Advanced probabilistic methods

Lecture 3: Multivariate Gaussian, Bayesian linear models

Pekka Marttinen

Aalto University

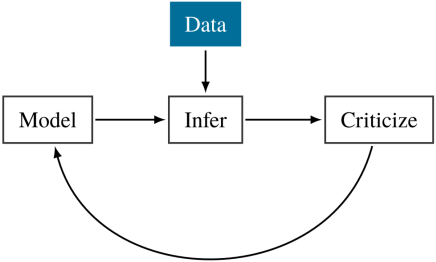
January, 2021

Lecture 3 overview

- Gaussian distribution
  - Bayesian parameter learning
- Multivariate Gaussian distribution
  - Characterization
  - Useful identities
- Bayesian Linear Parameter Models (LPMs)
  - Posterior computation (given fixed hyperparameters)
- Ch. 8 & 18 (until the end of Section 18.1.1) in Barber's book

Recall from lecture 1

- Tools for probabilistic modeling
  - **Models**: Bayesian networks, sparse Bayesian linear regression, Gaussian mixture models, latent linear models
  - **Methods for inference**: maximum likelihood, maximum a posteriori (MAP), analytical, Laplace approximation, expectation maximization (EM), Variational Bayes (VB), stochastic variational inference (SVI)
  - **Ways to select between models**



Box's loop (Blei, 2014)

What is a model?

- A model specifies a probability distribution for a random variable Y, and it is often affected by some parameter θ. The model can be denoted as $p(y|\theta)$.
- Fitting the model (i.e. inference) corresponds to learning the value (or the distribution) of θ, after some data y have been observed.

Prior, Likelihood, Posterior

- *Bayes' rule* tells us how to update our prior beliefs about variable θ in light of the data y to a posterior belief:

$$\underbrace{p(\theta|y)}_{\text{posterior}} = \frac{\overbrace{p(y|\theta)}^{}\overbrace{p(\theta)}^{}}{\underbrace{p(y)}_{\text{evidence}}}.$$

$$\underbrace{p(\theta|y)}_{\text{posterior}} = \frac{\underbrace{p(y|\theta)}_{\text{likelihood}}\,\underbrace{p(\theta)}_{\text{prior}}}{\underbrace{p(y)}_{\text{evidence}}}.$$

  The evidence is also called the marginal likelihood.

- $p(y|\theta)$ is the probability that the model generates the observed data y when using parameter θ
  - $L(\theta) \equiv p(y|\theta)$, with y held fixed, is called the *likelihood*
  - $f(y) \equiv p(y|\theta)$, with θ held fixed, is called the *observation model*

- "*Methods for inference*" = Bayes' rule + some algorithm to do the actual computations (on this course)

Point estimates for parameters

- The *Maximum A Posteriori (MAP)* parameter value, which maximizes the posterior

$$\theta_* = \arg\max_{\theta} p(\theta|y)$$

- The Maximum likelihood assignment (ML)

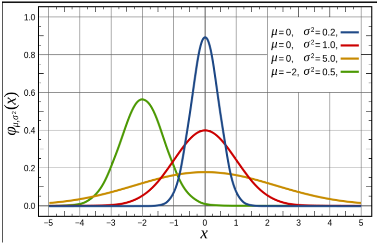
$$\theta_* = \arg\max_{\theta} p(y|\theta)$$

- The full posterior distribution $p(\theta|y)$ tells also of the uncertainty about the value of θ.

Gaussian distribution

- $X \sim N(\mu, \sigma^2)$
- Parameters: μ: mean, σ^2: variance
- Inverse of the variance, $\lambda = 1/\sigma^2$, is called the precision
- Standard deviation σ
- 95% credible interval equals approximately $[\mu - 2\sigma, \mu + 2\sigma]$
- PDF:

$$N(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$



Gaussian (or normal) distribution (wikip.)

