, \lambda_y^{-1}) = \overline{y}$$

- the posterior mean μ_N is a convex combination of the MLE $\overline{y}$ and the prior mean μ_0

Inferring an Unknown Scalar from Noisy Measurements

- posterior

$$p(x|\mathbf{y}) = \mathcal{N}(x|\mu_N, \lambda_N^{-1})$$
$$\lambda_N = \lambda_0 + N\lambda_y$$
$$\mu_N = \frac{N\lambda_y}{N\lambda_y + \lambda_0}\overline{y} + \frac{\lambda_0}{N\lambda_y + \lambda_0}\mu_0$$

- note that the posterior mean is written in terms of $N\lambda_y\overline{y}$
- having N measurements each of precision λ_y is equivalent to having one measurement $\overline{y}$ with a precision $N\lambda_y$, this means

$$p(x|\mathbf{y}, \lambda_y) = p(x|\overline{y}, N, \lambda_y)$$

in other words $(\overline{y}, N, \lambda_y)$ is a sufficient statistics for the problem

- the procedure can be easily used for an online estimation
- let $\Sigma_0 \triangleq \lambda_0^{-1}$, $\Sigma_{y|x} \triangleq \lambda_y^{-1}$ and $\Sigma_i \triangleq \lambda_i^{-1}$,
- if we have just a measurement, i.e. $N = 1$, one has

$$p(x|y) = \mathcal{N}(x|\mu_1, \Sigma_1)$$

$$\Sigma_1 = \left(\frac{1}{\Sigma_0} + \frac{1}{\Sigma_{y|x}}\right)^{-1} = \frac{\Sigma_0 \Sigma_{y|}}{\Sigma_0 + \Sigma_{y|x}}$$

$$\mu_1 = \Sigma_1 \left(\frac{\mu_0}{\Sigma_0} + \frac{y}{\Sigma_{y|x}}\right) = \mu_0 \frac{\Sigma_0}{\Sigma_0 + \Sigma_{y|x}} + y \frac{\Sigma_{y|x}}{\Sigma_0 + \Sigma_{y|x}}$$

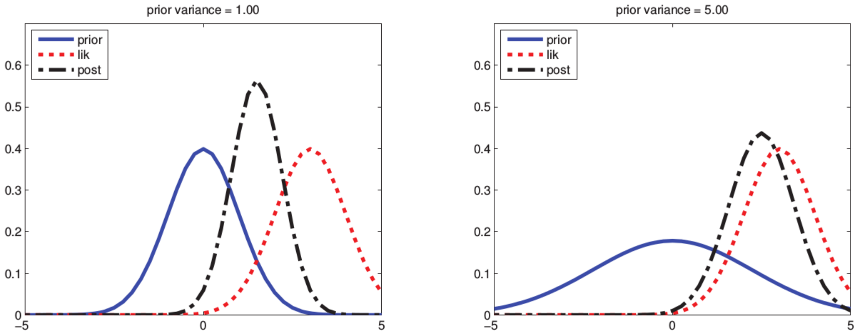
where the posterior μ_1 can be rewritten as

$$\mu_1 = \mu_0 + (y - \mu_0) \frac{\Sigma_0}{\Sigma_0 + \Sigma_{y|x}}$$

$$\mu_1 = y - (y - \mu_0) \frac{\Sigma_{y|x}}{\Sigma_0 + \Sigma_{y|x}}$$

- the third equation is called **shrinkage**: the data is adjusted towards the prior mean

Inferring an Unknown Scalar from Noisy Measurements



Outline

Inferring an Unknown Vector from Noisy Measurements

Problem

- suppose we make N **noisy measurements** $\mathbf{y}_i \in \mathbb{R}^D$ of some vector $\mathbf{x} \in \mathbb{R}^D$, i.e.

$$\mathbf{y}_i = \mathbf{x}_i + \boldsymbol{\epsilon}_i$$

  where $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{y|x})$

- the **likelihood** is

$$p(\mathbf{y}_i|\mathbf{x}) = \mathcal{N}(\mathbf{y}_i|\mathbf{x}, \boldsymbol{\Sigma}_{y|x})$$

  where $\mathbf{A} = \mathbf{I}$ and $\mathbf{b} = \mathbf{0}$

- we assume a **Gaussian prior**

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$$

- given $\mathcal{D} = \{\mathbf{y}_1, ..., \mathbf{y}_N\}$ we want then to compute the posterior $p(\mathbf{x}|\mathcal{D})$ by using a Bayesian approach

