Lecture 3 Gaussians

Outline:

1. Gaussian Priors, Gaussian Likelihoods
2. Central Limit Thm
3. Multivariate extension

1/ Task: Measure T in room

Meas noise $\sigma = 1K$

$y = x + n$

meas T (y), true T (x)

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

$\mathbb{E}[x] = \mu$

$Var[x] = \sigma^2$

Interestingly, the mean & mode & median are all the same

$$\Rightarrow p(y|x,N) = \mathcal{N}(y|x, N = 1^2K^2)$$

Prior knowledge: Rooms usually have $x_0 = 295K \pm 3K$

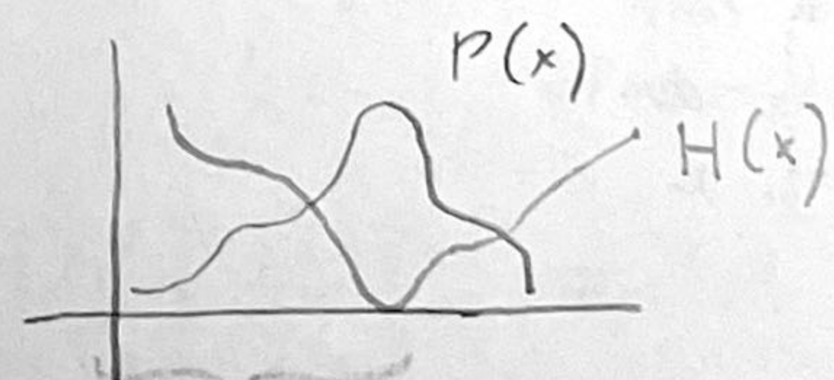
$$P(x) = \mathcal{N}(x|x_0, T = 9K^2)$$

Want to find $P(\underset{\text{true}}{x}|\underset{\text{meas}}{y}) = \frac{P(y|x)P(x)}{P(y)}$ (Bayes thm)

Information

To make the maths easier, instead of working w/ the likelihood directly, we'll work w/ the -ve log likelihood, $H(x)$

	Prob	-ve log likelihood
	$P(x) = e^{-H(x)}$	$H(x) :=$
chain rule	$P(x,y) = P(x$	$H($
Bayes rule	$P(x\|y) = \frac{P(y\|x)P(x)}{P(y)}$	$H(x\|y) = H(y\|x) + H(x) \underbrace{- H(y)}$



We are interested in the POSTERIOR on x

The evidence is just the NORMALIZATION & won't impact the shape of the posterior
↳ we'll ignore it (!)

Notation: $H(x) + H_0 \mathrel{\hat{=}} H(x)$

up to some const

$$H(y|x) = -\ln P(y|x)$$

$$= -\ln \frac{1}{\sqrt{2\pi N}} e^{-\frac{1}{2}\frac{(y-x)^2}{N}}$$

$$= \frac{1}{2}\ln(2\pi N) + \frac{1}{2}\frac{(y-x)^2}{N}$$

$$\hat{=} \frac{1}{2}\frac{(y-x)^2}{N}$$

Prior

$$H(x) = -\ln P(x) \hat{=} \frac{1}{2}\frac{(x-x_0)^2}{T}$$

Putting it together:

$$H(x|y) = H(y|x) + H(x) - H(y)$$

$$\hat{=} H(y|x) + H(x)$$

$$= \frac{1}{2}\frac{(y-x)^2}{N} + \frac{1}{2}\frac{(x-x_0)^2}{T}$$

Can we massage this into the same functional form as a Gaussian? (for x)

→ Method: Complete the square(!)

$$\hat{=} \frac{1}{2}\cancel{\frac{y^2}{N}} - \frac{yx}{N} + \frac{1}{2}\frac{x^2}{N} + \frac{1}{2}\frac{x^2}{T} - \frac{xx_0}{T} + \cancel{\frac{1}{2}\frac{x_0^2}{T}}$$

Dropping const terms that don't depend on x

$$= \frac{1}{2}\underbrace{\left(\frac{1}{N}+\frac{1}{T}\right)}_{\text{Let } \frac{1}{D}} x^2 - \left(\frac{y}{N}+\frac{x_0}{T}\right)x = \frac{1}{2}\underbrace{\left(\frac{1}{N}+\frac{1}{T}\right)}_{1/D}\left\{x^2 - 2\underbrace{\frac{\left(\frac{y}{N}+\frac{x_0}{T}\right)}{\frac{1}{N}+\frac{1}{T}}}_{m} x\right\}$$

$$\hat{=} \frac{1}{2D}(x-m)^2 + \cancel{\text{const}}$$

∴ The posterior also follows a Gaussian dist w/ mean

$$m = \frac{\frac{y}{N}+\frac{x_0}{T}}{\frac{1}{N}+\frac{1}{T}} = \frac{1/N}{\frac{1}{N}+\frac{1}{T}}\, y + \frac{1/T}{\frac{1}{N}+\frac{1}{T}}\, x_0$$

variance:

$$D = \frac{1}{\frac{1}{N}+\frac{1}{T}}$$

Nb: Error prop, the reciprocals of the std dev Σ in quadrature

$$\frac{1}{\sigma^2_{\text{posterior}}} = \frac{1}{\sigma^2_{\text{likelihood}}} + \frac{1}{\sigma^2_{\text{prior}}}$$

Where are we now?

