Probabilities and Estimations

Richard Yi Da Xu

School of Computing & Communication, UTS

September 27, 2018

Gaussian, or normal distribution

- 1-dimensional case:

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

- k-dimensional case:

$$p(X) = \mathcal{N}(X|\mu, \Sigma) = (2\pi)^{-k/2}|\Sigma|^{-\frac{1}{2}} \exp^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$$

$$
\begin{aligned}
\mathbb{E}[X] &= \int_X x(2\pi)^{-d/2}|\Sigma|^{-\frac{1}{2}} \exp^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)} \mathrm{d}x \qquad \text{let } z = x - \mu \\
&= (2\pi)^{-d/2}|\Sigma|^{-\frac{1}{2}} \int_Z \exp^{-\frac{1}{2}z^T\Sigma^{-1}z}(z+\mu)\mathrm{d}z \\
&= (2\pi)^{-d/2}|\Sigma|^{-\frac{1}{2}} \int_Z \underbrace{\exp^{-\frac{1}{2}z^T\Sigma^{-1}z}}_{\text{even}} \underbrace{z}_{\text{odd}} \mathrm{d}z + \mu \underbrace{\int_Z \exp^{-\frac{1}{2}z^T\Sigma^{-1}z} \mathrm{d}z}_{(2\pi)^{D/2}|\Sigma|^{\frac{1}{2}}} \\
&= \mu
\end{aligned}
$$

Second Order moment of multivariate Gaussian (1): pre-requesits

- Let $\triangle^2 = (x-\mu)^T\Sigma^{-1}(x-\mu)$ $\quad\quad$ $\triangle$ = mahalanobis distance
- Let $(\lambda_1, \mathbf{e}_1) \dots (\lambda_d, \mathbf{e}_d)$ be eigen (value, vector) pairs of Σ

$$\Sigma\mathbf{e}_i = \lambda_i\mathbf{e}_i$$

- $\Sigma = \sum_{i=1}^{d} \lambda_i \mathbf{e}_i \mathbf{e}_i^T$
- $\Lambda = \Sigma^{-1} = \sum_{i=1}^{d} \frac{1}{\lambda_i} \mathbf{e}_i \mathbf{e}_i^T$
- $|\Sigma|^{1/2} = \prod_{i=1}^{d} \lambda_i^{1/2}$
- $|\Sigma|^{-1/2} = \prod_{i=1}^{d} \lambda_i^{-1/2}$

Second Order moment of multivariate Gaussian (2): pre-requesits

Let's change the axis to make vector $x-\mu$ eigen-vector aligned:

- Let each dimension of Y, i.e., $y_i = \mathbf{e}_i^T(x-\mu)$
- $Y = \begin{bmatrix} y_1 \\ \dots \\ y_d \end{bmatrix} = \begin{bmatrix} \mathbf{e}_1^T \\ \dots \\ \mathbf{e}_d^T \end{bmatrix}(x-\mu) = E^T(x-\mu)$

$$p(Y) = \mathcal{N}\left(Y|0, \begin{bmatrix} \lambda_1 & \dots & 0 \\ \dots & \lambda_i & 0 \\ \dots & \dots & \lambda_d \end{bmatrix}\right) = \prod_{i=1}^{d} \frac{1}{(2\pi\lambda_i)^{-1/2}} \exp^{-\frac{y_i}{2\lambda_i}}$$

- $J = \frac{\partial X}{\partial Y} = \begin{bmatrix} \frac{\partial x_{11}}{\partial y_{11}} & \dots & \frac{\partial x_{1d}}{\partial y_{1d}} \\ \dots & \dots & \dots \\ \frac{\partial x_{d1}}{\partial y_{d1}} & \dots & \frac{\partial x_{dd}}{\partial y_{dd}} \end{bmatrix} = E \implies |J| = |E| = 1$

Second Order moment of multivariate Gaussian (3): pre-requesits

$$\left(\sum_{i=1}^{d} \mathbf{e}_i^T y_i\right) \Sigma^{-1} \left(\sum_{i=1}^{d} \mathbf{e}_i y_i\right) = \left(\sum_{i=1}^{d} \mathbf{e}_i^T y_i\right) \left(\sum_{k=1}^{d} \frac{1}{\lambda_k} \mathbf{e}_k \mathbf{e}_k^T\right) \left(\sum_{i=1}^{d} \mathbf{e}_i y_i\right)$$

$$= \sum_{k=1}^{d} \sum_{i=1}^{d} \sum_{j=1}^{d} \frac{y_i y_j}{\lambda_k} \left(\mathbf{e}_i^T \mathbf{e}_k\right) \left(\mathbf{e}_k^T \mathbf{e}_j\right) \text{ only terms remain is when } i = j = k$$

$$= \sum_{i=1}^{d} \frac{y_i y_i}{\lambda_i} \left(\mathbf{e}_i^T \mathbf{e}_i\right) \left(\mathbf{e}_i^T \mathbf{e}_i\right) = \sum_{i=1}^{d} \frac{y_i^2}{\lambda_i}$$

Second Order moment of multivariate Gaussian $\mathbb{E}(XX^T)$

$$\mathbb{E}[XX^T] = \int_X xx^T (2\pi)^{-d/2} |\Sigma|^{-\frac{1}{2}} \exp^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)} \mathrm{d}x \qquad \text{let } z = x - \mu$$

$$= (2\pi)^{-d/2} |\Sigma|^{-\frac{1}{2}} \int_Z \exp^{-\frac{1}{2} z^T \Sigma^{-1} z} (z+\mu)(z+\mu)^T \left| \frac{\partial x}{\partial z} \right| \mathrm{d}z$$

$$= (2\pi)^{-d/2} |\Sigma|^{-\frac{1}{2}} \int_Z \exp^{-\frac{1}{2} z^T \Sigma^{-1} z} \left(zz^T + \underbrace{z^T \mu}_{\text{odd}} + \underbrace{\mu^T z}_{\text{odd}} + \mu\mu^T \right) \mathrm{d}z$$

$$= \mu\mu^T + (2\pi)^{-d/2} |\Sigma|^{-\frac{1}{2}} \int_Z \exp^{-\frac{1}{2} z^T \Sigma^{-1} z} zz^T \mathrm{d}z$$

So, let's find out what $\int_Z \exp^{-\frac{1}{2} z^T \Sigma^{-1} z} zz^T \mathrm{d}z$ is!

