CSE 5523: Lecture Notes 9
Normal distributions

Normal distributions are widely used continuous distributions.

Contents

9.1 Normal distributions (De Moivre – Laplace Theorem)

A normal or Gaussian distribution is (w/in $O\left(\frac{1}{\sqrt{N}}\right)$ of) the limit of a binomial over N trials as $N \to \infty$ (this fits around the peak – e.g. binomials are > 0 only from 0 to N, Gaussians are > 0 everywhere):

$$\begin{aligned}
\lim_{N\to\infty} \binom{N}{n} p^n (1-p)^{N-n} &= \lim_{N\to\infty} \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n} && \text{def. of combination} \\
&\approx \lim_{N\to\infty} \frac{\sqrt{2\pi N}\left(\frac{N}{e}\right)^N}{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n \sqrt{2\pi(N-n)}\left(\frac{(N-n)}{e}\right)^{(N-n)}} p^n (1-p)^{N-n} && \text{Stirling's approx.} \\
&= \lim_{N\to\infty} \sqrt{\frac{2\pi}{(2\pi)^2} \frac{N}{n(N-n)}} \frac{e^{-N}}{e^{-n}e^{-(N-n)}} \frac{N^n p^n}{n^n} \frac{N^{N-n}(1-p)^{N-n}}{(N-n)^{N-n}} && \text{commutative axiom} \\
&= \lim_{N\to\infty} \sqrt{\frac{2\pi}{(2\pi)^2} \frac{N}{n(N-n)}} \left(\frac{Np}{n}\right)^n \left(\frac{N(1-p)}{N-n}\right)^{N-n} && \text{mult. inverses} \\
&= \lim_{N\to\infty} \sqrt{\frac{2\pi}{(2\pi)^2} \frac{N}{n(N-n)}} \exp\left\{ n \ln\left(\frac{Np}{n}\right) + (N-n) \ln\left(\frac{N(1-p)}{N-n}\right) \right\} && \text{def. of nat. log}
\end{aligned}$$

We now simplify the exponential:

$$\begin{aligned}
& n \ln\left(\frac{Np}{n}\right) + (N-n) \ln\left(\frac{N(1-p)}{N-n}\right) \\
&= -n \ln\left(\frac{n}{Np}\right) - (N-n) \ln\left(\frac{N-n}{N(1-p)}\right) && \text{log of inverses} \\
&= -n \ln\left(\frac{Np+n-Np}{Np}\right) - (N-n) \ln\left(\frac{N-Np-n+Np}{N(1-p)}\right) && \text{add inverses} \\
&= -n \ln\left(1 + \frac{n-Np}{Np}\right) - (N-n) \ln\left(1 + \frac{-(n-Np)}{N(1-p)}\right) && \text{multiplicative inverses}
\end{aligned}$$

$$= -((n-Np)+Np)\ln\left(1+\frac{n-Np}{Np}\right) - (N(1-p)-(n-Np))\ln\left(1+\frac{-(n-Np)}{N(1-p)}\right) \qquad \text{add inverses}$$

$$= -((n-Np)+Np)\left[\frac{(n-Np)}{Np} - \frac{1}{2}\frac{(n-Np)^2}{N^2p^2} + O\left(\frac{(n-Np)^3}{N^3}\right)\right]$$

$$- (N(1-p)-(n-Np))\left[-\frac{(n-Np)}{N(1-p)} - \frac{1}{2}\frac{(n-Np)^2}{N^2(1-p)^2} + O\left(\frac{(n-Np)^3}{N^3}\right)\right] \qquad \text{Maclauren (Taylor) series}$$

$$= -(n-Np)\frac{(n-Np)}{Np} + (n-Np)\frac{1}{2}\frac{(n-Np)^2}{N^2p^2} + O\left(\frac{(n-Np)^3}{N^3}\right)$$

$$- Np\frac{(n-Np)}{Np} + Np\frac{1}{2}\frac{(n-Np)^2}{N^2p^2} + O\left(\frac{(n-Np)^3}{N^3}\right)$$

$$+ N(1-p)\frac{(n-Np)}{N(1-p)} + N(1-p)\frac{1}{2}\frac{(n-Np)^2}{N^2(1-p)^2} + O\left(\frac{(n-Np)^3}{N^3}\right)$$

