Gaussian Process

Marshal Sinaga - marshal.sinaga@aalto.fi

Last update: March 11, 2024

This note aims to cover some materials on the Gaussian process. The primary references are Gaussian Process for Machine Learning by C. E. Rasmussen and CS-E4895 by Arno Solin.

1 Multivariate Normal Distribution

1.1 Linear transformation theorem for the multivariate normal distribution

Let x follow a multivariate normal distribution:

$$x \sim \mathcal{N}(\mu, \Sigma) \tag{1}$$

Then, any affine transformation of x is also multivariate normally distributed:

$$y = Ax + b \sim \mathcal{N}(A\mu + b, A\Sigma A^\top) \tag{2}$$

Proof:

The moment-generating function of random vector x is

$$M_x(t) = \mathbb{E}[\exp(t^T x)] \tag{3}$$

and therefore, the moment-generating function of the random vector y is given by

$$\begin{aligned} M_y(t) &= \mathbb{E}\left[\exp(t^T(Ax + b))\right] \\ &= \mathbb{E}[\exp(t^\top Ax)\exp(t^\top b)] \\ &= \exp(t^\top b)\mathbb{E}[\exp(t^\top Ax)] \\ &= \exp(t^\top b)M_x(A^\top t) \end{aligned} \tag{4}$$

The moment-generating function of the multivariate normal distribution is

$$M_x(t) = \exp(t^\top \mu + \frac{1}{2}t^\top \Sigma t) \tag{5}$$

and therefore, the moment-generating function of random vector y becomes

$$M_y(t) = \exp(t^\top(A\mu + b) + \frac{1}{2}t^\top A\Sigma A^\top t) \tag{6}$$

Since the moment-generating function and the probability density function of a random variable are equivalent, this demonstrates that y follows a multivariate normal distribution with mean $A\mu + b$ and covariance $A\Sigma A^\top$.

1.2 Marginal distribution of the multivariate normal distribution

Let x follow a multivariate normal distribution:

$$x \sim \mathcal{N}(\mu, \Sigma) \tag{7}$$

Then, the marginal distribution of any subset vector x_s is also a multivariate normal distribution

$$x_s \sim \mathcal{N}(\mu_s, \Sigma_s) \tag{8}$$

where μ_s drops the irrelevant variables (the ones not in the subset, i.e., marginalized out) from the mean vector μ and Σ_s drops the corresponding rows and columns from the covariance matrix Σ.

Proof: Define an $m \times n$ subset matrix S such that $s_{ij} = 1$, if the $j-$th element in x_s corresponds to the $i-$th element in x, and $s_{ij} = 0$ otherwise. Then,

$$x_s = Sx \tag{9}$$

and we can apply the linear transformation theorem to give

$$x_s \sim \mathcal{N}(S\mu, S\Sigma S^\top) \tag{10}$$

Finally, we see that $S\mu = \mu_s$ and $S\Sigma S^\top = \Sigma_s$

1.3 Conditional distribution of the multivariate normal distribution

Let x follow a multivariate normal distribution

$$x \sim \mathcal{N}(\mu, \Sigma) \tag{11}$$

Then, the conditional distribution of any subset vector x_1, given the complement vector x_2, is also a multivariate normal distribution

$$\begin{aligned} \mu_{1|2} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2) \\ \Sigma_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \end{aligned} \tag{12}$$

with block-wise mean and covariance defined as:

$$\begin{aligned} \mu &= \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \\ \Sigma &= \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \end{aligned} \tag{13}$$

Proof: Without loss of generality, we assume that in parallel to 13,

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{14}$$

where $x_1 \in \mathbb{R}^{n_1 \times 1}$, $x_2 \in \mathbb{R}^{n_2 \times 1}$, and $x \in \mathbb{R}^{n \times 1}$. The joint distribution of x_1 and x_2 is

$$x \sim \mathcal{N}(\mu, \Sigma) \tag{15}$$

Moreover, the marginal distribution of x_2 follows from 11 and 13 as

$$x_2 \sim \mathcal{N}(\mu_2, \Sigma_{22}) \tag{16}$$

According to conditional probability, it holds that

$$\begin{aligned} p(x_1|x_2) &= \frac{p(x_1, x_2)}{p(x_2)} \\ &= \frac{\mathcal{N}(\mu, \Sigma)}{\mathcal{N}(\mu_2, \Sigma_{22})} \end{aligned} \tag{17}$$

