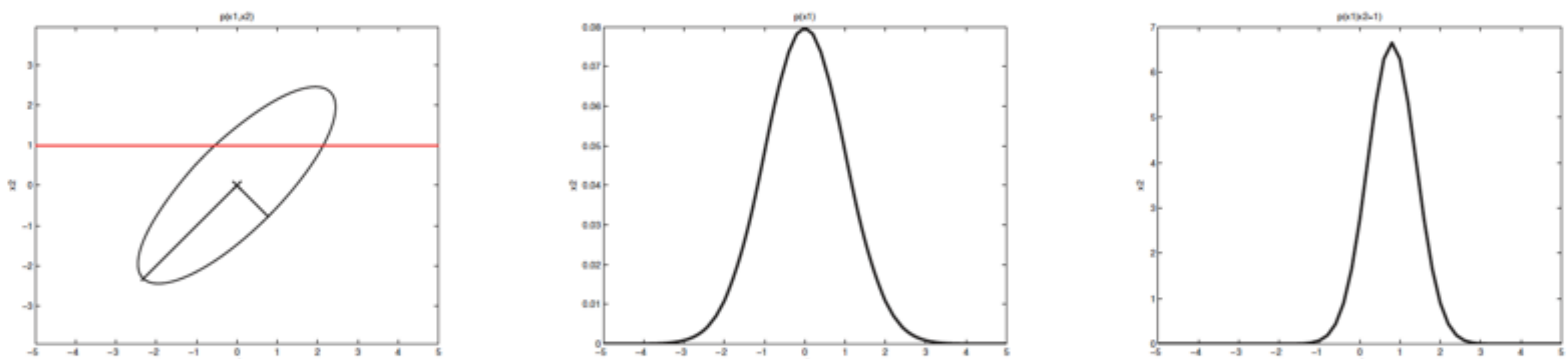
Gaussian models

Sep 2022

Murphy chap4

Inference in jointly Gaussian distributions

Given a joint distribution, $p(\mathbf{x}_1, \mathbf{x}_2)$, it is useful to be able to compute marginals $p(\mathbf{x}_1)$ and conditionals $p(\mathbf{x}_1|\mathbf{x}_2)$. We discuss how to do this below, and then give some applications. These



(a) (b) (c)

Inference in jointly Gaussian distributions

Theorem 4.3.1 (Marginals and conditionals of an MVN)**.** *Suppose* $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$ *is jointly Gaussian with parameters*

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}, \quad \boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} \end{pmatrix} \tag{4.67}$$

Then the marginals are given by

$$\begin{aligned} p(\mathbf{x}_1) &= \mathcal{N}(\mathbf{x}_1 | \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) \\ p(\mathbf{x}_2) &= \mathcal{N}(\mathbf{x}_2 | \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}) \end{aligned} \tag{4.68}$$

and the posterior conditional is given by

$$\boxed{\begin{aligned} p(\mathbf{x}_1 | \mathbf{x}_2) &= \mathcal{N}(\mathbf{x}_1 | \boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2}) \\ \boldsymbol{\mu}_{1|2} &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \\ &= \boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\Lambda}_{12}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \\ &= \boldsymbol{\Sigma}_{1|2}\left(\boldsymbol{\Lambda}_{11}\boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{12}(\mathbf{x}_2 - \boldsymbol{\mu}_2)\right) \\ \boldsymbol{\Sigma}_{1|2} &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} = \boldsymbol{\Lambda}_{11}^{-1} \end{aligned}} \tag{4.69}$$

Inference in jointly Gaussian distributions

Marginals and conditionals of a 2d Gaussian

Let us consider a 2d example. The covariance matrix is

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \tag{4.70}$$

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

$$p(x_1) = \mathcal{N}(x_1|\mu_1, \sigma_1^2) \tag{4.71}$$

Suppose we observe $X_2 = x_2$; the conditional $p(x_1|x_2)$ is obtained by "slicing" the joint distribution through the $X_2 = x_2$ line (see Figure 4.9):

$$p(x_1|x_2) = \mathcal{N}\left(x_1|\mu_1 + \frac{\rho\sigma_1\sigma_2}{\sigma_2^2}(x_2 - \mu_2),\ \sigma_1^2 - \frac{(\rho\sigma_1\sigma_2)^2}{\sigma_2^2}\right) \tag{4.72}$$

If $\sigma_1 = \sigma_2 = \sigma$, we get

$$p(x_1|x_2) = \mathcal{N}\left(x_1|\mu_1 + \rho(x_2 - \mu_2),\ \sigma^2(1 - \rho^2)\right) \tag{4.73}$$