$$- (n-Np)\frac{(n-Np)}{N(1-p)} - (n-Np)\frac{1}{2}\frac{(n-Np)^2}{N^2(1-p)^2} + O\left(\frac{(n-Np)^3}{N^3}\right) \qquad \text{distributive axiom}$$

$$= -\frac{(n-Np)^2}{Np} + \frac{1}{2}\frac{(n-Np)^3}{N^2p^2} + O\left(\frac{(n-Np)^3}{N^3}\right)$$

$$- (n-Np) + \frac{1}{2}\frac{(n-Np)^2}{Np} + O\left(\frac{(n-Np)^3}{N^3}\right)$$

$$+ (n-Np) + \frac{1}{2}\frac{(n-Np)^2}{N(1-p)} + O\left(\frac{(n-Np)^3}{N^3}\right)$$

$$- \frac{(n-Np)^2}{N(1-p)} - \frac{1}{2}\frac{(n-Np)^3}{N^2(1-p)^2} + O\left(\frac{(n-Np)^3}{N^3}\right) \qquad \text{combine like terms}$$

$$= -\frac{(n-Np)^2(1-p)}{Np(1-p)} + \frac{1}{2}\frac{(n-Np)^3(1-p)^2}{N^2p^2(1-p)^2} + O\left(\frac{(n-Np)^3}{N^3}\right)$$

$$- (n-Np) + \frac{1}{2}\frac{(n-Np)^2(1-p)}{Np(1-p)} + O\left(\frac{(n-Np)^3}{N^3}\right)$$

$$+ (n-Np) + \frac{1}{2}\frac{(n-Np)^2p}{N(1-p)p} + O\left(\frac{(n-Np)^3}{N^3}\right)$$

$$- \frac{(n-Np)^2p}{N(1-p)p} - \frac{1}{2}\frac{(n-Np)^3p^2}{N^2(1-p)^2p^2} + O\left(\frac{(n-Np)^3}{N^3}\right) \qquad \text{multiply inverses}$$

$$= -\frac{1}{2}\frac{(n-Np)^2}{Np(1-p)} + \frac{1}{2}\frac{(n-Np)^3(1-p)^2}{N^2p^2(1-p)^2} - \frac{1}{2}\frac{(n-Np)^3p^2}{N^2(1-p)^2p^2} + O\left(\frac{(n-Np)^3}{N^3}\right) \qquad \text{combine like terms}$$

$$= -\frac{1}{2}\frac{(n-Np)^2}{Np(1-p)} + O\left(\frac{(n-Np)^3}{N^2}\right) \qquad \text{subsume small terms}$$

and substitute back:

$$\lim_{N\to\infty}\binom{N}{n}p^n(1-p)^{N-n} = \lim_{N\to\infty}\sqrt{\frac{2\pi}{(2\pi)^2}\frac{N}{n(N-n)}}\exp\left\{-\frac{1}{2}\frac{(n-Np)^2}{Np(1-p)} + O\left(\frac{(n-Np)^3}{N^2}\right)\right\} \qquad \text{substitution}$$

$$
\begin{aligned}
&\approx \lim_{N\to\infty} \sqrt{\frac{2\pi}{(2\pi)^2}\frac{N}{n(N-n)}}\exp\left\{-\frac{1}{2}\frac{(n-Np)^2}{Np(1-p)}\right\} && \text{within } O\left(\frac{1}{\sqrt{N}}\right)\\
&= \lim_{N\to\infty} \sqrt{\frac{2\pi}{(2\pi)^2}\frac{N}{(Np+n-Np)(N-Np-n+Np)}}\exp\left\{-\frac{1}{2}\frac{(n-Np)^2}{Np(1-p)}\right\} && \text{add. inv.}\\
&= \lim_{N\to\infty} \left(\sqrt{\frac{2\pi}{(2\pi)^2}\frac{N}{Np(N-Np)}}\left[1+O\left(\frac{(n-Np)}{N}\right)\right]\right)\exp\left\{-\frac{1}{2}\frac{(n-Np)^2}{Np(1-p)}\right\} && \begin{array}{r}\text{differ by}\\ \pm(n-Np)\end{array}\\
&\approx \lim_{N\to\infty} \sqrt{\frac{2\pi}{(2\pi)^2}\frac{N}{Np(N-Np)}}\exp\left\{-\frac{1}{2}\frac{(n-Np)^2}{Np(1-p)}\right\} && \text{within } O\left(\frac{1}{\sqrt{N}}\right)\\
&= \lim_{N\to\infty} \sqrt{\frac{1}{2\pi Np(1-p)}}\exp\left\{-\frac{1}{2}\frac{(n-Np)^2}{Np(1-p)}\right\} && \text{mult. inverses}
\end{aligned}
$$