Using the probability density of multivariate-normal, this becomes

$$\begin{aligned} p(x_1|x_2) &= \frac{1/\sqrt{(2\pi)^n |\Sigma|} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right)}{1/\sqrt{(2\pi)^{n_2}|\Sigma_{22}|} \exp\left(-\frac{1}{2}(x-\mu_2)^\top \Sigma_{22}^{-1}(x-\mu_2)\right)} \\ &= 1/\sqrt{(2\pi)^{n-n_2}} \sqrt{\frac{|\Sigma_{22}|}{|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu) + \frac{1}{2}(x-\mu_2)^\top \Sigma_{22}^{-1}(x-\mu_2)\right) \end{aligned} \tag{18}$$

Writing the inverse Σ as

$$\Sigma^{-1} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \tag{19}$$

and applying 13 to 18, we obtain:

$$p(x_1|x_2) = 1/\sqrt{(2\pi)^{n-n_2}} \sqrt{\frac{|\Sigma_{22}|}{|\Sigma|}} \exp(-\frac{1}{2}\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}\right)^\top \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}\right) \tag{20}$$

$$+ \frac{1}{2}(x-\mu_2)^\top \Sigma_{22}^{-1}(x-\mu_2))$$

Multiplying within 20, we have

$$\begin{aligned} p(x_1|x_2) = 1/\sqrt{(2\pi)^{n-n_2}} \sqrt{\frac{|\Sigma_{22}|}{|\Sigma|}} \exp(-\frac{1}{2}((x_1-\mu_1)^\top \Sigma^{11}(x_1-\mu_1) + 2(x_1-\mu_1)\Sigma^{12}(x_2-\mu_2) \\ + (x_2-\mu_2)^\top \Sigma^{22}(x_2-\mu_2)) + \frac{1}{2}(x-\mu_2)^\top \Sigma_{22}^{-1}(x-\mu_2)) \end{aligned} \tag{21}$$

where we have used the fact that $\Sigma^{12} = \Sigma^{21^\top}$, because Σ^{-1} is symmetric. The inverse of a block matrix is

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} (A-BD^{-1}C)^{-1} & -(A-BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A-BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A-BD^{-1}C)^{-1}BD^{-1} \end{bmatrix} \tag{22}$$

Thus, the inverse of Σ^{-1} in 19 is

$$\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1} = \begin{bmatrix} (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} & -(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1} \\ -\Sigma_{22}^{-1}\Sigma_{21}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} & \Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma_{21}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1} \end{bmatrix} \tag{23}$$

Plugging this into 20, we have

$$\begin{aligned} p(x_1|x_2) =& \frac{1}{\sqrt{(2\pi)^{n-n_2}}} \cdot \sqrt{\frac{|\Sigma_{22}|}{|\Sigma|}} \cdot \\ & \exp\Bigg[-\frac{1}{2}\Big((x_1-\mu_1)^{\mathrm{T}}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}(x_1-\mu_1) \; - \\ & \qquad 2(x_1-\mu_1)^{\mathrm{T}}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2) + \\ & \qquad (x_2-\mu_2)^{\mathrm{T}}\left[\Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma_{21}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1}\right](x_2-\mu_2)\Big) \\ & \quad + \frac{1}{2}\Big((x_2-\mu_2)^{\mathrm{T}}\Sigma_{22}^{-1}(x_2-\mu_2)\Big)\Bigg] \; . \end{aligned} \tag{24}$$

Eliminating some terms, we have

$$\begin{aligned} p(x_1|x_2) =& \frac{1}{\sqrt{(2\pi)^{n-n_2}}} \cdot \sqrt{\frac{|\Sigma_{22}|}{|\Sigma|}} \cdot \\ & \exp\Bigg[-\frac{1}{2}\Big((x_1-\mu_1)^{\mathrm{T}}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}(x_1-\mu_1) \; - \\ & \qquad 2(x_1-\mu_1)^{\mathrm{T}}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2) + \\ & \qquad (x_2-\mu_2)^{\mathrm{T}}\Sigma_{22}^{-1}\Sigma_{21}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2)\Big)\Bigg] \; . \end{aligned} \tag{25}$$

