Linear Gaussian Systems & The Exponential Family

Zou Lexiao

Harbin Institute of Technology, Shenzhen

October 2022

Table of Contents

Linear Gaussian Systems

- Motivation:
  In example "Imputing missing values", we inferred the posterior over the hidden part under the condition of noise-free observations. **Linear Gaussian Systems** extends the approach to handle noisy observations.
- Example:
  Inferring an unknown scalar/vector by N noisy measurement, with the assumption that prior of the unknown source and the likelihood is Gaussian.

Problem Restatement

Let $\boldsymbol{z} \in \mathbb{R}^L$ be an unknown vector of values, and $\boldsymbol{y} \in \mathbb{R}^D$ be some noisy measurement of $\boldsymbol{z}$. We assume these variables are related by the following joint distribution:

$$p(\boldsymbol{z}) = \mathcal{N}\left(\boldsymbol{z} \mid \boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z\right)$$

$$p(\boldsymbol{y} \mid \boldsymbol{z}) = \mathcal{N}\left(\boldsymbol{y} \mid \mathbf{W}\boldsymbol{z} + \boldsymbol{b}, \boldsymbol{\Sigma}_y\right)$$

where $\mathbf{W}$ is a matrix of size $D \times L$

task compute the posterior $p(\boldsymbol{z}|\boldsymbol{y})$

Derivation

The log of the joint distribution is as follows (dropping irrelevant constants):

$$\log p(\boldsymbol{z}, \boldsymbol{y}) = -\frac{1}{2}(\boldsymbol{z} - \boldsymbol{\mu}_z)^T \boldsymbol{\Sigma}_z^{-1} (\boldsymbol{z} - \boldsymbol{\mu}_z) - \frac{1}{2}(\boldsymbol{y} - \mathbf{W}\boldsymbol{z} - \boldsymbol{b})^T \boldsymbol{\Sigma}_y^{-1} (\boldsymbol{y} - \mathbf{W}\boldsymbol{z} - \boldsymbol{b})$$

Since it is the exponential of quadratic form, this is a joint Gaussian distribution.

$$\mathcal{N}(\boldsymbol{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \triangleq \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{y} - \boldsymbol{\mu})\right]$$

Derivation

Furthermore, to compute the posterior of $p(z|y)$, we need the parameter of the joint distribution according to the conditionals of an MVN.

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

$$
\begin{aligned}
Q &= -\frac{1}{2}\boldsymbol{z}^T\boldsymbol{\Sigma}_z^{-1}\boldsymbol{z} - \frac{1}{2}\boldsymbol{y}^T\boldsymbol{\Sigma}_y^{-1}\boldsymbol{y} - \frac{1}{2}(\mathbf{W}\boldsymbol{z})^T\boldsymbol{\Sigma}_y^{-1}(\mathbf{W}\boldsymbol{z}) + \boldsymbol{y}^T\boldsymbol{\Sigma}_y^{-1}\mathbf{W}\boldsymbol{z} \\
&= -\frac{1}{2}\begin{pmatrix} \boldsymbol{z} \\ \boldsymbol{y} \end{pmatrix}^T \begin{pmatrix} \boldsymbol{\Sigma}_z^{-1} + \mathbf{W}^T\boldsymbol{\Sigma}_y^{-1}\mathbf{W} & -\mathbf{W}^T\boldsymbol{\Sigma}_y^{-1} \\ -\boldsymbol{\Sigma}_y^{-1}\mathbf{W} & \boldsymbol{\Sigma}_y^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{z} \\ \boldsymbol{y} \end{pmatrix} \\
&= -\frac{1}{2}\begin{pmatrix} \boldsymbol{z} \\ \boldsymbol{y} \end{pmatrix}^T \boldsymbol{\Sigma}^{-1} \begin{pmatrix} \boldsymbol{z} \\ \boldsymbol{y} \end{pmatrix}
\end{aligned}
$$

Derivation

According to the conditionals of an MVN

$$p\left(\boldsymbol{y}_1 \mid \boldsymbol{y}_2\right)=\mathcal{N}\left(\boldsymbol{y}_1 \mid \boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2}\right)$$
$$\boldsymbol{\mu}_{1|2}=\boldsymbol{\Sigma}_{1|2}\left(\boldsymbol{\Lambda}_{11}\boldsymbol{\mu}_1-\boldsymbol{\Lambda}_{12}\left(\boldsymbol{y}_2-\boldsymbol{\mu}_2\right)\right)$$
$$\boldsymbol{\Sigma}_{1|2}=\boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}=\boldsymbol{\Lambda}_{11}^{-1}$$