This is a normal or Gaussian with mean $\mu = Np$ and variance $\sigma^2 = Np(1-p)$.

9.2 Conjugacy for Gaussians

Gaussians are conjugate to Normal-Gamma priors (here $\tau = \frac{1}{\sigma^2}$):

$$
\begin{aligned}
\text{NormalGamma}_{\mu_0,\lambda,\alpha,\beta}(\tau,\mu) &= \overbrace{\frac{\beta^\alpha \tau^{\alpha-1}}{\Gamma(\alpha)}\exp(-\tau\beta)}^{\text{Gamma distrib. over } \tau}\ \overbrace{\frac{\sqrt{\lambda\tau}}{\sqrt{2\pi}}\exp\left(-\frac{\tau\lambda(\mu-\mu_0)^2}{2}\right)}^{\text{Normal distrib. over } \mu}\\
&= \frac{\beta^\alpha\sqrt{\lambda}}{\Gamma(\alpha)\sqrt{2\pi}}\tau^{\alpha-\frac{1}{2}}\exp\left(-\tau\beta-\frac{\tau\lambda(\mu-\mu_0)^2}{2}\right) && \text{product of exponentials}
\end{aligned}
$$

So the posterior has the same form as the prior:

$$
\begin{aligned}
\mathsf{P}(\tau,\mu\,|\,\mathcal{D}) &= \frac{\mathsf{P}(\tau,\mu)\cdot\mathsf{P}(\mathcal{D}\,|\,\tau,\mu)}{\mathsf{P}(\mathcal{D})} && \text{Bayes' rule}\\
&\propto \mathsf{P}(\tau,\mu)\cdot\mathsf{P}(\mathcal{D}\,|\,\tau,\mu) && \text{def. proportion}\\
&= \overbrace{\frac{\beta^\alpha\sqrt{\lambda}}{\Gamma(\alpha)\sqrt{2\pi}}\tau^{\alpha-\frac{1}{2}}\exp\left(-\tau\beta-\frac{\tau\lambda(\mu-\mu_0)^2}{2}\right)}^{\text{prior}}\ \overbrace{\prod_x\frac{\sqrt{\tau}}{\sqrt{2\pi}}\exp\left(-\frac{\tau(x-\mu)^2}{2}\right)}^{\text{likelihood}} && \text{substitute}\\
&\propto \tau^{\alpha-\frac{1}{2}}\exp\left(-\tau\beta-\frac{\tau\lambda(\mu-\mu_0)^2}{2}\right)\prod_x\sqrt{\tau}\exp\left(-\frac{\tau(x-\mu)^2}{2}\right) && \text{def. proportion}\\
&= \tau^{\alpha-\frac{1}{2}+\frac{N}{2}}\exp\left(-\tau\beta-\tau\frac{\lambda(\mu-\mu_0)^2}{2}-\tau\frac{\sum_x(x-\mu)^2}{2}\right) && \text{product of exps}\\
&= \tau^{\alpha-\frac{1}{2}+\frac{N}{2}}\exp\left(-\tau\beta-\tau\frac{\lambda(\mu^2-2\mu\mu_0+{\mu_0}^2)}{2}-\tau\frac{\sum_x(x^2-2x\mu+\mu^2)}{2}\right) && \text{expand square}
\end{aligned}
$$

$$= \tau^{\alpha-\frac{1}{2}+\frac{N}{2}} \exp\left(-\tau\beta - \tau\frac{\lambda\mu^2 - 2\lambda\mu\mu_0 + \lambda{\mu_0}^2}{2} - \tau\frac{\sum_x x^2 - 2\mu\sum_x x + N\mu^2}{2}\right) \quad \text{distrib. axiom}$$