Bayesian estimation of the mean of a Gaussian (1/2)

- Suppose we have observations $x = (x_1, \ldots, x_n)$ from $N(\mu, \sigma^2)$, where σ^2 is known.
- To learn μ, we specify a prior

$$\mu \sim N(\mu_0, \tau_0^2)$$

- Posterior

$$\begin{aligned}
p(\mu|x) &= \frac{p(x|\mu)p(\mu)}{p(x)} \propto p(\mu)p(x|\mu) \\
&= \frac{1}{\sqrt{2\pi}\tau_0} e^{-\frac{1}{2\tau_0^2}(\mu-\mu_0)^2} \times \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2} \\
&\propto e^{-\frac{1}{2\tau_0^2}(\mu-\mu_0)-\frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2} \\
&= \ldots \text{(details in BDA course)}
\end{aligned}$$

Bayesian estimation of the mean of a Gaussian (2/2)

- Posterior

$$
\begin{aligned}
p(\mu|x) &\propto e^{-\frac{1}{2\tau_n^2}(\mu-\mu_n)^2} \\
&\propto N(\mu|\mu_n, \tau_n^2)
\end{aligned}
$$

  where

$$
\mu_n = \frac{\frac{1}{\tau_0^2}\mu_0 + \frac{n}{\sigma^2}\overline{x}}{\frac{n}{\sigma^2} + \frac{1}{\tau_0^2}} \quad \text{and} \quad \frac{1}{\tau_n^2} = \frac{n}{\sigma^2} + \frac{1}{\tau_0^2}.
$$

- Posterior precision $1/\tau_n^2$: sum of prior precision $1/\tau_0^2$ and data precision n/σ^2
- Posterior mean μ_n: precision weighted average of prior mean μ_0 and data mean $\overline{x}$.

Conjugate prior distributions (1/2)

- In the previous example

$$\text{Prior: } \mu \sim N(\mu_0, \tau_0^2)$$
$$\text{Posterior: } \mu \sim N(\mu_n, \tau_n^2).$$

If the prior and posterior belong to the same family of distributions, we say that the prior is conjugate to the likelihood used.

  - For example, normal prior $\mu \sim N(\mu_0, \tau_0^2)$ is conjugate to the normal likelihood $N(x|\mu, \sigma^2)$.

- Conjugacy is useful, because it makes computations easy.

Conjugate prior distributions (2/2)

- With conjugate prior, the posterior is available in a closed form

$$p(\theta|x) \propto p(x|\theta)p(\theta)$$

  - Drop all terms not depending on θ
  - Recognize the result as a density function belonging to the same family of distributions as the prior $p(\theta)$, but with different parameters.

- Examples (likelihood - conjugate prior):
  - Likelihood for normal mean - Normal prior
  - Likelihood for normal variance - Inverse-Gamma prior
  - Bernoulli - Beta
  - Binomial - Beta
  - Exponential - Gamma
  - Poisson - Gamma

Gaussian distribution, unknown mean and precision (1/2)

- Suppose we have observations $x = (x_1, \ldots, x_n)$ from $N(\mu, \lambda^{-1})$, where both the mean μ and the precision λ are unknown.
- The conjugate prior distribution is the normal-gamma distribution

$$
\begin{aligned}
p(\mu, \lambda | \mu_0, \beta, a, b) &= N(\mu | \mu_0, (\beta\lambda)^{-1}) \mathrm{Gam}(\lambda | a, b) \\
&\equiv \text{Normal-Gamma}(\mu, \lambda | \mu_0, \beta, a, b)
\end{aligned}
$$

Note the dependency of the prior of μ on the value of λ.