Second Order moment of multivariate Gaussian $\mathbb{E}(XX^T) = \Sigma$

Let $Y = \begin{bmatrix} y_1 \\ \dots \\ y_d \end{bmatrix} = \begin{bmatrix} \mathbf{e}_1^T \\ \dots \\ \mathbf{e}_d^T \end{bmatrix} \underbrace{(x - \mu)}_{\mathbf{z}} = E^T \mathbf{Z}$, then, $\mathbf{Z} = [\mathbf{e}_1, \dots, \mathbf{e}_d]\ Y = \sum_{i=1}^d \mathbf{e}_i y_i$

$$
\begin{aligned}
&\int_{\mathbf{z}} \exp^{-\frac{1}{2}\mathbf{z}^T \Sigma^{-1} \mathbf{z}} \mathbf{Z}\mathbf{Z}^T d\mathbf{Z} \\
&= \int_Y \exp^{-\frac{1}{2}\left(\sum_{i=1}^d \mathbf{e}_i y_i\right)^T \Sigma^{-1} \left(\sum_{i=1}^d \mathbf{e}_i y_i\right)} \left(\sum_{i=1}^d \mathbf{e}_i y_i\right) \left(\sum_{i=1}^d \mathbf{e}_i y_i\right)^T \left|\frac{\partial \mathbf{Z}}{\partial Y}\right| dY \\
&= \int_Y \exp^{-\frac{1}{2}\left(\sum_{k=1}^d \frac{y_k^2}{\lambda_k}\right)} \left(\sum_{i=1}^d \mathbf{e}_i y_i\right) \left(\sum_{i=1}^d \mathbf{e}_i^T y_i\right) dY \\
&= \sum_{i=1}^d \sum_{j=1}^d \mathbf{e}_j \mathbf{e}_i^T \int_Y \exp^{-\frac{1}{2}\left(\sum_{k=1}^d \frac{y_k^2}{\lambda_k}\right)} y_i y_j \, dY \\
&= \sum_{i=1}^d \mathbf{e}_i \mathbf{e}_i^T \underbrace{\int_{y_i} \exp^{-\frac{1}{2}\left(\frac{y_i^2}{\lambda_i}\right)} y_i^2 \, dy_i}_{\lambda_i (2\pi\lambda_i)^{1/2}} \underbrace{\left(\prod_{k=1,k\neq i}^d \int_{y_k} \exp^{-\frac{1}{2}\left(\frac{y_k^2}{\lambda_k}\right)} dy_k\right)}_{\prod_{k=1,k\neq i}^d (2\pi\lambda_i)^{1/2}} \quad \text{only terms } i = j \text{ remain} \\
&= (2\pi)^{d/2} |\Sigma|^{1/2} \sum_{i=1}^d \mathbf{e}_i \mathbf{e}_i^T \lambda_i
\end{aligned}
$$

Most of the distributions we are going to look at are from **exponential family**
exponential family can be expressed in terms of its natural parameters:

$$\exp\left(T(x)^T\eta - A(\eta) - B(x)\right)$$

Think about why is this representation useful?

Always have in mind ask yourself where are the **support** of these distributions, i.e., where $p(X) > 0$?

More about Gaussian 1-d: Natural Parameter Representation

$$
\begin{aligned}
\mathcal{N}(x;\mu,\sigma^2) &= (2\pi\sigma^2)^{-1/2}\exp^{-\frac{(x-\mu)^2}{2\sigma^2}} \\
&= \exp\left(-\frac{x^2-2x\mu+\mu^2}{2\sigma^2}-\frac{1}{2}\ln(2\pi\sigma^2)\right) \\
&= \exp\left(-\frac{1}{2\sigma^2}x^2+\frac{\mu}{\sigma^2}x-\frac{\mu^2}{2\sigma^2}-\frac{1}{2}\ln(2\pi\sigma^2)\right) \\
&= \exp\left(\begin{bmatrix} x \\ x^2 \end{bmatrix}^T \begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix} - \frac{\mu^2}{2\sigma^2}-\frac{1}{2}\ln(2\pi\sigma^2)\right) \\
&= \exp\left(T(x)^T\eta - \left(\frac{-\eta_1^2}{4\eta_2}-\frac{1}{2}\ln(-2\eta_2)\right)-\frac{1}{2}\ln(2\pi)\right)
\end{aligned}
$$

- $T(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}$ $\qquad A(\eta) = \frac{-\eta_1^2}{4\eta_2} - \frac{1}{2}\ln(-2\eta_2)$
- $\eta = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix}$ $\qquad$ Reverse is: $\theta = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix} = \begin{bmatrix} \frac{-\eta_1}{2\eta_2} \\ -\frac{1}{2\eta_2} \end{bmatrix}$

- Gamma Distribution

$$p(X) = \text{Gamma}(X|a,b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp^{-bx}$$

```
>> a = 1; b = 2; gamrnd(a,b, 10)
```