$$= \tau^{\alpha-\frac{1}{2}+\frac{N}{2}} \exp\left(-\tau\beta - \tau\frac{\lambda\mu^2 - 2\lambda\mu\mu_0 + \lambda{\mu_0}^2 + \sum_x x^2 - 2\mu\sum_x x + N\mu^2}{2}\right) \quad \text{distrib. axiom}$$

$$= \tau^{\alpha-\frac{1}{2}+\frac{N}{2}} \exp\left(-\tau\beta - \tau\frac{\lambda\mu^2 - 2\lambda\mu\mu_0 + \lambda{\mu_0}^2 + \sum_x x^2 - 2\mu N\bar{x} + N\mu^2}{2}\right) \quad \text{subst } \bar{x} = \frac{\sum_x x}{N}$$

$$= \tau^{\alpha-\frac{1}{2}+\frac{N}{2}} \exp\left(-\tau\beta - \tau\frac{(\lambda+N)\mu^2 - 2(\lambda\mu_0+N\bar{x})\mu + \sum_x x^2 + \lambda{\mu_0}^2}{2}\right) \quad \text{distrib. axiom}$$

$$= \tau^{\alpha-\frac{1}{2}+\frac{N}{2}} \exp\left(-\tau\beta - \tau\frac{(\lambda+N)\left(\mu^2 - 2\frac{\lambda\mu_0+N\bar{x}}{\lambda+N}\mu\right) + \sum_x x^2 + \lambda{\mu_0}^2}{2}\right) \quad \text{distrib. axiom}$$

$$= \tau^{\alpha-\frac{1}{2}+\frac{N}{2}} \exp\left(-\tau\beta - \tau\frac{(\lambda+N)\left(\mu^2 - 2\frac{\lambda\mu_0+N\bar{x}}{\lambda+N}\mu + \frac{\lambda\mu_0+N\bar{x}}{\lambda+N}^2 - \frac{\lambda\mu_0+N\bar{x}}{\lambda+N}^2\right) + \sum_x x^2 + \lambda{\mu_0}^2}{2}\right) \quad \text{add inverses}$$

$$= \tau^{\alpha-\frac{1}{2}+\frac{N}{2}} \exp\left(-\tau\beta - \tau\frac{(\lambda+N)\left(\left(\mu - \frac{\lambda\mu_0+N\bar{x}}{\lambda+N}\right)^2 - \frac{\lambda\mu_0+N\bar{x}}{\lambda+N}^2\right) + \sum_x x^2 + \lambda{\mu_0}^2}{2}\right) \quad \text{complete square}$$

$$= \tau^{\alpha-\frac{1}{2}+\frac{N}{2}} \exp\left(-\tau\beta - \tau\frac{(\lambda+N)\left(\mu - \frac{\lambda\mu_0+N\bar{x}}{\lambda+N}\right)^2 - \frac{(\lambda\mu_0+N\bar{x})^2}{\lambda+N} + \sum_x x^2 + \lambda{\mu_0}^2}{2}\right) \quad \text{distrib. axiom}$$

$$= \text{NormalGamma}_{\frac{\lambda\mu_0+N\bar{x}}{\lambda+N},\ \lambda+N,\ \alpha+\frac{N}{2},\ \beta+\frac{1}{2}\left(-\frac{(\lambda\mu_0+N\bar{x})^2}{\lambda+N}+\sum_x x^2+\lambda{\mu_0}^2\right)}(\tau,\mu) \quad \text{def. of NG}$$

$$= \text{NormalGamma}_{\frac{\lambda\mu_0+N\bar{x}}{\lambda+N},\ \lambda+N,\ \alpha+\frac{N}{2},\ \beta+\frac{1}{2}\left(-\frac{(\lambda\mu_0+N\bar{x})^2}{\lambda+N}+\frac{(\lambda+N)\lambda{\mu_0}^2}{\lambda+N}+\sum_x x^2\right)}(\tau,\mu) \quad \text{mult. inverses}$$