Gaussian distribution, unknown mean and precision (2/2)

- The conjugate prior distribution is the normal-gamma distribution

$$p(\mu, \lambda | \mu_0, \beta, a, b) = \text{Normal-Gamma}(\mu, \lambda | \mu_0, \beta, a, b)$$

- Posterior

$$p(\mu, \lambda | x) = \text{Normal-Gamma}(\mu, \lambda | \mu_n, \beta_n, a_n, b_n),$$

with

$$\begin{aligned}
\mu_n &= \frac{\beta\mu_0 + n\overline{x}}{\beta + n} \\
\beta_n &= \beta + n \\
a_n &= a + \frac{n}{2} \\
b_n &= b + \frac{1}{2}\left(ns + \frac{\beta n(\overline{x} - \mu_0)^2}{\beta + n}\right)
\end{aligned}$$

Gaussian distribution, unknown mean and precision, example (1/2)

- Simulate samples from $N(\mu = 2, \sigma^2 = 0.25)$
  - precision $\lambda = 4$
- Try to learn μ and λ
- Specify prior

$$p(\mu, \lambda | \mu_0, \beta, a, b) = \text{Normal-Gamma}(\mu, \lambda | \mu_0, \beta, a, b)$$

with

$$\mu_0 = 0, \quad \beta = 0.001, \quad a = 0.01, \quad b = 0.01$$

- See: *normal_example.m*

Gaussian distribution, unknown mean and precision, example (2/2)

- When μ and λ have distribution

  $$\text{Normal-Gamma}(\mu, \lambda | \mu_n, \beta_n, a_n, b_n) = N(\mu | \mu_n, (\beta_n \lambda)^{-1}) \text{Gam}(\lambda | a_n, b_n),$$

  marginal distribution of λ can be plotted using the PDF of $\text{Gam}(\lambda | a_n, b_n)$
- To plot the marginal distribution of μ, we need to take the dependence on λ into account.
  - we compute the marginal distribution of μ by averaging over $N(\mu | \mu_n, (\beta_n \lambda_i)^{-1})$, for multiple λ_i simulated from $\text{Gam}(\lambda | a_n, b_n)$
  - (could also be done analytically...)

Consistency

- If $p(x|\theta_t)$ is the true data generating mechanism, and A is a neighborhood of θ_t, then

$$p(\theta \in A|x) \overset{n\to\infty}{\longrightarrow} 1.$$

- The posterior distribution concentrates around the true value (if such a value exists!). See the *normal_example.m*
- It follows that

$$\overline{\theta}_{MAP} \overset{n\to\infty}{\longrightarrow} \theta_t \quad \text{and} \quad \overline{\theta}_{ML} \overset{n\to\infty}{\longrightarrow} \theta_t$$

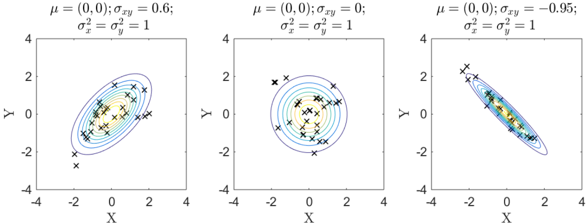
Multivariate Gaussian distribution

$$N_D(x|\mu,\Sigma) \equiv (2\pi)^{-\frac{D}{2}}|\Sigma|^{-\frac{1}{2}}e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

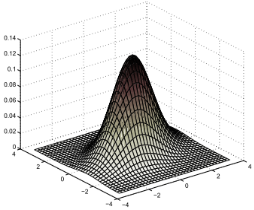
- D: dimension, μ: mean, Σ: covariance matrix. With $D = 2$:

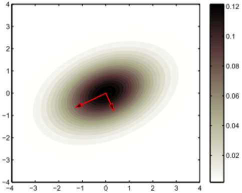
$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix}$$

- $\sigma_{12} = \sigma_{21}$: covariance between x_1 and x_2. (tells direction of dependency)
- $\rho_{12} = \sigma_{12}/(\sigma_1\sigma_2)$:correlation between x_1 and x_2. (direction and strength)



Multivariate Gaussian - characterization (1/2)