Inference in jointly Gaussian distributions

Marginals and conditionals of a 2d Gaussian

Let us consider a 2d example. The covariance matrix is

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \tag{4.70}$$

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

$$p(x_1) = \mathcal{N}(x_1|\mu_1, \sigma_1^2) \tag{4.71}$$

Suppose we observe $X_2 = x_2$; the conditional $p(x_1|x_2)$ is obtained by "slicing" the joint distribution through the $X_2 = x_2$ line (see Figure 4.9):

$$p(x_1|x_2) \quad = \quad \mathcal{N}\left(x_1|\mu_1 + \frac{\rho\sigma_1\sigma_2}{\sigma_2^2}(x_2 - \mu_2),\ \sigma_1^2 - \frac{(\rho\sigma_1\sigma_2)^2}{\sigma_2^2}\right) \tag{4.72}$$

If $\sigma_1 = \sigma_2 = \sigma$, we get

$$p(x_1|x_2) \quad = \quad \mathcal{N}\left(x_1|\mu_1 + \rho(x_2 - \mu_2),\ \sigma^2(1 - \rho^2)\right) \tag{4.73}$$

Inference in jointly Gaussian distributions

Suppose we want to estimate a 1d function, defined on the interval $[0, T]$, such that $y_i = f(t_i)$ for N observed points t_i. We assume for now that the data is noise-free, so we want to **interpolate** it, that is, fit a function that goes exactly through the data. (See Section 4.4.2.3 for the noisy data case.) The question is: how does the function behave in between the observed

We start by discretizing the problem. First we divide the support of the function into D equal subintervals. We then define

$$x_j = f(s_j), \quad s_j = jh, \quad h = \frac{T}{D}, \; 1 \leq j \leq D \tag{4.74}$$

We can encode our smoothness prior by assuming that x_j is an average of its neighbors, x_{j-1} and x_{j+1}, plus some Gaussian noise:

$$x_j = \frac{1}{2}(x_{j-1} + x_{j+1}) + \epsilon_j, \quad 2 \leq j \leq D-2 \tag{4.75}$$

Inference in jointly Gaussian distributions

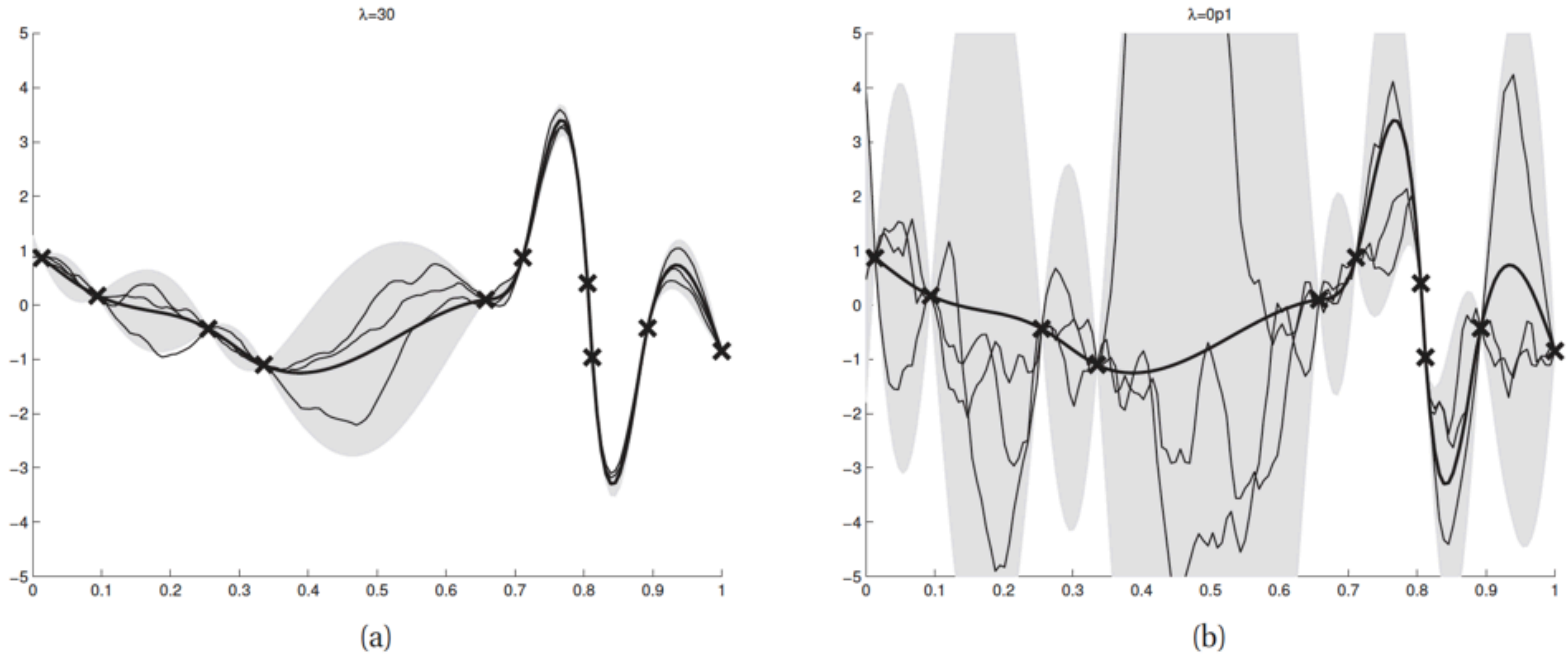


Figure 4.10 Interpolating noise-free data using a Gaussian with prior precision λ. (a) $\lambda = 30$. (b) $\lambda = 0.01$. See also Figure 4.15. Based on Figure 7.1 of (Calvetti and Somersalo 2007). Figure generated by `gaussInterpDemo`.