we get the **Bayes rule for Gaussians**

Bayes rule for Gaussians

$$p(\boldsymbol{z} \mid \boldsymbol{y})=\mathcal{N}\left(\boldsymbol{z} \mid \boldsymbol{\mu}_{z|y}, \boldsymbol{\Sigma}_{z|y}\right)$$
$$\boldsymbol{\Sigma}_{z|y}^{-1}=\boldsymbol{\Sigma}_z^{-1}+\mathbf{W}^{\top}\boldsymbol{\Sigma}_y^{-1}\mathbf{W}$$
$$\boldsymbol{\mu}_{z|y}=\boldsymbol{\Sigma}_{z|y}\left[\mathbf{W}^{\top}\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{y}-\boldsymbol{b})+\boldsymbol{\Sigma}_z^{-1}\boldsymbol{\mu}_z\right]$$

Bayes rule for Gaussians

Bayes rule for Gaussians

$$p(\boldsymbol{z} \mid \boldsymbol{y}) = \mathcal{N}\left(\boldsymbol{z} \mid \boldsymbol{\mu}_{z|y}, \boldsymbol{\Sigma}_{z|y}\right)$$

$$\boldsymbol{\Sigma}_{z|y}^{-1} = \boldsymbol{\Sigma}_z^{-1} + \mathbf{W}^\top \boldsymbol{\Sigma}_y^{-1} \mathbf{W}$$

$$\boldsymbol{\mu}_{z|y} = \boldsymbol{\Sigma}_{z|y}\left[\mathbf{W}^\top \boldsymbol{\Sigma}_y^{-1}(\boldsymbol{y} - \boldsymbol{b}) + \boldsymbol{\Sigma}_z^{-1} \boldsymbol{\mu}_z\right]$$

We see that the Gaussian prior $p(\boldsymbol{z})$, combined with the Gaussian likelihood $p(\boldsymbol{y} \mid \boldsymbol{z})$, results in a Gaussian posterior $p(\boldsymbol{z} \mid \boldsymbol{y})$. Thus Gaussians are closed under Bayesian conditioning. To describe this more generally, we say that the Gaussian prior is a **conjugate prior** for the Gaussian likelihood, since the posterior distribution has the same type as the prior.

Example 1: Inferring an unknown scalar

- Background:
  Suppose we make N noisy but independent measurements y_i of some underlying quantity z;
- Assumption:
  - Measurement noise has fixed precision $\lambda_y = 1/\sigma^2$
  - The likelihood and prior are Gaussian.

$$p\left((y_1, \ldots, y_N) \mid z\right) = \mathcal{N}\left(\boldsymbol{y} \mid ((z, \ldots, z)), \operatorname{diag}\left(\sigma^2 \mathbf{I}\right)\right)$$

$$p(z) = \mathcal{N}\left(z \mid \mu_0, \lambda_0^{-1}\right)$$

- Methodology:
  Defining $\mathbf{W} = \mathbf{1}_N$, and $\boldsymbol{\Sigma}_y^{-1} = \operatorname{diag}\left(\lambda_y \mathbf{I}\right)$, apply the Bayes rule of Gaussians

$$\begin{aligned}
p(z \mid \boldsymbol{y}) &= \mathcal{N}\left(z \mid \mu_N, \lambda_N^{-1}\right) \\
\lambda_N &= \lambda_0 + N\lambda_y \\
\mu_N &= \frac{N\lambda_y \bar{y} + \lambda_0 \mu_0}{\lambda_N} = \frac{N\lambda_y}{N\lambda_y + \lambda_0}\bar{y} + \frac{\lambda_0}{N\lambda_y + \lambda_0}\mu_0
\end{aligned}$$

Example 1: Inferring an unknown scalar

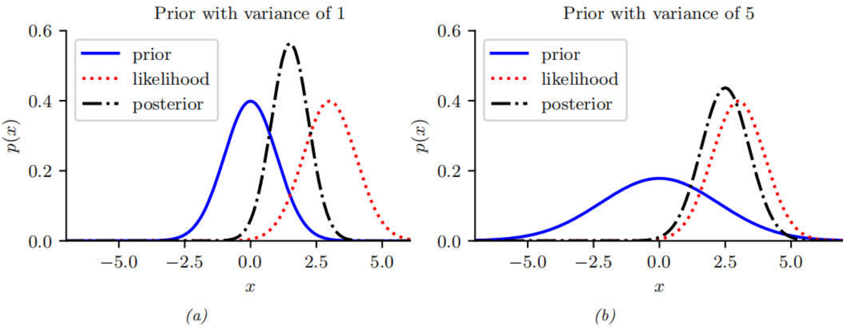

Figure 3.8: Inference about z given a noisy observation $y = 3$. (a) Strong prior $\mathcal{N}(0, 1)$. The posterior mean is "shrunk" towards the prior mean, which is 0. (b) Weak prior $\mathcal{N}(0, 5)$. The posterior mean is similar to the MLE. Generated by gauss_infer_1d.ipynb.