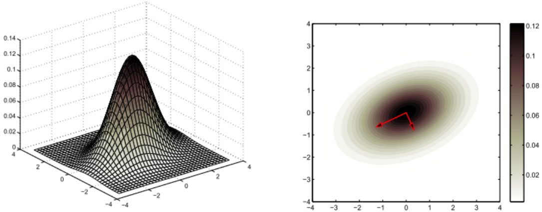
- Eigendecomposition

$$\Sigma = E \Lambda E^T,$$

  where $E^T E = I$ and $\Lambda =$diag$(\lambda_1, \ldots, \lambda_D)$.
- Now the transformation

$$y = \Lambda^{-\frac{1}{2}} E^T (x - \mu)$$

  can be shown to have the distribution $N_D(0, I)$ (product of D independent standard Gaussians)



- Thus, $x = E\Lambda^{\frac{1}{2}}y + \mu$ with distribution $N_D(\mu, \Sigma)$ is obtained from standard independent Gaussians y by
  - *scaling* by the square roots of eigenvalues
  - *rotating* by the eigenvectors
  - *shifting* by adding the mean

Marginalization and conditioning (1/2)

- Let $z \sim N(\mu, \Sigma)$ and consider partitioning it as:

$$z = \begin{pmatrix} x \\ y \end{pmatrix}$$

with

$$\mu = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} \quad \text{and} \quad \Sigma = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{pmatrix}.$$

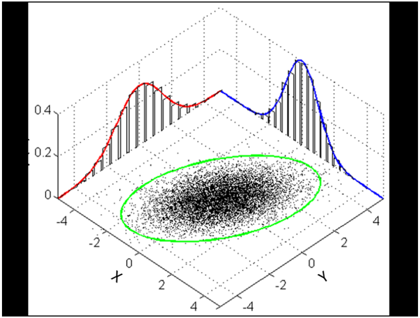
Marginalization and conditioning (2/2)

- Then

$$p(x) \sim N(\mu_x, \Sigma_{xx}) \quad \textbf{(marginalization)}$$

$$p(x|y) = N(\mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y), \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}) \quad \textbf{(conditioning)}$$

$\Longrightarrow$Marginals and conditionals of M-V Gaussians are still M-V Gaussian.



Important identities related to the multivariate Gaussian

- **Linear transformation**: if

$$y = Mx + \eta,$$

where $x \sim N(\mu_x, \Sigma_x)$ and $\eta \sim N(\mu, \Sigma)$,then

$$p(y) = N(y|M\mu_x + \mu, M\Sigma_x M^T + \Sigma)$$

- **Completing the square**:

$$\frac{1}{2}x^T Ax - b^T x = \frac{1}{2}(x - A^{-1}b)^T A(x - A^{-1}b) - \frac{1}{2}b^T A^{-1} b$$

From which one can derive, for example

$$\int \exp(-\frac{1}{2}x^T Ax + b^T x)dx = \sqrt{\det(2\pi A^{-1})}\exp(\frac{1}{2}b^T A^{-1} b)$$

Multivariate Gaussian - ML fitting

- Let $x = (x_1, \ldots, x_n)$ be from $N(\mu, \Sigma)$ with unknown μ and Σ. Log-likelihood, assuming data are *i.i.d.*:

$$
\begin{aligned}
L(\mu, \Sigma) &= \sum_{i=1}^{N} \log p(x_i | \mu, \Sigma) \\
&= -\frac{1}{2} \sum_{i=1}^{N} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu) - \frac{N}{2} \log \det(2\pi\Sigma)
\end{aligned}
$$

Multivariate Gaussian - ML fitting

- Differentiate $L(\mu, \Sigma)$ w.r.t. the vector μ:

$$\nabla_{\mu} L(\mu, \Sigma) = \sum_{i=1}^{N} \Sigma^{-1}(x_i - \mu)$$

  Equating to zero gives

$$\sum_{i=1}^{N} \Sigma^{-1} x_i = N\Sigma^{-1}\mu.$$

  Thus we get

$$\widehat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

- Similarly one can derive:

$$\widehat{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \overline{x})(x_i - \overline{x})^T$$