Inference in jointly Gaussian distributions

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, (1/\lambda)\mathbf{I})$. The precision term λ controls how much we think the function will vary: a large λ corresponds to a belief that the function is very smooth, a small λ corresponds to a belief that the function is quite "wiggly". In vector form, the above equation can be written as follows:

$$\mathbf{L}\mathbf{x} = \boldsymbol{\epsilon} \tag{4.76}$$

where $\mathbf{L}$ is the $(D-2) \times D$ second order **finite difference matrix**

$$\mathbf{L} = \frac{1}{2}\begin{pmatrix} -1 & 2 & -1 & & & & \\ & -1 & 2 & -1 & & & \\ & & & \ddots & & & \\ & & & & -1 & 2 & -1 \end{pmatrix} \tag{4.77}$$

The corresponding prior has the form

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{0}, (\lambda^2\mathbf{L}^T\mathbf{L})^{-1}) \propto \exp\left(-\frac{\lambda^2}{2}||\mathbf{L}\mathbf{x}||_2^2\right) \tag{4.78}$$

We will henceforth assume we have scaled $\mathbf{L}$ by λ so we can ignore the λ term, and just write $\boldsymbol{\Lambda} = \mathbf{L}^T\mathbf{L}$ for the precision matrix.

Inference in jointly Gaussian distributions

Now let $\mathbf{x}_2$ be the N noise-free observations of the function, and $\mathbf{x}_1$ be the $D-N$ unknown function values. Without loss of generality, assume that the unknown variables are ordered first, then the known variables. Then we can partition the $\mathbf{L}$ matrix as follows:

$$\mathbf{L} = [\mathbf{L}_1,\ \mathbf{L}_2],\ \mathbf{L}_1 \in \mathbb{R}^{(D-2)\times(D-N)},\ \mathbf{L}_2 \in \mathbb{R}^{(D-2)\times(N)} \tag{4.79}$$

We can also partition the precision matrix of the joint distribution:

$$\boldsymbol{\Lambda} = \mathbf{L}^T\mathbf{L} = \begin{pmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{L}_1^T\mathbf{L}_1 & \mathbf{L}_1^T\mathbf{L}_2 \\ \mathbf{L}_2^T\mathbf{L}_1 & \mathbf{L}_2^T\mathbf{L}_2 \end{pmatrix} \tag{4.80}$$

Using Equation 4.69, we can write the conditional distribution as follows:

$$p(\mathbf{x}_1|\mathbf{x}_2) = \mathcal{N}(\boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2}) \tag{4.81}$$

$$\boldsymbol{\mu}_{1|2} = -\boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\Lambda}_{12}\mathbf{x}_2 = -\mathbf{L}_1^T\mathbf{L}_2\mathbf{x}_2 \tag{4.82}$$

$$\Sigma_{1|2} = \boldsymbol{\Lambda}_{11}^{-1} \tag{4.83}$$

Inference in jointly Gaussian distributions

It is also interesting to plot the 95% **pointwise marginal credibility intervals**, $\mu_j \pm 2\sqrt{\Sigma_{1|2,jj}}$, shown in grey. We see that the variance goes up as we move away from the data. We also see that the variance goes up as we decrease the precision of the prior, λ. In-

$$
\begin{aligned}
\Pr(\mu - 1\sigma \le X \le \mu + 1\sigma) &\approx 68.27\% \\
\Pr(\mu - 2\sigma \le X \le \mu + 2\sigma) &\approx 95.45\% \\
\Pr(\mu - 3\sigma \le X \le \mu + 3\sigma) &\approx 99.73\%
\end{aligned}
$$

Linear Gaussian system

Suppose we have two variables, $\mathbf{x}$ and $\mathbf{y}$. Let $\mathbf{x} \in \mathbb{R}^{D_x}$ be a hidden variable, and $\mathbf{y} \in \mathbb{R}^{D_y}$ be a noisy observation of $\mathbf{x}$. Let us assume we have the following prior and likelihood:

$$\boxed{\begin{aligned} p(\mathbf{x}) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x) \\ p(\mathbf{y}|\mathbf{x}) &= \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \boldsymbol{\Sigma}_y) \end{aligned}} \tag{4.124}$$

where $\mathbf{A}$ is a matrix of size $D_y \times D_x$. This is an example of a **linear Gaussian system**. We can represent this schematically as $\mathbf{x} \rightarrow \mathbf{y}$, meaning $\mathbf{x}$ generates $\mathbf{y}$. In this section, we show how to "invert the arrow", that is, how to infer $\mathbf{x}$ from $\mathbf{y}$. We state the result below, then give