Example 2: Inferring an unknown vector

- Background:
  Suppose we make N noisy but independent measurements $\boldsymbol{y_i}$ of an unknown quantity of interest $\boldsymbol{z}, \boldsymbol{z} \in \mathbb{R}^D$;
- Assumption:
  - Σ_y is given.
  - The likelihood and prior are Gaussian.

$$p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N \mid \boldsymbol{z}) = \prod_{n=1}^{N} \mathcal{N}\left(\boldsymbol{y}_n \mid \boldsymbol{z}, \boldsymbol{\Sigma}_y\right) = \mathcal{N}\left(\overline{\boldsymbol{y}} \mid \boldsymbol{z}, \frac{1}{N}\boldsymbol{\Sigma}_y\right)$$

$$p(\boldsymbol{z}) = \mathcal{N}\left(\boldsymbol{z} \mid \mu_z, \Sigma_z\right)$$

- Methodology:
  Setting $\boldsymbol{W} = \boldsymbol{I}, \boldsymbol{b} = 0$, apply the Bayes rule of Gaussian

$$p\left(\boldsymbol{z} \mid \boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\right) = \mathcal{N}(\boldsymbol{z} \mid \widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\Sigma}})$$

$$\widehat{\boldsymbol{\Sigma}}^{-1} = \boldsymbol{\Sigma}_z^{-1} + N_{\mathcal{D}} \boldsymbol{\Sigma}_y^{-1}$$

$$\widehat{\boldsymbol{\mu}} = \widehat{\boldsymbol{\Sigma}}\left(\boldsymbol{\Sigma}_y^{-1}\left(N_{\mathcal{D}} \overline{\boldsymbol{y}}\right) + \boldsymbol{\Sigma}_z^{-1} \boldsymbol{\mu}_z\right)$$

Example 2: Inferring an unknown vector

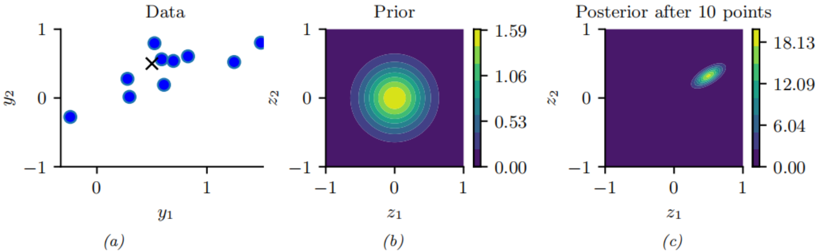

Figure 3.9: Illustration of Bayesian inference for a 2d Gaussian random vector $\boldsymbol{z}$. (a) The data is generated from $\boldsymbol{y}_n \sim \mathcal{N}(\boldsymbol{z}, \boldsymbol{\Sigma}_y)$, where $\boldsymbol{z} = [0.5, 0.5]^{\mathsf{T}}$ and $\boldsymbol{\Sigma}_y = 0.1[2, 1; 1, 1])$. We assume the sensor noise covariance $\boldsymbol{\Sigma}_y$ is known but $\boldsymbol{z}$ is unknown. The black cross represents $\boldsymbol{z}$. (b) The prior is $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\mathbf{0}, 0.1\mathbf{I}_2)$. (c) We show the posterior after 10 data points have been observed. Generated by gauss_infer_2d.ipynb.

Example 3: sensor fusion

Background:
Extending *Example 2*, now we have multiple measurements which comes from different sensors with different reliability(Σ).

$$p(\boldsymbol{z}, \boldsymbol{y}) = p(\boldsymbol{z}) \prod_{m=1}^{M} \prod_{n=1}^{N_m} \mathcal{N}\left(\boldsymbol{y}_{n,m} \mid \boldsymbol{z}, \boldsymbol{\Sigma}_m\right)$$

where M is the number of sensors (measurement devices), and N_m is the number of observations from sensor m, and $\boldsymbol{y} = \boldsymbol{y}_{1:N,1:M} \in \mathbb{R}^K$. Our goal is to combine the evidence together, to compute $p(\boldsymbol{z} \mid \boldsymbol{y})$. This is known as **sensor fusion**.