- Inverse Gamma Distribution

$$p(X) = \text{IG}(X|a,b) = \frac{b^a}{\Gamma(a)} x^{-a-1} \exp^{-b/x}$$

$X \sim \text{Gamma}(a,b) \implies \frac{1}{X} \sim \text{IG}(a,b)$

Positive Matrix Distributions

- Support $\mathbf{X} \in \mathbb{S}^p_{++}$
- Wishart Distribution:

$$p(\mathbf{X}) = \text{Wishart}(\mathbf{X}; \Psi, \nu) = \frac{|\mathbf{X}|^{\frac{\nu-p-1}{2}} \exp^{-\frac{\text{tr}(\Psi^{-1}\mathbf{X})}{2}}}{2^{\frac{\nu p}{2}} |\Psi|^{\frac{\nu}{2}} \Gamma_p\left(\frac{\nu}{2}\right)}$$

$\mathbb{E}(\mathbf{X}) = \nu\Psi$

```
>> Psi = [1 0; 0 1]; nv = 10; wishrnd(Psi,nv)
```

Larger $n \implies X \to nV \implies \mathbb{VAR}(X) \to 0$

- Inverse Wishart Distribution:

$$P(\mathbf{X}) = IW(\mathbf{X}; \Psi, \nu) = \frac{|\Psi|^{\frac{\nu}{2}}}{2^{\frac{\nu p}{2}} \Gamma_p(\frac{\nu}{2})} |\mathbf{X}|^{-\frac{\nu+p+1}{2}} e^{-\frac{1}{2}\text{tr}(\Psi\mathbf{X}^{-1})}$$

Weight distributions

- k-dimensional Dirichlet Distribution
- Support: $\sum_{i=1}^{k} p_i = 1$

$$\text{Dir}(p_1, \dots, p_k | \alpha_1, \dots, \alpha_k) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} p_i^{\alpha_i - 1}$$

- Beta Distribution
- Support: $0 \leq p \leq 1$

$$\text{Beta}(p|\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha - 1}(1 - p)^{\beta - 1}$$

Discrete distributions - modelling K-class number of occurance

- k-dimensional Multinomial Distribution

$$\text{Mult}(n_1, \dots, n_k | p_1, \dots p_k) = \frac{(\sum n_i)!}{n_1! \dots n_k!} \prod_{i=1}^{k} p_i^{n_i}$$

- Binomial Distribution

$$\text{Binomial}(n_1, n_2 | p) = \frac{(n_1 + n_2)!}{n_1! n_2!} p^{n_1} (1 - p)^{n_2}$$

- Bernoulli Distribution

$$\text{Bernoulli}(x | p) = p^x (1 - p)^{1-x}$$

Some very useful property of Dirichlet-Multinomial (1)

We let:

$$\int_{\Omega_{\mathbf{u}}} p\left(\mathbf{z}_{i \in A*} |\mathbf{u}\right) p(\mathbf{z}_{i \in \mathbf{A}}|\mathbf{u}) p(\mathbf{u}) d\mathbf{u} = \frac{\exp\left[\sum_{j=1}^{K} \ln \Gamma(\bar{z}^j_{\mathbf{A}*} + \bar{z}^j_{\mathbf{A}} + \beta\pi_j) - \ln \Gamma\left(\mathcal{Z}_{A*} + \mathcal{Z}_{\mathbf{A}} + \beta\right)\right]}{C^*_{\text{mul}} C^{\mathbf{A}}_{\text{mul}} Z_D(\beta\pi)}$$

where:

- $\mathbf{u} = p_1, \ldots p_k \sim \text{DIR}(\beta\pi_1, \ldots, \beta\pi_k)$ and $\sum_i^k \pi_i = 1$
- Dirichlet constants: $Z_D(\beta\pi) = \frac{\prod_{j=1}^K \Gamma(\beta\pi_j)}{\Gamma(\beta)}$
- Component-wise summations: $\bar{z}^j_{\mathbf{A}*} = \sum_{i=1}^{|A^*|} z_{ij}$ $\quad \bar{z}^j_{\mathbf{A}} = \sum_{i=1}^{|\mathbf{A}|} z_{ij}$
- Constants: $\mathcal{Z}_{A*} = \sum_j^K \bar{z}^j_{A*}$ $\quad \mathcal{Z}_{\mathbf{A}} = \sum_j^K \bar{z}^j_{\mathbf{A}}$
- multinomial constants: $C^*_{\text{mul}} = \prod_{i=1}^{|A^*|} \left(\frac{\prod_{j=1}^K z_{ij}!}{N_i!}\right)$ $\quad C^{\mathbf{A}}_{\text{mul}} = \prod_{i=1}^{|\mathbf{A}|} \left(\frac{\prod_{j=1}^K z_{ij}!}{N_i!}\right)$

We also let:

$$\int_{\Omega_{\mathbf{u}}} p(\mathbf{z}_{i \in \mathbf{A}}|\mathbf{u}) p(\mathbf{u}) d\mathbf{u} = \frac{\exp\left[\sum_{j=1}^{K} \ln \Gamma(\bar{z}^j_{\mathbf{A}} + \beta\pi_j) - \ln \Gamma\left(\mathcal{Z}_{\mathbf{A}} + \beta\right)\right]}{C^{\mathbf{A}}_{\text{mul}} Z_D(\beta\pi)}$$