Linear Gaussian system

Theorem 4.4.1 (Bayes rule for linear Gaussian systems)**.** *Given a linear Gaussian system, as in Equation 4.124, the posterior $p(\mathbf{x}|\mathbf{y})$ is given by the following:*

$$\boxed{\begin{aligned} p(\mathbf{x}|\mathbf{y}) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{x|y}, \boldsymbol{\Sigma}_{x|y}) \\ \boldsymbol{\Sigma}_{x|y}^{-1} &= \boldsymbol{\Sigma}_x^{-1} + \mathbf{A}^T\boldsymbol{\Sigma}_y^{-1}\mathbf{A} \\ \boldsymbol{\mu}_{x|y} &= \boldsymbol{\Sigma}_{x|y}[\mathbf{A}^T\boldsymbol{\Sigma}_y^{-1}(\mathbf{y}-\mathbf{b}) + \boldsymbol{\Sigma}_x^{-1}\boldsymbol{\mu}_x] \end{aligned}} \tag{4.125}$$

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

$$\boxed{p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu}_x + \mathbf{b}, \boldsymbol{\Sigma}_y + \mathbf{A}\boldsymbol{\Sigma}_x\mathbf{A}^T)} \tag{4.126}$$

Proof: derive joint distribution and then use results from last section

Linear Gaussian system (proof)

In more detail, we proceed as follows. The log of the joint distribution is as follows (dropping irrelevant constants):

$$\log p(\mathbf{x}, \mathbf{y}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_x)^T \boldsymbol{\Sigma}_x^{-1}(\mathbf{x} - \boldsymbol{\mu}_x) - \frac{1}{2}(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b})^T \boldsymbol{\Sigma}_y^{-1}(\mathbf{y} - \mathbf{A}\mathbf{x} - \mathbf{b}) \quad (4.150)$$

This is clearly a joint Gaussian distribution, since it is the exponential of a quadratic form.

Expanding out the quadratic terms involving $\mathbf{x}$ and $\mathbf{y}$, and ignoring linear and constant terms, we have

$$\begin{aligned}
Q &= -\frac{1}{2}\mathbf{x}^T \boldsymbol{\Sigma}_x^{-1}\mathbf{x} - \frac{1}{2}\mathbf{y}^T \boldsymbol{\Sigma}_y^{-1}\mathbf{y} - \frac{1}{2}(\mathbf{A}\mathbf{x})^T \boldsymbol{\Sigma}_y^{-1}(\mathbf{A}\mathbf{x}) + \mathbf{y}^T \boldsymbol{\Sigma}_y^{-1}\mathbf{A}\mathbf{x} && (4.151)\\
&= -\frac{1}{2}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}^T \begin{pmatrix}\boldsymbol{\Sigma}_x^{-1} + \mathbf{A}^T\boldsymbol{\Sigma}_y^{-1}\mathbf{A} & -\mathbf{A}^T\boldsymbol{\Sigma}_y^{-1}\\ -\boldsymbol{\Sigma}_y^{-1}\mathbf{A} & \boldsymbol{\Sigma}_y^{-1}\end{pmatrix}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix} && (4.152)\\
&= -\frac{1}{2}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix}^T \boldsymbol{\Sigma}^{-1}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix} && (4.153)
\end{aligned}$$

where the precision matrix of the joint is defined as

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix}\boldsymbol{\Sigma}_x^{-1} + \mathbf{A}^T\boldsymbol{\Sigma}_y^{-1}\mathbf{A} & -\mathbf{A}^T\boldsymbol{\Sigma}_y^{-1}\\ -\boldsymbol{\Sigma}_y^{-1}\mathbf{A} & \boldsymbol{\Sigma}_y^{-1}\end{pmatrix} \triangleq \boldsymbol{\Lambda} = \begin{pmatrix}\boldsymbol{\Lambda}_{xx} & \boldsymbol{\Lambda}_{xy}\\ \boldsymbol{\Lambda}_{yx} & \boldsymbol{\Lambda}_{yy}\end{pmatrix} \quad (4.154)$$

Linear Gaussian system (proof)

where the precision matrix of the joint is defined as

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \boldsymbol{\Sigma}_x^{-1} + \mathbf{A}^T\boldsymbol{\Sigma}_y^{-1}\mathbf{A} & -\mathbf{A}^T\boldsymbol{\Sigma}_y^{-1} \\ -\boldsymbol{\Sigma}_y^{-1}\mathbf{A} & \boldsymbol{\Sigma}_y^{-1} \end{pmatrix} \triangleq \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{xx} & \boldsymbol{\Lambda}_{xy} \\ \boldsymbol{\Lambda}_{yx} & \boldsymbol{\Lambda}_{yy} \end{pmatrix} \tag{4.154}$$