$$= \text{NormalGamma}_{\frac{\lambda\mu_0+N\bar{x}}{\lambda+N},\ \lambda+N,\ \alpha+\frac{N}{2},\ \beta+\frac{1}{2}\left(\frac{-(\lambda\mu_0)^2-2\lambda\mu_0 N\bar{x}-(N\bar{x})^2+\lambda\lambda{\mu_0}^2+N\lambda{\mu_0}^2}{\lambda+N}+\sum_x x^2\right)}(\tau,\mu) \quad \text{distrib. axiom}$$

$$= \text{NormalGamma}_{\frac{\lambda\mu_0+N\bar{x}}{\lambda+N},\ \lambda+N,\ \alpha+\frac{N}{2},\ \beta+\frac{1}{2}\left(\frac{-2\lambda\mu_0 N\bar{x}-(N\bar{x})^2+N\lambda{\mu_0}^2}{\lambda+N}+\sum_x x^2\right)}(\tau,\mu) \quad \text{sum inverses}$$

$$= \text{NormalGamma}_{\frac{\lambda\mu_0+N\bar{x}}{\lambda+N},\ \lambda+N,\ \alpha+\frac{N}{2},\ \beta+\frac{1}{2}\left(\frac{N\lambda{\mu_0}^2-2\lambda\mu_0 N\bar{x}-(N\bar{x})^2}{\lambda+N}+\sum_x x^2\right)}(\tau,\mu) \quad \text{comm. axiom}$$

$$= \text{NormalGamma}_{\frac{\lambda\mu_0+N\bar{x}}{\lambda+N},\ \lambda+N,\ \alpha+\frac{N}{2},\ \beta+\frac{1}{2}\left(\frac{N\lambda{\mu_0}^2-2\lambda\mu_0 N\bar{x}+\lambda N\bar{x}^2-\lambda N\bar{x}^2-(N\bar{x})^2}{\lambda+N}+\sum_x x^2\right)}(\tau,\mu) \quad \text{add inverses}$$

$$= \text{NormalGamma}_{\frac{\lambda\mu_0+N\bar{x}}{\lambda+N},\ \lambda+N,\ \alpha+\frac{N}{2},\ \beta+\frac{1}{2}\left(\frac{N\lambda(\mu_0-\bar{x})^2-\lambda N\bar{x}^2-(N\bar{x})^2}{\lambda+N}+\sum_x x^2\right)}(\tau,\mu) \quad \text{complete square}$$

$$= \text{NormalGamma}_{\frac{\lambda\mu_0+N\bar{x}}{\lambda+N},\ \lambda+N,\ \alpha+\frac{N}{2},\ \beta+\frac{1}{2}\left(\frac{N\lambda(\mu_0-\bar{x})^2}{\lambda+N}-N\bar{x}^2+\sum_x x^2\right)}(\tau,\mu) \quad \text{distrib. axiom}$$

$$= \text{NormalGamma}_{\frac{\lambda\mu_0+N\bar{x}}{\lambda+N},\ \lambda+N,\ \alpha+\frac{N}{2},\ \beta+\frac{1}{2}\left(\frac{N\lambda(\mu_0-\bar{x})^2}{\lambda+N}+\sum_x(x^2-\bar{x}^2)\right)}(\tau,\mu) \quad \text{sum over indep. var}$$

$$= \text{NormalGamma}_{\frac{\lambda\mu_0+N\bar{x}}{\lambda+N},\ \lambda+N,\ \alpha+\frac{N}{2},\ \beta+\frac{1}{2}\left(\frac{N\lambda(\mu_0-\bar{x})^2}{\lambda+N}+\sum_x(x^2-2x\bar{x}+\bar{x}^2)\right)}(\tau,\mu) \quad \sum_x x = \sum_x \bar{x}$$

$$= \text{NormalGamma}_{\frac{\lambda\mu_0+N\bar{x}}{\lambda+N},\ \lambda+N,\ \alpha+\frac{N}{2},\ \beta+\frac{1}{2}\left(\frac{N\lambda(\mu_0-\bar{x})^2}{\lambda+N}+\sum_x(x-\bar{x})^2\right)}(\tau,\mu) \quad \text{complete square}$$

(The extra 'NormalGamma' steps aren't necessary to show conjugacy, just to clean up parameters.)