Rearranging the terms, we have

$$\begin{aligned} p(x_1|x_2) =& \frac{1}{\sqrt{(2\pi)^{n-n_2}}} \cdot \sqrt{\frac{|\Sigma_{22}|}{|\Sigma|}} \cdot \exp\Bigg[-\frac{1}{2} \cdot \\ & \left[(x_1-\mu_1) - \Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2)\right]^{\mathrm{T}}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\left[(x_1-\mu_1) - \Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2)\right]\Bigg] \\ =& \frac{1}{\sqrt{(2\pi)^{n-n_2}}} \cdot \sqrt{\frac{|\Sigma_{22}|}{|\Sigma|}} \cdot \exp\Bigg[-\frac{1}{2} \cdot \\ & \left[x_1 - \left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2)\right)\right]^{\mathrm{T}}(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}\left[x_1 - \left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2-\mu_2)\right)\right]\Bigg] \end{aligned} \tag{26}$$

where we used the fact that $\Sigma_{21} = \Sigma_{12}^{\top}$. The determinant of a block matrix is

$$\begin{vmatrix} A & B \\ C & D \end{vmatrix} = |D| \cdot |A - BD^{-1}C| \; , \tag{27}$$

such that we have for Σ that

$$\begin{vmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{vmatrix} = |\Sigma_{22}| \cdot |\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}| \tag{28}$$

with this and $n - n_2 = n_1$, we finally arrive at

$$\begin{aligned} p(x_1|x_2) =& \frac{1}{\sqrt{(2\pi)^{n_1}|\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}|}} \cdot \exp\Big[-\frac{1}{2} \cdot \\ & \left[x_1 - \left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)\right)\right]^{\mathrm{T}} (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1} \left[x_1 - \left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)\right)\right]\Big] \end{aligned} \tag{29}$$

which is the pdf of a multivariate normal distribution

$$p(x_1|x_2) = \mathcal{N}(x_1; \mu_{1|2}, \Sigma_{1|2}) \tag{30}$$

with mean $\mu_{1|2}$ and covariance $\Sigma_{1|2}$ given by 12.

2 Kernel Theory

2.1 Hilbert Space

- A vector space $\mathcal{V}$ is a set of closed vectors under addition and scalar multiplication.
- If $\mathcal{V}$ is equipped with a norm $\|.\|_{\mathcal{V}} \in \mathbb{R}$, it is a norm space.
- A Hilbert space $\mathcal{H}$ is a complete inner product space, with inner product $\langle . \rangle_{\mathcal{H}}$ and induced norm $\|x\| = \sqrt{\langle x, x \rangle_{\mathcal{H}}}$.

2.2 Kernel Function and Reproducing Kernel Hilbert Space (RKHS)

- A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel function if and only if there exists a Hilbert space $\mathcal{H}$ and a map $\phi : \mathcal{X} \to \mathcal{H}$ such that:

  $$k(x, y) = \langle \phi(x), \phi(y) \rangle \tag{31}$$

  for all $x, y \in \mathcal{X}$.
- Let $\phi : \mathcal{X} \to \mathbb{R}^{\mathcal{X}}$ and let us define:

  $$k_x \coloneqq \phi(x) = k(x, .) \tag{32}$$

  Therefore, we have $k_x(y) = k(x, y)$.
- Let $\mathcal{G}$ denote a vector space with span based on the images $\{k_x | x \in \mathcal{X}\}$, i.e.,

  $$\{\mathcal{G} \coloneqq \sum_{i=1}^{m} \alpha_i k_{x_i} | \alpha_i \in \mathbb{R}, m \in \mathbb{N}, x_i \in \mathcal{X}\} \tag{33}$$

- By the definition of the kernel function, the inner product on $\mathcal{G}$ is defined as follows:

$$\langle k_x, k_y \rangle := k(x, y) \tag{34}$$

Recall that $k_x = k(x, .)$, hence, $\langle k_x, k_y \rangle = \langle k(x, .), k(y, .) \rangle$.