Therefore:

$$\frac{\int_{\Omega_{\mathbf{u}}} p\left(\mathbf{z}_{i\in A_k}|\mathbf{u}\right) p(\mathbf{z}_{i\in \mathbf{A}}|\mathbf{u})p(\mathbf{u})d\mathbf{u}}{\int_{\Omega_{\mathbf{u}}} p(\mathbf{z}_{i\in \mathbf{A}}|\mathbf{u})p(\mathbf{u})d\mathbf{u}}$$

$$= \frac{1}{C^*_{\text{mul}}} \frac{\exp\left[\sum_{j=1}^{K} \ln\Gamma(\bar{z}^j_{\mathbf{A}*} + \bar{z}^j_{\mathbf{A}} + \beta\pi_j) - \ln\Gamma\left(\mathcal{Z}_{A*} + \mathcal{Z}_{\mathbf{A}} + \beta\right)\right]}{\exp\left[\sum_{j=1}^{K} \ln\Gamma(\bar{z}^j_{\mathbf{A}} + \beta\pi_j) - \ln\Gamma\left(\mathcal{Z}_{\mathbf{A}} + \beta\right)\right]}$$

$$= \frac{1}{C^*_{\text{mul}}} \exp\left[\sum_{j=1}^{K}\left[\ln\Gamma(\bar{z}^j_{\mathbf{A}*} + \bar{z}^j_{\mathbf{A}} + \beta\pi_j) - \ln\Gamma(\bar{z}^j_{\mathbf{A}} + \beta\pi_j)\right] - \ln\Gamma\left(\mathcal{Z}_{A*} + \mathcal{Z}_{\mathbf{A}} + \beta\right) + \ln\Gamma\left(\mathcal{Z}_{\mathbf{A}} + \beta\right)\right]$$

- Poisson Distribution

$$\text{Poisson}(x|\lambda) = \frac{\lambda^x}{x!} \exp(-\lambda)$$

Relationship between Binomial and Poisson

- Imagine you increase the number of independant Bernoulli draws (e.g. hours to seconds), i.e., n increase.
- The probablity (p) per time interval (e.g. prob. car appears) decreases.
- However, there is a constant relationship $\lambda = np$

Using identity:

$$\exp(x) = \lim_{n\to\infty}\left(1 + \frac{x}{n}\right)^n$$

$$\begin{aligned}
\text{Binomial}(x|n,p) &= \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!}\frac{\lambda}{n}^x (1 - \frac{\lambda}{n})^{n-x} \\
&= \underbrace{\frac{\lambda^x}{x!}}_{\text{constant}} \underbrace{\frac{n!}{(n-x)!}\frac{1}{n^x}} \left(1 - \frac{\lambda}{n}\right)^{n-x} \\
&= \frac{\lambda^x}{x!} \frac{\overbrace{n(n-1),\ldots(n-x+1)}^{n \text{ terms}}}{n^x} \left(1 - \frac{\lambda}{n}\right)^{n-x} \\
&= \frac{\lambda^x}{x!}\frac{n}{n}\frac{n-1}{n}\cdots\frac{n-x+1}{n}\left(1 - \frac{\lambda}{n}\right)^n\left(1 - \frac{\lambda}{n}\right)^{-x} \\
&= \frac{\lambda^x}{x!} 1 \left(1 - \frac{1}{n}\right) \cdots \left(1 - \frac{x+1}{n}\right)\left(1 - \frac{\lambda}{n}\right)^n\left(1 - \frac{\lambda}{n}\right)^{-x}
\end{aligned}$$

$$\lim_{n\to\infty} \text{Binomial}(x|n,p) = \lim_{n\to\infty}\binom{n}{x} p^x (1-p)^{n-x}$$

$$= \frac{\lambda^x}{x!}\lim_{n\to\infty}\left(1 - \frac{1}{n}\right)\cdots\lim_{n\to\infty}\left(1 - \frac{x+1}{n}\right)\lim_{n\to\infty}\left(1 - \frac{\lambda}{n}\right)^n\lim_{n\to\infty}\left(1 - \frac{\lambda}{n}\right)^{-x} = \frac{\lambda^x}{x!}\exp(-\lambda)$$

Relationship between Multinomial distribution and Poisson

$$\text{Poisson}(x|\lambda) = \frac{\lambda^x}{x!}\exp(-\lambda) \qquad \text{Mult}(n_1, \dots, n_k|p_1, \dots p_k) = \frac{(\sum n_i)!}{n_1! \dots n_k!}\prod_{i=1}^{k} p_i^{n_i}$$

suppose:

- $x_1 \sim \text{Poisson}(x|\lambda_1), \dots, x_k \sim \text{Poisson}(x|\lambda_k) \implies$
- The above generated two random variables:

1st random variable: $\left(n = \sum_{i=1}^{k} x_i\right) \sim \text{Poisson}(\lambda_1 + \lambda_2 + \dots + \lambda_k)$

2nd random variable: $\mathbf{x} = (x_1, \dots, x_k)|n \sim \text{Mult}(n, p_1, \dots p_k)$ where $p_i = \frac{\lambda_i}{\sum_{j=1}^{k}\lambda_j}$

- $X \sim \text{Poisson}(\lambda)$
- T denote the length of time until k arrivals.

Extend this Relationship to **Process**

- Grouped data $x_1, \dots x_J$ for any measurable disjoint partition $A_1, \dots A_Q$ of Ω,
- Jointly model the count random variables $\{X_j(A_q)\}$.
- Poisson process $X_j \sim \text{PP}(G)$, with a shared Completely Random Measure G on $\Omega : X_j(A) \sim Pois(G(A))$
- $X_j \sim \text{PP}(G)$
  $\equiv X_j \sim \text{MP}(X_j(\Omega), \tilde{G}), \qquad X_j(\Omega) \sim Pois(G(\Omega)) \qquad \text{where } \tilde{G} = \frac{G}{G(\Omega)}$

$$
\begin{aligned}
X_j &\sim \text{NBP}\left(G_0, \frac{1}{c+1}\right) = \int_G \text{PP}(X_j|G)\text{GaP}(c, G_0)\text{d}G \\
&\sim \text{NBP}\left(G_0, p\right) = \int_G \text{PP}(X_j|G)\text{GaP}\left(\frac{J(1-p)}{p}, G_0\right)\text{d}G
\end{aligned}
$$

Non-exponential family distribution

They often can be constructed from two exponential family distributions:

- Student-t distribution

$$
\begin{aligned}
t(x|\mu, a, b) &= \int_\lambda \mathcal{N}(x; \mu, \lambda^{-1})\text{Gamma}(\lambda; a, b) \\
&= \int_\lambda \left(\frac{\lambda}{2\pi}\right)^{1/2} \exp\left\{-\frac{\lambda}{2}(\lambda - \mu)^2\right\} \frac{b^a}{\Gamma(a)} \lambda^{a-1} \exp^{-b\lambda} \\
&= \frac{b^a}{\Gamma(a)} \left(\frac{1}{2\pi}\right)^{1/2} \int_\lambda \lambda^{1/2} \exp\left\{-\frac{\lambda}{2}(\lambda - \mu)^2\right\} \lambda^{a-1} \exp^{-b\lambda} \\
&= \frac{b^a}{\Gamma(a)} \left(\frac{1}{2\pi}\right)^{1/2} \int_\lambda \lambda^{a+1/2-1} \exp\left\{-\left[b + \frac{1}{2}(\lambda - \mu)^2\right]\lambda\right\} \\
&= \frac{b^a}{\Gamma(a)} \left(\frac{1}{2\pi}\right)^{1/2} \frac{\Gamma(a+1/2)}{\left[b + \frac{1}{2}(x-\mu)^2\right]^{a+1/2}} \\
&= \frac{\Gamma(a+1/2)}{\Gamma(a)} \left(\frac{1}{2\pi}\right)^{1/2} \left(b + \frac{1}{2}(x-\mu)^2\right)^{-(a+1/2)} \underbrace{\left(\frac{1}{b}\right)^{-(a+1/2)} \left(\frac{1}{b}\right)^{1/2}}_{b^a} \\
&= \frac{\Gamma(a+1/2)}{\Gamma(a)} \left(\frac{1}{2\pi b}\right)^{1/2} \left(1 + \frac{1}{2b}(x-\mu)^2\right)^{-(a+1/2)}
\end{aligned}
$$

Conjugacy

Looking at the posterior, prior relationship:

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{\int_\theta p(X|\theta)p(\theta)} \propto p(X|\theta)p(\theta)$$

- Wouldn't it be good if $p(\theta|X)$ and $p(\theta)$ are the same family of distributions?
- Many conjugacy exist

For example:

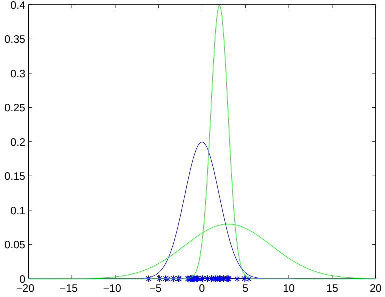
- the prior $p(\mu) = \mathcal{N}(\mu_0, \sigma_0^2)$
- and the likelihood $p(X|\mu) = \mathcal{N}(\mu, \sigma)$.
- and the posterior $p(\mu|X)$ is also a Gaussian distribution
- Exercise, derive the above

Multinomial-Dirichlet

$$
\begin{aligned}
& P(p_1, \dots, p_k | n_1, \dots, n_k) \\
& \propto \underbrace{\frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k p_i^{\alpha_i - 1}}_{\text{Dir}(p_1, \dots, p_k | \alpha_1, \dots, \alpha_k)} \underbrace{\frac{n!}{n_1! \dots n_k!} \prod_{i=1}^k p_i^{n_i}}_{\text{Mult}(n_1, \dots, n_k | p_1, \dots p_k)} \\
& \propto \prod_{i=1}^k p_i^{\alpha_i - 1} \prod_{i=1}^k p_i^{n_i} = \prod_{i=1}^k p_i^{\alpha_i - 1 + n_i} \\
& = \text{Dir}(p_1, \dots p_k | \alpha_i + n_i, \dots \alpha_k + n_k)
\end{aligned}
$$

Normal distributed data

- You believe data = $X = \{x_1, \dots x_N\}$ are Normal distributed:



Maximum Likelihood Estimation

- which “normal” distribution parameter $\theta = (\mu, \sigma)$ is more likely?
- It appears that the blue distribution is more likely than the green distribution. But why?
- In terms of probability, we find a particular θ that maximises the likelihood $p(X|\theta)$

$$\begin{aligned}\theta^{\text{MLE}} &= \arg\max_{\theta} \left(p(X|\theta)\right)\\ &= \arg\max_{\theta} \left(\prod_{i=1}^{N} \mathcal{N}(x_i; \mu, \sigma)\right)\end{aligned}$$

- How to solve this “argmax”? It depends on the distribution. But in the case of Gaussian, it's simple

MLE - log-likelihood

Instead of perform $\theta^{\text{MLE}} = \arg\max_\theta \left(p(X|\theta)\right)$, we perform:

$$
\begin{aligned}
\theta^{\text{MLE}} &= \arg\max_\theta \left(\underbrace{\log[p(X|\theta)]}_{\mathcal{L}(\theta)} \right) \\
&= \arg\max_\theta \left(\sum_{i=1}^{N} \log(\mathcal{N}(x_i; \mu, \sigma)) \right)
\end{aligned}
$$

$\mathcal{L}(\theta|X) = \log[p(X|\theta)]$ is called the log-likelihood **function**. It's NOT a probability distribution.

Why is log chosen?