From Equation 4.69, and using the fact that $\boldsymbol{\mu}_y = \mathbf{A}\boldsymbol{\mu}_x + \mathbf{b}$, we have

$$\begin{aligned} p(\mathbf{x}|\mathbf{y}) &= \mathcal{N}(\boldsymbol{\mu}_{x|y}, \boldsymbol{\Sigma}_{x|y}) && (4.155) \\ \boldsymbol{\Sigma}_{x|y} &= \boldsymbol{\Lambda}_{xx}^{-1} = (\boldsymbol{\Sigma}_x^{-1} + \mathbf{A}^T\boldsymbol{\Sigma}_y^{-1}\mathbf{A})^{-1} && (4.156) \\ \boldsymbol{\mu}_{x|y} &= \boldsymbol{\Sigma}_{x|y}\left(\boldsymbol{\Lambda}_{xx}\boldsymbol{\mu}_x - \boldsymbol{\Lambda}_{xy}(\mathbf{y} - \boldsymbol{\mu}_y)\right) && (4.157) \\ &= \boldsymbol{\Sigma}_{x|y}\left(\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\mu} + \mathbf{A}^T\boldsymbol{\Sigma}_y^{-1}(\mathbf{y} - \mathbf{b})\right) && (4.158) \end{aligned}$$

$$\begin{aligned} p(\mathbf{x}_1|\mathbf{x}_2) &= \mathcal{N}(\mathbf{x}_1|\boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2}) \\ \boldsymbol{\mu}_{1|2} &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \\ &= \boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\Lambda}_{12}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \\ &= \boldsymbol{\Sigma}_{1|2}\left(\boldsymbol{\Lambda}_{11}\boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{12}(\mathbf{x}_2 - \boldsymbol{\mu}_2)\right) \\ \boldsymbol{\Sigma}_{1|2} &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} = \boldsymbol{\Lambda}_{11}^{-1} \end{aligned} \tag{4.69}$$

Linear Gaussian system (example)

Inferring an unknown scalar from noisy measurements

Suppose we make N noisy measurements y_i of some underlying quantity x; let us assume the measurement noise has fixed precision $\lambda_y = 1/\sigma^2$, so the likelihood is

$$p(y_i|x) = \mathcal{N}(y_i|x, \lambda_y^{-1}) \tag{4.127}$$

Now let us use a Gaussian prior for the value of the unknown source:

$$p(x) = \mathcal{N}(x|\mu_0, \lambda_0^{-1}) \tag{4.128}$$

We want to compute $p(x|y_1, \ldots, y_N, \sigma^2)$. We can convert this to a form that lets us apply Bayes rule for Gaussians by defining $\mathbf{y} = (y_1, \ldots, y_N)$, $\mathbf{A} = \mathbf{1}_N^T$ (an $1 \times N$ row vector of 1's), and $\mathbf{\Sigma}_y^{-1} = \text{diag}(\lambda_y \mathbf{I})$. Then we get

$$p(x|\mathbf{y}) = \mathcal{N}(x|\mu_N, \lambda_N^{-1}) \tag{4.129}$$

$$\lambda_N = \lambda_0 + N\lambda_y \tag{4.130}$$

$$\mu_N = \frac{N\lambda_y \overline{y} + \lambda_0 \mu_0}{\lambda_N} = \frac{N\lambda_y}{N\lambda_y + \lambda_0}\overline{y} + \frac{\lambda_0}{N\lambda_y + \lambda_0}\mu_0 \tag{4.131}$$

Linear Gaussian system (example)

$$p(x) = \mathcal{N}(x \mid \mu_0, \lambda_0^{-1})$$
$$p(y|x) = \mathcal{N}(y \mid Ax, \Sigma_y)$$
$$y = (y_1, ..., y_N)^T, A = (1, ..., 1)^T, \Sigma_y = diag(\lambda_y^{-1}\mathbf{I})$$

Apply Bayesian rule

$$p(x|y) = \mathcal{N}(y \mid \mu_{x|y}, \Sigma_{x|y})$$
$$\Sigma_{x|y} = (\Sigma_x^{-1} + A^T\Sigma_y^{-1}A)^{-1}$$
$$\mu_{x|y} = \Sigma_{x|y}[A^T\Sigma_y^{-1}y + \Sigma_x^{-1}\mu_x]$$

$$\Sigma_{x|y} = (\lambda_0 + N\lambda_y)^{-1}$$
$$\mu_{x|y} = \frac{N\lambda_y\bar{y} + \lambda_0\mu_0}{\lambda_0 + N\lambda_y}$$
$$= \frac{N\lambda_y}{\lambda_0 + N\lambda_y}\bar{y} + \frac{\lambda_0}{\lambda_0 + N\lambda_y}\mu_0$$