- in order to use the theorem 2, we can introduce a variable $\tilde{\mathbf{y}} \triangleq [\mathbf{y}_1, ..., \mathbf{y}_N] \in \mathbb{R}^N$, a matrix

$$\tilde{\mathbf{A}} \triangleq \begin{bmatrix} \mathbf{A} \\ \vdots \\ \mathbf{A} \end{bmatrix} = \begin{bmatrix} \mathbf{I} \\ \vdots \\ \mathbf{I} \end{bmatrix}$$

  and $\mathbf{\Sigma}_{\tilde{y}|x} = \mathrm{diag}(\mathbf{\Sigma}_{y|x})$
- then we get the posterior

$$p(\mathbf{x}|\tilde{\mathbf{y}}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_N, \mathbf{\Sigma}_N)$$
$$\mathbf{\Sigma}_N^{-1} = \mathbf{\Sigma}_0^{-1} + N\mathbf{\Sigma}_{y|x}^{-1}$$
$$\mu_N = \mathbf{\Sigma}_N(\mathbf{\Sigma}_{y|x}^{-1}(N\overline{\mathbf{y}}) + \mathbf{\Sigma}_0^{-1}\boldsymbol{\mu}_0)$$

  where $\overline{\mathbf{y}} \triangleq \frac{1}{N}\sum_i \mathbf{y}_i$
- in this case the MLE estimate of $\mathbf{x}$ is exactly $\mathbf{x}_{MLE} = \overline{\mathbf{y}}$
- the expression of the posterior mean $\boldsymbol{\mu}_N$ is very similar to the scalar case

Outline

- assume we have N noisy observations $y_i \in \mathbb{R}$
- each y_i corresponds to a distinct linear combination of a vector $\mathbf{x} \in \mathbb{R}^D$
- for each y_i we have a noise $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
- we can model this setup as a linear Gaussian system

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \boldsymbol{\epsilon}$$

where $\mathbf{y} = [y_1, ..., y_N]^T \in \mathbb{R}^N$, $\boldsymbol{\epsilon} = [\epsilon_1, ..., \epsilon_N]^T \in \mathbb{R}^N$, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_y)$ and $\boldsymbol{\Sigma}_y = \sigma^2 \mathbf{I}$

- the matrix $\mathbf{A} \in \mathbb{R}^{N \times D}$ is known and can be used for selecting out certain components, for instance if $N = 2$ and $D = 4$

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

- we again assume a smoothness prior

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x) = \mathcal{N}(\mathbf{x}|\mathbf{0}, (\lambda \mathbf{L}^T \mathbf{L})^{-1})$$

where $\boldsymbol{\Lambda}_x = \mathbf{L}^T \mathbf{L}$ defines an improper prior known as **intrinsic Gaussian random field**

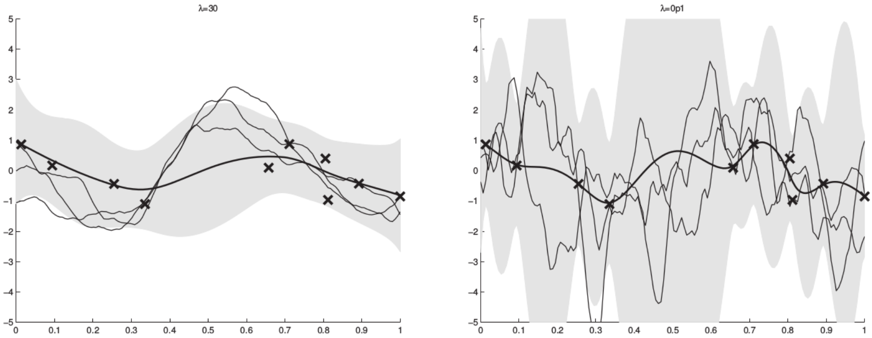
- linear Gaussian system

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \boldsymbol{\epsilon}$$

- smoothness prior

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x) = \mathcal{N}(\mathbf{x}|\mathbf{0}, (\lambda \mathbf{L}^T \mathbf{L})^{-1})$$

- we can apply theorem 2 in order to compute the posterior $p(\mathbf{y}|\mathbf{x})$



- *left*: interpolation by using $\lambda = 30$
- strong prior(large λ) $\Longrightarrow$ smooth estimate and low uncertainty
- *right*: interpolation by using $\lambda = 0.01$
- weak prior(small λ) $\Longrightarrow$ wiggly estimate and high uncertainty
- N.B.: the precision λ affects the posterior mean as well as the posterior variance

- a MAP solution can be found by maximizing the posterior, i.e.

$$\hat{\mathbf{x}}_{MAP} = \underset{\mathbf{x}}{\text{argmax}} \ \log p(\mathbf{x}|\mathbf{y}) = \underset{\mathbf{x}}{\text{argmax}} \left[\log p(\mathbf{y}|\mathbf{x}) + \log p(\mathbf{x}) \right]$$

- in the case $\mathbf{A} = \mathbf{I}$, we can equivalently solve the following optimization problem

$$\hat{\mathbf{x}}_{MAP} = \underset{\mathbf{x}}{\text{argmin}} \ \frac{1}{2\sigma^2} \sum_{i=1}^{N} (x_i - y_i)^2 + \frac{\lambda}{2} \sum_{i=1}^{D} \left[(x_j - x_{j-1})^2 + (x_j - x_{j+1})^2 \right]$$

  where we define $x_0 = x_1$ and $x_{D+1} = x_D$ for simplicity of notation
- the previous equation is a discrete approximation to the following problem

$$\underset{f}{\text{argmin}} \ \frac{1}{2\sigma^2} \int (f(t) - y(t))^2 dt + \frac{\lambda}{2} \int f'(t) dt$$

  where $f'(t)$ is the first time derivative of the function f
- the first term measures the fit to the data and the second term penalizes function that are too wiggly (**Tikhonov regularization** problem)

Credits

- Kevin Murphy's book