- Firslty, log is a monotonically increasing function: $A \geq B \implies \log(A) \geq \log(B)$
- Secondly, log transforms multiplication into addition: $\log(AB) = \log(A) + \log(B)$

When need to perform MLE over Gaussian. Substitute Gaussian definition into:

$$\begin{aligned}
\theta^{\text{MLE}} &= \arg\max_{\theta}[\mathcal{L}(\theta|X)] = \arg\max_{\theta}\left(\sum_{i=1}^{N}\log(\mathcal{N}(x_i;\mu,\sigma))\right) \\
&= \arg\max_{\theta}\left(\sum_{i=1}^{N}\log\left[\frac{1}{\sigma\sqrt{2\pi}}\exp^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right]\right)
\end{aligned}$$

- Taking derivative with respect to both μ and σ^2
- Which one first? In Gaussian, only works if we take derivative with respect to μ first

MLE - Gaussian μ_{MLE}

When need to perform MLE over Gaussian. Substitute Gaussian definition into:

- Taking derivative with respect to both μ and σ^2
- Which one first? In Gaussian, only works if we take derivative with respect to μ first

$$= \frac{\partial \left(\sum_{i=1}^{N} \log \left[\frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \right] \right)}{\partial \mu}$$

$$= \frac{\partial \left(\sum_{i=1}^{N} \log \left[\exp^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \right] \right)}{\partial \mu} = \frac{\partial \left(\sum_{i=1}^{N} -\frac{(x_i-\mu)^2}{2\sigma^2} \right)}{\partial \mu}$$

$$= \sum_{i=1}^{N} \frac{(x_i - \mu)}{\sigma^2}$$

$$= \sum_{i=1}^{N} \frac{(x_i - \mu)}{\sigma^2} = 0 \implies \sum_{i=1}^{N} x_i = N\mu \implies \mu_{\text{MLE}} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

MLE - Gaussian σ^2_{MLE}

Once obtained μ_{MLE}, we substitute it into the $\mathcal{L}(\theta|X)$ function:

$$= \frac{\partial \left(\sum_{i=1}^N \log \left[\frac{1}{\sigma\sqrt{2\pi}} \exp^{-\frac{(x_i - \mu_{\text{MLE}})^2}{2\sigma^2}} \right] \right)}{\partial \sigma^2}$$

$$= \frac{-\partial \sum_{i=1}^N \log \sigma\sqrt{2\pi}}{\partial \sigma^2} + \frac{\partial \left(\sum_{i=1}^N \log \left[\exp^{-\frac{(x_i - \mu_{\text{MLE}})^2}{2\sigma^2}} \right] \right)}{\partial \sigma^2}$$

$$= \frac{-\frac{N}{2}\partial \log(\sigma^2\sqrt{2\pi})}{\partial \sigma^2} + \frac{\partial \left(\sum_{i=1}^N - \frac{(x_i - \mu_{\text{MLE}})^2}{2\sigma^2} \right)}{\partial \sigma^2}$$

$$= -\frac{N}{2\sigma^2} - \frac{1}{2} \left(\sum_{i=1}^N (x_i - \mu_{\text{MLE}})^2 \right) \frac{\partial \left(\frac{1}{\sigma^2} \right)}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2} \left(\sum_{i=1}^N (x_i - \mu_{\text{MLE}})^2 \right) \frac{1}{(\sigma^2)^2}$$

$$= \frac{1}{2\sigma^2} \left(-N + \left(\sum_{i=1}^N (x_i - \mu_{\text{MLE}})^2 \right) \frac{1}{\sigma^2} \right)$$

$$-N + \left(\sum_{i=1}^N (x_i - \mu_{\text{MLE}})^2 \right) \frac{1}{\sigma^2} = 0 \implies \left(\sum_{i=1}^N (x_i - \mu_{\text{MLE}})^2 \right) \frac{1}{\sigma^2} = N$$

$$\implies \sigma^2_{\text{MLE}} = \frac{\sum_{i=1}^N (x_i - \mu_{\text{MLE}})^2}{N}$$

MLE - Multinomial

- Think about the observations 1425 12351222 122124
- equivalently, $n_1 = 5$, $n_2 = 8$, $n_3 = 1$, $n_4 = 2$, $n_5 = 2$
- Why $\left(\frac{5}{16}\right)^5 \left(\frac{8}{16}\right)^8 \left(\frac{1}{16}\right)^1 \left(\frac{2}{16}\right)^2 \left(\frac{2}{16}\right)^2$ gives maximum likelihood?

$$\text{Mult}(n_1, \dots, n_k | p_1, \dots p_k) = \frac{(\sum n_i)!}{n_1! \dots n_k!} \prod_{i=1}^{k} p_i^{n_i}$$

$$\implies \operatorname*{arg\,max}_{p_1, \dots p_k} \ln \left(\Pr(n_1, \dots, n_k | p_1, \dots p_k) \right) = \operatorname*{arg\,max}_{p_1, \dots p_k} \sum_{i=1}^{k} n_i \ln(p_i)$$

$$\implies \text{LM}(\lambda, p_1, \dots p_k) = \sum_{i=1}^{k} n_i \ln(p_i) + \lambda \left(\sum_{i=1}^{k} p_i - 1 \right)$$

$$\frac{\partial \text{LM}(\lambda, p_1, \dots p_k)}{\partial p_i} = \frac{n_i}{p_i} - \lambda = 0 \implies p_i = \frac{n_i}{\lambda}$$

$$\frac{\partial \text{LM}(\lambda, p_1, \dots p_k)}{\partial \lambda} = \sum_{i=1}^{k} p_i - 1 = 0 \implies \sum_{i=1}^{k} \frac{n_i}{\lambda} = 1 \implies \lambda_{\text{ML}} = \sum_{i=1}^{k} = N$$

$$\implies p_{i\text{ML}} = \frac{n_i}{N}$$

Taking geometric mean-alike operations:

$$
\begin{aligned}
&(p_1 p_4 p_2 p_5)^{a_1} (p_1 p_2 p_3 p_5 p_1 p_2 p_2 p_2)^{a_2} (p_1 p_2 p_2 p_1 p_2 p_4)^{a_3} \\
=& p_1^{(a_1+2a_2+2a_3)} p_2^{(a_1+4a_2+3a_3)} p_3^{(a_2)} p_4^{(a_1+a_3)} p_5^{(a_1+a_2)} \\
=& p_1^{(\bar{n}_1)} p_2^{(\bar{n}_2)} p_3^{(\bar{n}_3)} p_4^{(\bar{n}_4)} p_5^{(\bar{n}_5)}
\end{aligned}
$$