9.3 Multivariate normal or Gaussians

For mean vector $\boldsymbol{\mu} \in \mathbb{R}^V$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{V \times V}$:

$$\mathcal{N}_{\mu,\Sigma}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^V |\boldsymbol{\Sigma}|}} \exp\left\{ -\frac{1}{2} \overbrace{(\mathbf{x}-\boldsymbol{\mu})^\top \underbrace{\boldsymbol{\Sigma}^{-1}}_{\text{concentration/precision matrix}} (\mathbf{x}-\boldsymbol{\mu})}^{\text{square of Mahalanobis distance (number of standard deviations)}} \right\}$$

For observations $\mathbf{X} \in \mathbb{R}^{N \times V}$ (usually samples are rows, variables are columns — as in our .csv's):

- the **mean vector** can be estimated: $\boldsymbol{\mu} = \frac{1}{N} \overbrace{\mathbf{X}^\top \mathbf{1}^N}^{\text{sum over items}}$; and
- the **covariance matrix** can be estimated: $\boldsymbol{\Sigma} = \overbrace{\mathbf{X'}^\top \mathbf{X'}}^{\text{sum over items of each pair of variable values}}$ from centered data: $\mathbf{X'} = \mathbf{X} - \overbrace{\mathbf{1}^N \boldsymbol{\mu}^\top}^{\text{'broadcast' mean to all items}}$.

The **concentration matrix** $\boldsymbol{\Sigma}^{-1}$ is the inverse of the covariance matrix $\boldsymbol{\Sigma}$.

Now decompose $\mathbf{X'}$ via orthogonal (all-perpendicular) transforms $\mathbf{U}$ to a space with unit variance:

$$\mathbf{X'} = \text{diag}(\mathbf{s})\,\mathbf{U}$$

Since these are orthogonal, the inverse matrices give us the form of a distance metric:

$$\begin{aligned}
(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu}) &= (\mathbf{x}-\boldsymbol{\mu})^\top (\mathbf{X'}^\top \mathbf{X'})^{-1} (\mathbf{x}-\boldsymbol{\mu}) && \text{substitution} \\
&= (\mathbf{x}-\boldsymbol{\mu})^\top ((\text{diag}(\mathbf{s})\,\mathbf{U})^\top \text{diag}(\mathbf{s})\,\mathbf{U})^{-1} (\mathbf{x}-\boldsymbol{\mu}) && \text{substitution} \\
&= (\mathbf{x}-\boldsymbol{\mu})^\top \left(\mathbf{U}^\top \text{diag}\left(\mathbf{s}^2\right) \mathbf{U}\right)^{-1} (\mathbf{x}-\boldsymbol{\mu}) && \text{transp of prod is reverse prod of transps} \\
&= (\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{U}^{-1} \text{diag}\left(\frac{1}{\mathbf{s}^2}\right) (\mathbf{U}^\top)^{-1} (\mathbf{x}-\boldsymbol{\mu}) && \text{inv of prod is reverse prod of invs} \\
&= (\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{U}^\top \text{diag}\left(\frac{1}{\mathbf{s}^2}\right) \mathbf{U}\,(\mathbf{x}-\boldsymbol{\mu}) && \text{inv of orthogonal matrix is transp} \\
&= \left(\text{diag}\left(\frac{1}{\mathbf{s}}\right) \mathbf{U}\,(\mathbf{x}-\boldsymbol{\mu})\right)^\top \text{diag}\left(\frac{1}{\mathbf{s}}\right) \mathbf{U}\,(\mathbf{x}-\boldsymbol{\mu}) && \text{transp of prod is reverse prod of transp} \\
&= \sum_v \left(\text{diag}\left(\frac{1}{\mathbf{s}}\right) \mathbf{U}\,(\mathbf{x}-\boldsymbol{\mu})\right)_{[v]} \cdot \left(\text{diag}\left(\frac{1}{\mathbf{s}}\right) \mathbf{U}\,(\mathbf{x}-\boldsymbol{\mu})\right)_{[v]} && \text{definition of inner product}
\end{aligned}$$

So a concentration matrix gives a **distance metric** in orthonormal space (dist. in std. deviations).