Multivariate Gaussian - Bayesian learning*

- Gaussian-Wishart is the conjugate prior, when $X_i \sim N(\mu, \Lambda)$ and both mean μ and precision Λ are unknown:

$$p(\mu, \Lambda | \mu_0, \beta, W, \nu) = N(\mu | \mu_0, (\beta\Lambda)^{-1}) \mathcal{W}(\Lambda | W, \nu)$$

- If X_i are scalar, this is equivalent to the Gaussian-Gamma distribution.
- Posterior

$$p(\mu, \Lambda | x) = N(\mu | \mu_n, (\beta_n \Lambda)^{-1}) \mathcal{W}(\Lambda | W_n, \nu_n)$$

Wishart distribution*

- Wishart distribution is a distribution for nonnegative-definite matrix-valued random variables

$$\Lambda \sim \mathcal{W}(\Lambda | W, \nu)$$

$$E(\Lambda) = \nu W$$
$$\mathrm{Var}(\Lambda_{ij}) = n(w_{ij}^2 + w_{ii} w_{jj})$$

- Further: exercises...

Linear models with Gaussian noise

- **Data** $\mathcal{D} = \{(\mathbf{x}_i, y_i), i = 1, \ldots, N\}$
  - $\mathbf{x}_i$: the input
  - y_i: the output
- **Model**:

$$y = \underbrace{f(\mathbf{w}, \mathbf{x})}_{\text{clean output}} + \underbrace{\eta}_{\text{noise}}, \quad \eta \sim N(0, \beta^{-1})$$

- In the simplest case

$$\begin{aligned} f(\mathbf{w}, \mathbf{x}) &= \mathbf{w}^T \mathbf{x} \\ &= w_1 x_1 + \ldots + w_D x_D \end{aligned}$$

- The *parameters* w_i are also called the *weights*

Bayesian linear parameter models

- A prior distribution $p(\mathbf{w}|\alpha)$ is placed on the weights $\mathbf{w}$.
- The posterior distribution $p(\mathbf{w}|\mathcal{D}, \Gamma)$ can be computed, and reflects the uncertainty of the parameters.

Prior distribution

- A Gaussian prior distribution may placed on **w**:

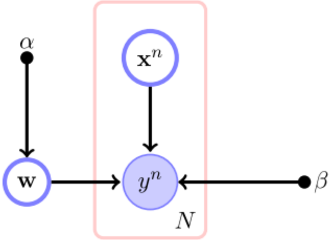
$$p(\mathbf{w}|\alpha) = N(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I})$$

$$= \prod_{i=1}^{D} N(w_i|0, \alpha^{-1}) = \left(\frac{\alpha}{2\pi}\right)^{\frac{D}{2}} e^{-\frac{\alpha}{2}\sum_i w_i^2}$$

- Posterior

$$\log p(\mathbf{w}|\Gamma, \mathcal{D}) = -\frac{\beta}{2}\sum_{i=1}^{N}\left[y_i - \mathbf{w}^T\mathbf{x}_i\right]^2 - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w} + \text{const}$$

Hyperparameters



- α: *precision* of the *regression weights*
  - determines the amount of regularization
  - large precision $\rightarrow$ small variance $\rightarrow$ weights are close to zero
- β: *precision* of the noise
- $\Gamma = \{\alpha, \beta\}$ are called the **hyperparameters** (in the course book...)