Some pattern matching with previous slide shows:

$$
\implies p_{i\,\mathrm{ML}} = \frac{\bar{n}_i}{\sum_{i=1}^{k} \bar{n}_i}
$$

Solve MLE using Natural Parameters

$$\mathcal{N}(x;\mu,\sigma^2) = \mathcal{N}_{\text{nat}}(\eta_1,\eta_2) = \exp\left(T(x)^T\eta - \left(\frac{-\eta_1^2}{4\eta_2} - \frac{1}{2}\ln(-2\eta_2)\right) - \frac{1}{2}\ln(2\pi)\right)$$

$$\ln(\mathcal{N}_{\text{nat}}(x_i;\eta_1,\eta_2)) = T(x)^T\eta - \left(\frac{-\eta_1^2}{4\eta_2} - \frac{1}{2}\ln(-2\eta_2)\right) - \frac{1}{2}\ln(2\pi)$$

$$\implies \sum_{i=1}^{n}\ln(\mathcal{N}_{\text{nat}}(x_i;\eta_1,\eta_2)) = \mathrm{T}(\mathbf{x})^T\eta - \left(\frac{-\eta_1^2}{4\eta_2} - \frac{1}{2}\ln(-2\eta_2)\right)n - \frac{n}{2}\ln(2\pi)$$

$$\implies \frac{\partial\left(\sum_{i=1}^{n}\ln(\mathcal{N}_{\text{nat}}(x_i;\eta_1,\eta_2))\right)}{\partial\eta} = 0 \implies \frac{\partial\left(\frac{-\eta_1^2}{4\eta_2} - \frac{1}{2}\ln(-2\eta_2)\right)n}{\partial\eta} = \mathrm{T}(\mathbf{x})$$

Solve MLE using Natural Parameters (2)

- $\eta = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{bmatrix}$ Reverse is: $\theta = \begin{bmatrix} \mu \\ \sigma^2 \end{bmatrix} = \begin{bmatrix} \frac{-\eta_1}{2\eta_2} \\ -\frac{1}{2\eta_2} \end{bmatrix}$

$$
\begin{aligned}
\frac{\partial \left(\frac{-\eta_1^2}{4\eta_2} - \frac{1}{2} \ln(-2\eta_2) \right) n}{\partial \eta} &= \mathrm{T}(\mathbf{x}) \\
\implies \begin{bmatrix} -\frac{\eta_1}{2\eta_2} \\ \frac{\eta_1^2}{4\eta_2^2} - \frac{1}{2\eta_2} \end{bmatrix} &= \begin{bmatrix} \frac{\sum_{i=1}^n x_i}{n} \\ \frac{\sum_{i=1}^n x_i^2}{n} \end{bmatrix} \\
\implies \begin{bmatrix} \mu \\ \mu^2 + \sigma^2 \end{bmatrix} &= \begin{bmatrix} \frac{\sum_{i=1}^n x_i}{n} \\ \frac{\sum_{i=1}^n x_i^2}{n} \end{bmatrix}
\end{aligned}
$$

which is same as without using natural parameters

Maxmimum A Posterior Example: 1-d Gaussian

- What if I have some prior knowledge of of μ, for example, $\mu \sim \mathcal{N}(\mu_0, \sigma_0)$. This type of estimation is called Maximum a Posterior (MAP):

$$\theta_{\text{MAP}} = \arg\max_{\theta} \left(\log[p(X|\theta)p(\theta)]\right)$$

  Say what you need is to find the mean, i.e.,

$$\mu_{\text{MAP}} = \arg\max_{\mu} \left(\sum_{i=1}^{N} \log[\mathcal{N}(x_i|\mu, \sigma)\mathcal{N}(\mu; \mu_0, \sigma_0)] \right)$$

- How to solve "argmax"? Well easy, take the deriviative and let it equal zero. Works in the Gaussian case.

Does conjugacy always for Exponential family distribution?

- Prior

$$P(\theta, \Theta|\beta, \gamma) = \exp\left(\beta^T\theta + \beta^T\Theta\beta - \gamma A(\theta, \Theta) \underbrace{-\lambda_\theta\|\theta\|_2^2 - \lambda_\Theta\|\text{vec}(\Theta)\|_1}_{h(\theta,\Theta)}\right)$$

- Likelihood

$$\text{PMRF}(x|\theta, \Theta) = \exp\left(\theta^T x + x^T\Theta x \underbrace{- \sum_{s=1}^{p} \ln(x_s!)}_{h(x)} - A(\theta, \Theta)\right)$$

- Posterior

$$P(\theta, \Theta|x) \propto \exp\left(\underbrace{(x+\beta)}_{\hat{\beta}}{}^T\theta + \underbrace{(x+\beta)}_{\hat{\beta}}{}^T\Theta\underbrace{(x+\beta)}_{\hat{\beta}} - \underbrace{(\gamma+1)}_{\hat{\gamma}}A(\theta, \Theta) \underbrace{-\lambda_\theta\|\theta\|_2^2 - \lambda_\Theta\|\text{vec}(\Theta)\|_1}_{h(\theta,\Theta)}\right)$$

$$P(\mathbf{w}, \boldsymbol{\theta}_{1\ldots k}, \Theta_{1\ldots k}|\mathbf{x}) = P\left(\mathbf{x}|\mathbf{w}, \boldsymbol{\theta}_{1\ldots k}, \Theta_{1\ldots k}\right) P(\boldsymbol{\theta}_{1\ldots k}, \Theta_{1\ldots k}|\mathbf{w}) P_{Dir}(\mathbf{w}) \tag{1}$$