- Therefore, for any $f, g \in \mathcal{G}$, with $f = \sum_i \alpha_i k_{x_i}$ and $g = \sum_j \beta_j k_{y_j}$, we have:

$$\begin{aligned} \langle f, g \rangle &= \langle \sum_i \alpha_i k_{x_i}, \sum_j \beta_j k_{y_j} \rangle & (35) \\ &= \sum_{ij} \alpha_i \beta_j \langle k_{x_i}, k_{y_j} \rangle & (36) \\ &= \sum_{ij} \alpha_i \beta_j k(x_i, y_j) & (37) \end{aligned}$$

- To make $\mathcal{G}$ a Hilbert space, we need to make it complete, i.e., ensure all Cauchy sequences converge.

- **Definition 1.** *Let $\mathcal{H}$ be a Hilbert space of real function f defined on an index set $\mathcal{X}$. Then $\mathcal{H}$ is called a reproducing kernel Hilbert space endowed with an inner product $\langle ., . \rangle_{\mathcal{H}}$ if there exists a kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with the following properties:*

  - *For every $x \in \mathcal{X}, k_x(y) = k(x, y)$ as function of $y \in \mathcal{X}$ belongs to $\mathcal{H}$, and*
  - *k has the reproducing property.*

- Reproducing property:

$$\begin{aligned} \langle k_x, f \rangle &= \langle k_x, \sum_i \alpha_i k_{x_i} \rangle & (38) \\ &= \sum_i \alpha_i \langle k_x, k_{x_i} \rangle = \sum_i k(x, x_i) = f(x) & (39) \end{aligned}$$

- Moore-Aronszajn theorem: Given a kernel, there is a unique RKHS, Given an RKHS, there is a unique kernel.

2.3 Representer Theorem

Settings:

- We are given kernel k and denote the corresponding RKHS at $\mathcal{H}$.
- We want to learn a linear function $f(\mathbf{x})$ from a finite data set $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$

Theorem 1. *Consider the risk minimization problem of the form:*

$$\min_{f \in \mathcal{H}} \underbrace{R_n(\mathbf{y}, \mathbf{f})}_{\text{Empirical Risk}} + \underbrace{\lambda \Omega(\|f\|_{\mathcal{H}})}_{\text{Regularizer}} \tag{40}$$

where $\mathbf{f} = \{f(\mathbf{x}_1), \cdots, f(\mathbf{x}_n)\}, \mathbf{y} = \{y_1, \cdots, y_n\}$*, and* λ *is a scaling parameter. Then 40 always has an optimal solution of the form:*

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}_i, \mathbf{x}) \tag{41}$$

3 Spectral Kernel

3.1 Fourier Transforms

- Fourier transform $S(\omega)$ of a function $f(x)$,

$$S(\omega) = \int_{-\infty}^{\infty} f(x) \exp(-2\pi i x\omega) dx \tag{42}$$

- Inverse Fourier transform $f(x)$ of a spectral density $S(\omega)$

$$f(x) = \int_{-\infty}^{\infty} S(\omega) \exp(2\pi i x\omega) d\omega \tag{43}$$

- Euler's identity:

$$\exp(ix) = \cos x + i \sin x \tag{44}$$

  Hence

$$\exp(\pm 2\pi i x\omega) = \cos(2\pi x\omega) \pm i \sin(2\pi x\omega) \tag{45}$$

3.2 Fourier Duals

Theorem 2. *Bochner's theorem: Any stationary kernel* $k : \mathbb{R}^D \to \mathbb{R}$ *and its spectral density* $S : \mathbb{R}^D \to \mathbb{R}$ *are Fourier duals*

$$k(x - x') \equiv k(\tau) = \int_{-\infty}^{\infty} S(\omega) \exp(2\pi i x\omega^\top \tau) d\omega$$
$$S(\omega) = \int_{-\infty}^{\infty} k(\tau) \exp(-2\pi i x\omega^\top \tau) d\tau$$