$$\lim_{N\to\infty}\Sigma_{x|y} = 0, \ \lim_{N\to\infty}\mu_{x|y} = \bar{y}$$

Linear Gaussian system (example)

$$
\begin{aligned}
\mu_{x|y} &= \frac{N\lambda_y}{\lambda_0 + N\lambda_y}\bar{y} + \frac{\lambda_0}{\lambda_0 + N\lambda_y}\mu_0 \\
&= \frac{1}{\frac{\lambda_0}{N\lambda_y} + 1}\bar{y} + \frac{1}{1 + \frac{N\lambda_y}{\lambda_0}}\mu_0 \\
&= \frac{1}{\frac{\sigma_y}{N\sigma_0} + 1}\bar{y} + \frac{1}{1 + \frac{N\sigma_0}{\sigma_y}}\mu_0
\end{aligned}
$$

When sigma_o is large, then the MLE estimate has more weights, posterior is close to MLE

Linear Gaussian system (example)

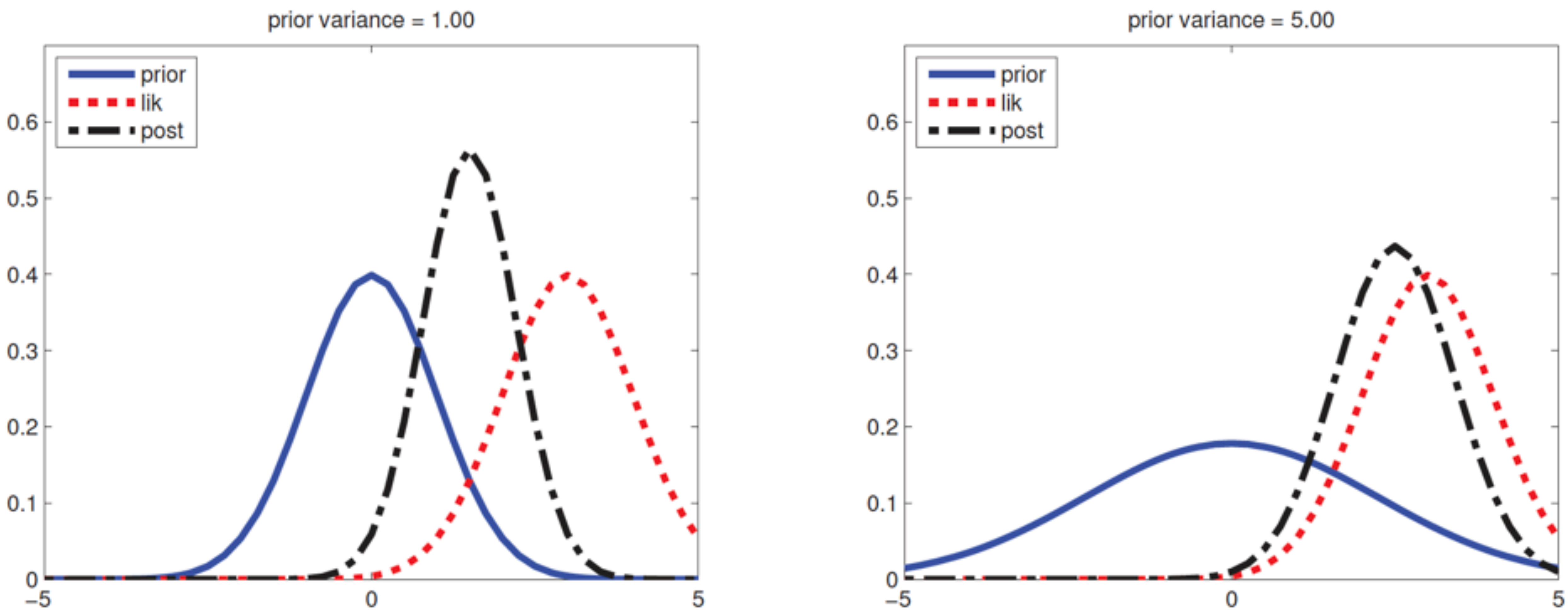


Figure 4.12 Inference about x given a noisy observation $y = 3$. (a) Strong prior $\mathcal{N}(0, 1)$. The posterior mean is "shrunk" towards the prior mean, which is 0. (a) Weak prior $\mathcal{N}(0, 5)$. The posterior mean is similar to the MLE. Figure generated by `gaussInferParamsMean1d`.