$$\propto \underbrace{\exp\left\{\left(\sum_{j=1}^{k} w_j \boldsymbol{\theta}_j\right)^T \mathbf{x} + \mathbf{x}^T \left(\sum_{j=1}^{k} w_j \Theta_j\right)\mathbf{x} - \sum_{s=1}^{p} \ln(x_s!)\right\}}_{\mathrm{PMRF}(\mathbf{x}|\mathbf{w},\boldsymbol{\theta}_{1\ldots k},\Theta_{1\ldots k})} \times \underbrace{\frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_k)} \prod_{i=1}^{k} w_i^{\alpha_i - 1}}_{P_{Dir}(\mathbf{w})} \tag{2}$$

$$\times \underbrace{\prod_{j=1}^{k} \exp\left\{\beta^T w_j \boldsymbol{\theta}_j + \beta^T w_j \Theta_j \beta - \gamma A(w_j\boldsymbol{\theta}_j, w_j\Theta_j) - \lambda_{\boldsymbol{\theta}} \|w_j\boldsymbol{\theta}_j\|_2^2 - \lambda\|\mathrm{vec}(w_j\Theta_j)\|_1\right\}}_{P(\boldsymbol{\theta}_{1\ldots k},\Theta_{1\ldots k}|\mathbf{w})} \tag{3}$$

$$\propto \exp\left\{\left(\sum_{j=1}^{k} w_j \boldsymbol{\theta}_j\right)^T \mathbf{x} + \mathbf{x}^T \left(\sum_{j=1}^{k} w_j \boldsymbol{\Theta}_j\right)\mathbf{x} + \left(\sum_{j=1}^{k} w_j \boldsymbol{\theta}_j\right)^T \beta + \beta^T \left(\sum_{j=1}^{k} w_j \boldsymbol{\Theta}_j\right)\beta \right. \tag{4}$$

$$\left. - \sum_{j=1}^{k}\left(\gamma A(w_j\boldsymbol{\theta}_j, w_j\Theta_j) + \lambda_{\boldsymbol{\theta}}\|w_j\boldsymbol{\theta}_j\|_2^2 + \lambda\|\mathrm{vec}(w_j\Theta_j)\|_1\right) + \sum_{j=1}^{k}(\alpha_i - 1)\ln w_i \right\} \tag{5}$$

$$\propto \exp\left\{\left(\sum_{j=1}^{k} w_j \boldsymbol{\theta}_j\right)^T (\mathbf{x}+\beta) + (\mathbf{x}+\beta)^T \left(\sum_{j=1}^{k} w_j \boldsymbol{\Theta}_j\right)(\mathbf{x}+\beta) \right. \tag{6}$$

$$\underbrace{\left. - \sum_{j=1}^{k}\left(\gamma A(w_j\boldsymbol{\theta}_j, w_j\Theta_j) + \lambda_{\boldsymbol{\theta}}\|w_j\boldsymbol{\theta}_j\|_2^2 + \lambda\|\mathrm{vec}(w_j\Theta_j)\|_1\right) + \sum_{j=1}^{k}(\alpha_i - 1)\ln w_i \right\}}_{\eta(\theta,\Theta)} \tag{7}$$

A case study (2)

$$= \exp\left\{ \left(\sum_{j=1}^{k} w_j \boldsymbol{\theta}_j \right)^T \tilde{\mathbf{x}} + \tilde{\mathbf{x}}^T \left(\sum_{j=1}^{k} w_j \boldsymbol{\Theta}_j \right) \tilde{\mathbf{x}} - \eta(\theta, \Theta) \right. \tag{8}$$

$$= \exp\left\{ \left[\sum_{j=1}^{k} w_j \boldsymbol{\theta}_j + \left(\sum_{j=1}^{k} w_j \boldsymbol{\Theta}_j \right)^T \tilde{\mathbf{x}} \right]^T \tilde{\mathbf{x}} - \eta(\theta, \Theta) \right. \tag{9}$$

$$\tag{10}$$

MAP Example Conti.

- Same trick applies: take the derivative with respect of μ and let it equal zero
- If you write out the expression for Gaussian fully, you will get:

$$\mu_{\text{MAP}} = \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\left(\frac{1}{n}\sum_{j=1}^{n} x_i\right) + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0$$

- see what happens if $\sigma_0 \to \infty$

Sum of variables - discrete case

$$G(z) = \mathrm{E}(z^X) = \sum_{x=0}^{\infty} p(x) z^x$$

logarithmic distribution:

$$Y_n \sim \mathrm{Log}(p) = p(k; r, p) = \frac{-p^k}{k \ln(1-p)} \qquad N \sim \mathrm{Poisson}(N; -r\ln(1-p))$$

$$G_N(z) = \sum_{N=0}^{\infty} \frac{(-r\ln(1-p))^N e^{r\ln(1-p)}}{N!} z^N = \exp^{(-r\ln(1-p))(z-1)}$$

Then $\left(X = \sum_{n=1}^{N} Y_n\right) \sim \mathrm{NB}(r, p)$

stochastic process

- A stochastic process $\{N(t), t \leq 0\}$ is said to be a counting process if $N(t)$ represents the total number of **events** that have occurred up to time t.
- $X_1, X_2, \ldots$ are times between events (or **life times**, or **inter-arrival times**).
- $S_n = X_1 + \cdots + X_n$ is the time of the n^{th} event.

Definition implies:

- $N(t) \leq 0$
- $N(t)$ is integer valued
- If $s < t$, then $N(s) \leq N(t)$
- For $s < t$, $N(t) - N(s)$ equals the number of events in $(s, t]$.