Example 3: sensor fusion

- Assumption:
  - Each Σ_m is given
  - The likelihood and prior are Gaussian.

$$p(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N \mid \boldsymbol{z}) = \mathcal{N}\left(\boldsymbol{y} \mid (\boldsymbol{z}, \ldots, \boldsymbol{z}), \begin{bmatrix} \sum_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sum_m \end{bmatrix}\right)$$

$$p(\boldsymbol{z}) = \mathcal{N}\left(\boldsymbol{z} \mid \mu_z, \Sigma_z\right)$$

- Methodology: Setting

$\boldsymbol{W} = [\boldsymbol{I}; \ldots; \boldsymbol{I}], \boldsymbol{b} = 0, \boldsymbol{\Sigma} = \begin{bmatrix} \sum_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sum_m \end{bmatrix}$, apply the Bayes rule of Gaussian

Example 3: sensor fusion

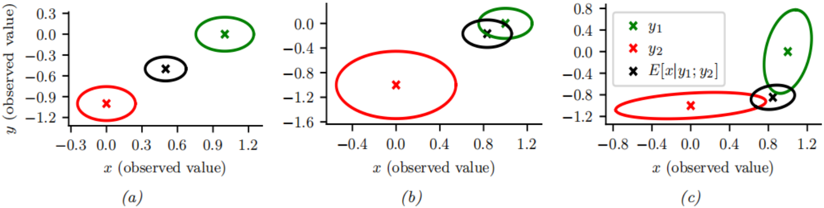

Figure 3.10: We observe $\boldsymbol{y}_1 = (0, -1)$ (red cross) and $\boldsymbol{y}_2 = (1, 0)$ (green cross) and estimate $\mathbb{E}[\boldsymbol{z}|\boldsymbol{y}_1, \boldsymbol{y}_2]$ (black cross). (a) Equally reliable sensors, so the posterior mean estimate is in between the two circles. (b) Sensor 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. Generated by sensor_fusion_2d.ipynb.

Table of Contents

Definition

Consider a family of probability distributions parameterized by $\eta \in \mathbb{R}^K$ with fixed support over $\mathcal{Y}^D \subseteq \mathbb{R}^D$. We say that the distribution $p(\boldsymbol{y} \mid \boldsymbol{\eta})$ is in the exponential family if its density can be written in the following way:

$$p(\boldsymbol{y} \mid \boldsymbol{\eta}) \triangleq \frac{1}{Z(\boldsymbol{\eta})} h(\boldsymbol{y}) \exp\left[\boldsymbol{\eta}^\top \mathcal{T}(\boldsymbol{y})\right] = h(\boldsymbol{y}) \exp\left[\boldsymbol{\eta}^\top \mathcal{T}(\boldsymbol{y}) - A(\boldsymbol{\eta})\right]$$

- $h(\boldsymbol{y})$, **scaling constant** (also known as the **base measure**, often 1)
- $\mathcal{T}(\boldsymbol{y}) \in \mathbb{R}^K$, **sufficient statistics**
- η, **natural parameters** or **canonical parameters**
- $Z(\boldsymbol{\eta})$, **partition function**
- $A(\boldsymbol{\eta}) = \log Z(\boldsymbol{\eta})$, **log partition function**
- An exponential family is **minimal** if there is no $\boldsymbol{\eta} \in \mathbb{R}^K \backslash \{0\}$such that $\boldsymbol{\eta}^\top \mathcal{T}(\boldsymbol{y}) = 0$.

Definition

The former equation can be generalized by defining $\boldsymbol{\eta} = f(\phi)$, where ϕ is some other, possibly smaller, set of parameters. In this case, the distribution has the form

$$p(\boldsymbol{y} \mid \phi) = h(\boldsymbol{y}) \exp\left[f(\phi)^\top \mathcal{T}(\boldsymbol{y}) - A(f(\phi))\right]$$

- If the mapping from ϕ to $\boldsymbol{\eta}$ is nonlinear, we call this a **curved exponential family**.
- If $\boldsymbol{\eta} = f(\phi) = \phi$, the model is said to be in **canonical form**.
- If $\mathcal{T}(\boldsymbol{y}) = \boldsymbol{y}$, we say this is a **natural exponential family** or **NEF**.