Red: Gaussian with MLE mean and variance

Linear Gaussian system (example)

Inferring an unknown vector from noisy measurements

Now consider N vector-valued observations, $\mathbf{y}_i \sim \mathcal{N}(\mathbf{x}, \mathbf{\Sigma}_y)$, and a Gaussian prior, $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_0, \mathbf{\Sigma}_0)$. Setting $\mathbf{A} = \mathbf{I}$, $\mathbf{b} = \mathbf{0}$, and using $\overline{\mathbf{y}}$ for the effective observation with precision $N\mathbf{\Sigma}_y^{-1}$, we have

$$p(\mathbf{x}|\mathbf{y}_1, \ldots, \mathbf{y}_N) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_N, \mathbf{\Sigma}_N) \tag{4.142}$$

$$\mathbf{\Sigma}_N^{-1} = \mathbf{\Sigma}_0^{-1} + N\mathbf{\Sigma}_y^{-1} \tag{4.143}$$

$$\boldsymbol{\mu}_N = \mathbf{\Sigma}_N(\mathbf{\Sigma}_y^{-1}(N\overline{\mathbf{y}}) + \mathbf{\Sigma}_0^{-1}\boldsymbol{\mu}_0) \tag{4.144}$$

As before, mu_N is a combination of prior mean and MLE

Linear Gaussian system (example)

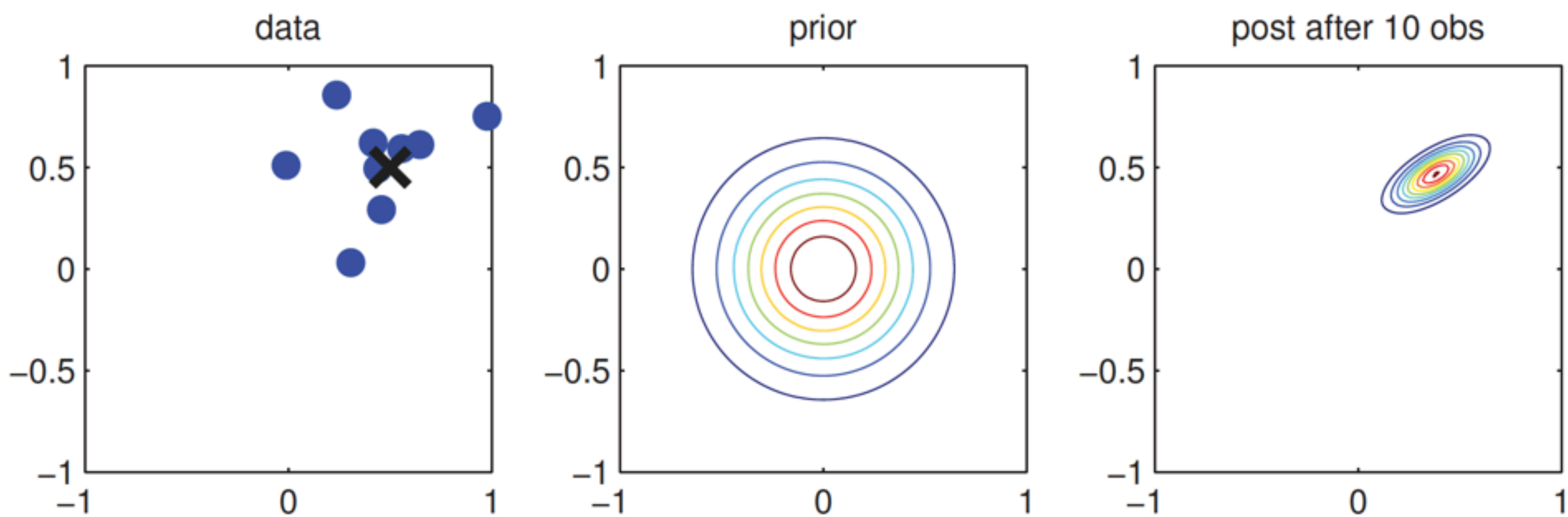


Figure 4.13 Illustration of Bayesian inference for the mean of a 2d Gaussian. (a) The data is generated from $\mathbf{y}_i \sim \mathcal{N}(\mathbf{x}, \mathbf{\Sigma}_y)$, where $\mathbf{x} = [0.5, 0.5]^T$ and $\mathbf{\Sigma}_y = 0.1[2, 1; 1, 1]$). We assume the sensor noise covariance $\mathbf{\Sigma}_y$ is known but $\mathbf{x}$ is unknown. The black cross represents $\mathbf{x}$. (b) The prior is $p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mathbf{0}, 0.1\mathbf{I}_2)$. (c) We show the posterior after 10 data points have been observed. Figure generated by `gaussInferParamsMean2d`.

sensor covariance is known

Linear Gaussian system (example)