4 Marginal Likelihood via Laplace Approximation

- Marginal likelihood to do model selection:

$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{f})\, p(\mathbf{f})\, d\mathbf{f} \tag{46}$$

- Let $\psi(\mathbf{f}) = \log h(\mathbf{f}) = \log(p(\mathbf{y}|f)p(\mathbf{f}))$

$$\psi(\mathbf{f}) = \log p(\mathbf{y}|\mathbf{f}) - \frac{N}{2} \log 2\pi - \frac{1}{2} \log |\mathbf{K}| - \frac{1}{2} \mathbf{f}^\top \mathbf{K}^{-1} \mathbf{f} \tag{47}$$

- Second order Taylor approximation around the mode $\hat{\mathbf{f}}$

$$\psi(\mathbf{f}) = \psi(\hat{\mathbf{f}}) - \frac{1}{2} (\mathbf{f} - \hat{\mathbf{f}})^\top \mathbf{A} (\mathbf{f} - \hat{\mathbf{f}}) \tag{48}$$

- Substituting back

$$
\begin{aligned}
p(\mathbf{y}) \approx q(\mathbf{y}) &= \int \exp(\psi(\hat{\mathbf{f}}) - \frac{1}{2}(\mathbf{f} - \hat{\mathbf{f}})^\top \mathbf{A}(\mathbf{f} - \hat{\mathbf{f}}))d\mathbf{f} && (49)\\
&= \exp(\psi(\hat{\mathbf{f}})) \int \exp(-\frac{1}{2}(\mathbf{f} - \hat{\mathbf{f}})^\top \mathbf{A}(\mathbf{f} - \hat{\mathbf{f}}))d\mathbf{f} && (50)\\
&= \exp(\psi(\hat{\mathbf{f}}))(2\pi)^{N/2}|\mathbf{A}^{-1}|^{1/2} && (51)\\
&= \exp(\log p(\mathbf{y}|\hat{\mathbf{f}}) - \frac{N}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{K}| - \frac{1}{2}\hat{\mathbf{f}}^\top \mathbf{K}^{-1}\hat{\mathbf{f}})\\
&\quad (2\pi)^{N/2}|\mathbf{A}^{-1}|^{1/2} && (52)
\end{aligned}
$$

- Taking the log of $q(\mathbf{y})$

$$
\begin{aligned}
\log q(\mathbf{y}) &= \log p(\mathbf{y}|\hat{\mathbf{f}}) - \frac{N}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{K}| - \frac{1}{2}\hat{\mathbf{f}}^\top \mathbf{K}^{-1}\hat{\mathbf{f}}\\
&\quad + \frac{N}{2}\log 2\pi + \frac{1}{2}\log|\mathbf{A}|^{-1} && (53)\\
&= \log p(\mathbf{y}|\hat{f}) - \frac{1}{2}\log|\mathbf{K}| - \frac{1}{2}\hat{f}^\top \mathbf{K}^{-1}\hat{\mathbf{f}} + \frac{1}{2}|\mathbf{A}^{-1}| && (54)
\end{aligned}
$$

- We can now use the fact that $|\mathbf{A}^{-1}| = |\mathbf{A}|^{-1}$

$$
\log q(\mathbf{y}) = \log p(\mathbf{y}|\hat{\mathbf{f}}) - \frac{1}{2}\log|\mathbf{K}| - \frac{1}{2}\hat{f}^\top \mathbf{K}^{-1}\hat{\mathbf{f}} - \frac{1}{2}|\mathbf{A}| \qquad (55)
$$

- Recall that $\mathbf{A} = \mathbf{K}^{-1} + \mathbf{W}$

$$
\log q(\mathbf{y}) = \log p(\mathbf{y}|\hat{\mathbf{f}}) - \frac{1}{2}\log|\mathbf{K}| - \frac{1}{2}\hat{f}^\top \mathbf{K}^{-1}\hat{\mathbf{f}} - \frac{1}{2}|\mathbf{K}^{-1} + \mathbf{W}| \qquad (56)
$$

- We optimize $\log q(\mathbf{y})$ using gradient based methods to choose hyperparamters.