Example: Bernoulli distribution

According to chapter ahead, we get

$$\begin{aligned}\mathrm{Ber}(y \mid \mu) &= \mu^y(1-\mu)^{1-y} \\ &= \exp[y\log(\mu) + (1-y)\log(1-\mu)]\end{aligned}$$

where $\mathcal{T}(y) = [\mathbb{I}(y=1), \mathbb{I}(y=0)], \boldsymbol{\eta} = [\log(\mu), \log(1-\mu)]$
Since there is a linear dependence between the features, this is an **over-complete representation**. If the representation is overcomplete, η is not uniquely identifiable. It is common to use a minimal representation, which means there is a unique η associated with the distribution. In this case, we can just define

$$\mathrm{Ber}(y \mid \mu) = \exp\left[y\log\left(\frac{\mu}{1-\mu}\right) + \log(1-\mu)\right]$$

where $\mathcal{T}(y) = y, \boldsymbol{\eta} = \log(\frac{\mu}{1-\mu}), A(\eta) = \log(1-\mu)$

Log partition function is cumulant generating function

The r th moment of a real-valued random variable X with density $f(x)$ is

$$\mu_r = E(X^r) = \int_{-\infty}^{\infty} x^r f(x) dx$$

for integer $r = 0, 1, \ldots$. The value is assumed to be finite. Provided that it has a Taylor expansion about the origin, the moment generating function

$$\begin{aligned} M(\xi) = E\left(e^{\xi X}\right) &= E\left(1 + \xi X + \cdots + \xi^r X^r / r! + \cdots\right) \\ &= \sum_{r=0}^{\infty} \mu_r \xi^r / r! \end{aligned}$$

is an easy way to combine all of the moments into a single expression. The r th moment is the r th derivative of M at the origin.
The cumulants κ_r are the coefficients in the Taylor expansion of the cumulant generating function about the origin

$$K(\xi) = \log M(\xi) = \sum_r \kappa_r \xi^r / r!$$

Log partition function is cumulant generating function

$$\nabla A(\boldsymbol{\eta}) = \mathbb{E}[\mathcal{T}(\boldsymbol{y})]$$
$$\nabla^2 A(\boldsymbol{\eta}) = \mathrm{Cov}[\mathcal{T}(\boldsymbol{y})]$$

- Hessian of $A(\eta)$ is positive definite, which means $A(\eta)$ has a minimum
- $\log p(\boldsymbol{y} \mid \boldsymbol{\eta}) = \boldsymbol{\eta}^\top \mathcal{T}(\boldsymbol{y}) - A(\boldsymbol{\eta}) + \mathrm{const}$ has a unique global maximum(Applied in **MLE**).

Maximum entropy derivation of the exponential family

Suppose we want to find a distribution $p(\boldsymbol{x})$ to describe some data, where all we know are the expected values (F_k) of certain features or functions $f_k(\boldsymbol{x})$:

$$\int d\boldsymbol{x} p(\boldsymbol{x}) f_k(\boldsymbol{x}) = F_k$$

To formalize what we mean by "least number of assumptions", we will search for the distribution that is as close as possible to our prior $q(\boldsymbol{x})$, in the sense of KL divergence (Section 6.2), while satisfying our constraints:

$$p = \underset{p}{\operatorname{argmin}} D_{\mathbb{K}L}(p \| q), \text{subject to constraints}$$

For discrete distributions, the KL divergence is defined as follows:

$$D_{\mathbb{K}L}(p \| q) \triangleq \sum_{k=1}^{K} p_k \log \frac{p_k}{q_k}$$

This naturally extends to continuous distributions as well:

$$D_{\mathbb{K}L}(p \| q) \triangleq \int dx p(x) \log \frac{p(x)}{q(x)}$$

Maximum entropy derivation of the exponential family

Assuming that $\mathbf{x}$ is discrete, we minimize the KL subject to the constraints that $p(\boldsymbol{x}) \geq 0$ and $\sum_{\boldsymbol{x}} p(\boldsymbol{x}) = 1$. The Lagrangian is given by

$$J(p, \boldsymbol{\lambda}) = -\sum_{\boldsymbol{x}} p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} + \lambda_0 \left(1 - \sum_{\boldsymbol{x}} p(\boldsymbol{x})\right) + \sum_k \lambda_k \left(F_k - \sum_{\boldsymbol{x}} p(\boldsymbol{x}) f_k(\boldsymbol{x})\right)$$

Then we have

$$\frac{\partial J}{\partial p_c} = -1 - \log \frac{p(x=c)}{q(x=c)} - \lambda_0 - \sum_k \lambda_k f_k(x=c)$$

Setting $\frac{\partial J}{\partial p_c} = 0$ for each c yields

$$p(\boldsymbol{x}) = \frac{q(\boldsymbol{x})}{Z} \exp\left(-\sum_k \lambda_k f_k(\boldsymbol{x})\right)$$

where we have defined $Z \triangleq e^{1+\lambda_0}$.