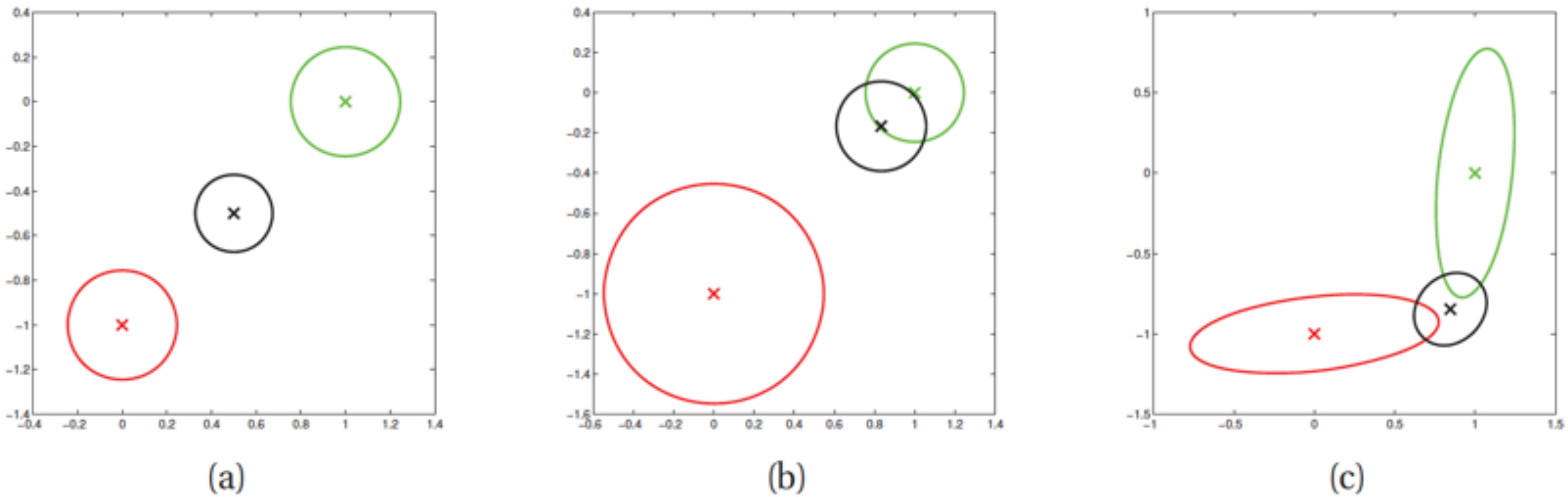


Figure 4.14 We observe $\mathbf{y}_1 = (0, -1)$ (red cross) and $\mathbf{y}_2 = (1, 0)$ (green cross) and infer $E(\boldsymbol{\mu}|\mathbf{y}_1, \mathbf{y}_2, \boldsymbol{\theta})$ (black cross). (a) Equally reliable sensors, so the posterior mean estimate is in between the two circles. (b) Sensor 2 is more reliable, so the estimate shifts more towards the green circle. (c) Sensor 1 is more reliable in the vertical direction, Sensor 2 is more reliable in the horizontal direction. The estimate is an appropriate combination of the two measurements. Figure generated by `sensorFusion2d`.

sensor covariance is known

Inference about MVN

Inferring the parameters of an MVN

So far, we have discussed inference in a Gaussian assuming the parameters $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are known. We now discuss how to infer the parameters themselves. We will assume the data has

the form $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for $i = 1 : N$ and is fully observed, so we have no missing data (see Section 11.6.1 for how to estimate parameters of an MVN in the presence of missing values). To simplify the presentation, we derive the posterior in three parts: first we compute $p(\boldsymbol{\mu}|\mathcal{D}, \boldsymbol{\Sigma})$; then we compute $p(\boldsymbol{\Sigma}|\mathcal{D}, \boldsymbol{\mu})$; finally we compute the joint $p(\boldsymbol{\mu}, \boldsymbol{\Sigma}|\mathcal{D})$.

Reading

Inference about MVN

Posterior distribution of $\boldsymbol{\mu}$

We have discussed how to compute the MLE for $\boldsymbol{\mu}$; we now discuss how to compute its posterior, which is useful for modeling our uncertainty about its value.

The likelihood has the form

$$p(\mathcal{D}|\boldsymbol{\mu}) = \mathcal{N}(\overline{\mathbf{x}}|\boldsymbol{\mu}, \frac{1}{N}\boldsymbol{\Sigma}) \tag{4.171}$$

For simplicity, we will use a conjugate prior, which in this case is a Gaussian. In particular, if $p(\boldsymbol{\mu}) = \mathcal{N}(\boldsymbol{\mu}|\mathbf{m}_0, \mathbf{V}_0)$ then we can derive a Gaussian posterior for $\boldsymbol{\mu}$ based on the results in Section 4.4.2.2. We get

$$p(\boldsymbol{\mu}|\mathcal{D}, \boldsymbol{\Sigma}) = \mathcal{N}(\boldsymbol{\mu}|\mathbf{m}_N, \mathbf{V}_N) \tag{4.172}$$

$$\mathbf{V}_N^{-1} = \mathbf{V}_0^{-1} + N\boldsymbol{\Sigma}^{-1} \tag{4.173}$$

$$\mathbf{m}_N = \mathbf{V}_N(\boldsymbol{\Sigma}^{-1}(N\overline{\mathbf{x}}) + \mathbf{V}_0^{-1}\mathbf{m}_0) \tag{4.174}$$

Same as linear Gaussian system