Posterior distribution

- Posterior distribution is obtained by completing the square (left as an exercise):

$$p(\mathbf{w}|\Gamma, \mathcal{D}) = N(\mathbf{w}|\mathbf{m}, S)$$

where

$$S = \left(\alpha I + \beta \sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i{}^T\right)^{-1}, \qquad \mathbf{m} = \beta S \sum_{i=1}^{N} y_i \mathbf{x}_i$$

- Mean prediction

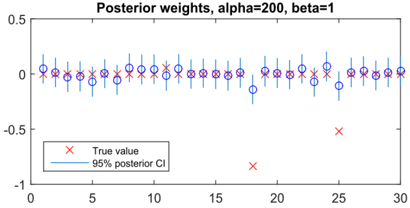
$$\widetilde{y} = \int \mathbf{w}^T \mathbf{x} \times p(\mathbf{w}|\Gamma, \mathcal{D}) d\mathbf{w} = \mathbf{m}^T \mathbf{x}$$

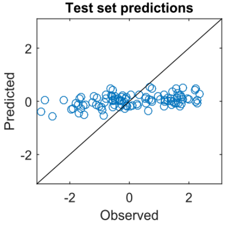
Example, impact of hyperparameters (1/3)

- Setup: simulate $y = \mathbf{w}_{true}^T \mathbf{x} + \epsilon$, where $\epsilon \sim N(0, \beta^{-1})$ and $\beta = 1$
- The goal is to investigate how hyperparameter α affects the posterior distribution of the parameters $\mathbf{w}$

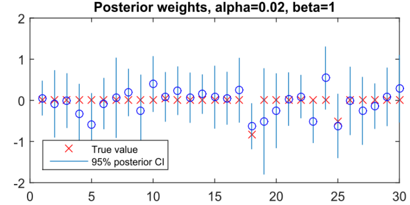
Example, impact of hyperparameters (2/3)

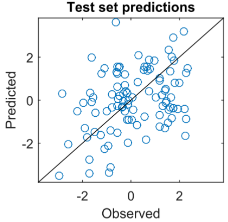
- Too large α, $Var(y - \tilde{y}) = 1.54$ (Original $Var(y) = 1.75$)





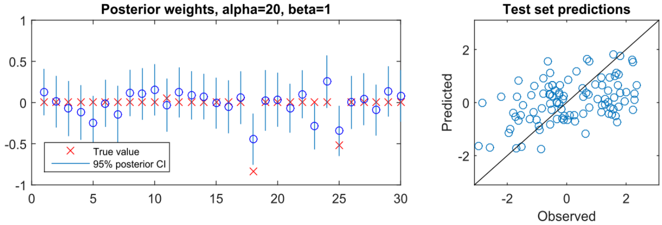
- Too small α, $Var(y - \tilde{y}) = 2.48$



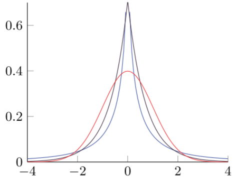


Example, impact of hyperparameters (3/3)

- About good α, $Var(y - \tilde{y}) = 1.46$
- A compromise between bias and variance

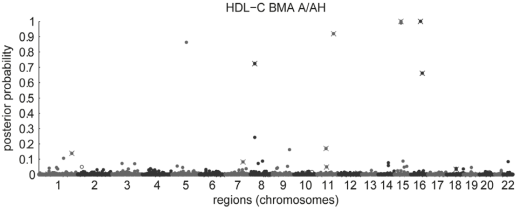


- Other sparse priors (e.g., Laplace, horse-shoe, spike-and-slab):



Example: genetic association studies

- Analysis of $\sim 1,000,000$ genetic polymorphisms in $\sim 50,000$ genomic regions (Peltola et al., 2012, *PLoS ONE*).
- *Spike-and-slab* prior on regression weights



Important points

- Bayesian learning of the Gaussian distribution using conjugate priors
- Multivariate Gaussian
  - Characterization
  - Marginal & conditional distributions
  - Linear transformation & completing the square
- By placing a Gaussian prior on the parameters of linear regression, the posterior is also Gaussian.
- Meaning and impact of hyperparameters in Bayesian linear regression.