Prior	Likelihood	Posterior	
Beta	Bernoulli	Beta	(last lecture)
Gamma	Poisson	Gamma	(tutorial)
★ Gaussian	Gaussian	Gaussian	

We'll implement this ex in the tutorial later(!)

2/ Central Limit Thm

"independent & identically distributed"

Consider a sequence of N iid random variables x_i w/ finite mean μ & variance σ^2 of any distribution

$$\bar{x}_N := \frac{1}{N}\sum_{i=1}^{N} x_i$$

For $N \to \infty$

Mean μ

$$P(\bar{x}_N) \to \frac{\sqrt{N}}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2}\frac{(\bar{x}_N - \mu)^2}{\sigma^2/N}\right]$$

Proof in §7.4 of Wackerly et. al.

Variance σ^2/N

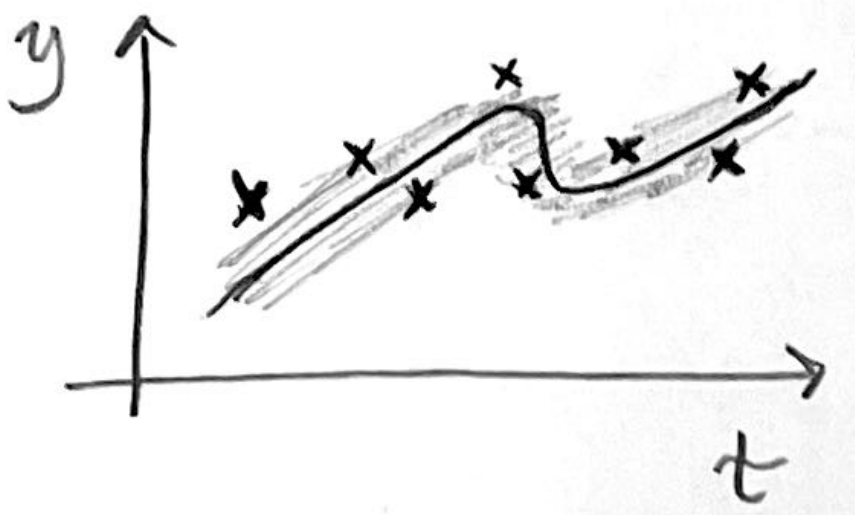
What does this MEAN?

→ More samples ($N \to \infty$) the distribution becomes narrower

Linear Regression

Model: $y_i = x_i + n_i$ $\leftarrow$ Degree of the polynomial

$$x_i = f(t_i) \approx \sum_{j=0}^{d} t_i^j a_j$$



Putting this together $\rightarrow$ $y_i = \sum_{j=0}^{M} t_i^j a_j + n_i$ (1) $\leftarrow$ Data points

N: # data points

Let's summarize more succinctly in matrix notation:

$$\vec{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix} \quad \vec{t} = \begin{pmatrix} t_1 \\ \vdots \\ t_N \end{pmatrix} \quad \vec{n} = \begin{pmatrix} n_1 \\ \vdots \\ n_N \end{pmatrix} ; \quad a$$

Want to repn powers of t as a matrix

Let $R_{ij} = t_i^j$

$$\underline{\underline{R}} = \overbrace{\begin{pmatrix} 1 & t_1 & t_1^2 & \cdots \\ 1 & \vdots & & \\ & \vdots & & \\ & t_d & t_d^2 & \cdots \end{pmatrix}}^{d+1 \,:\, \text{degree of poly}} \Bigg\} \text{N: \# data points} \rightarrow \text{Design Matrix}$$

So writing ⊛ more succinctly in matrix notation:

$$\vec{y} = \underline{R}\,\vec{a} + \vec{n}$$

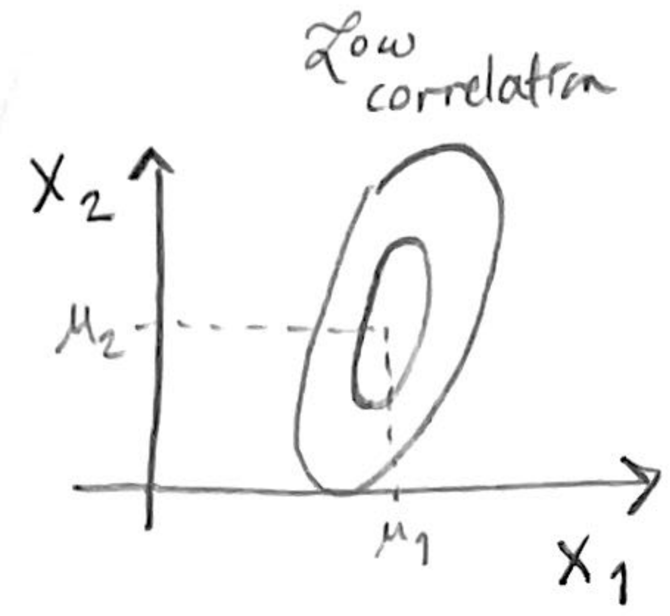
Multivariate Gaussian

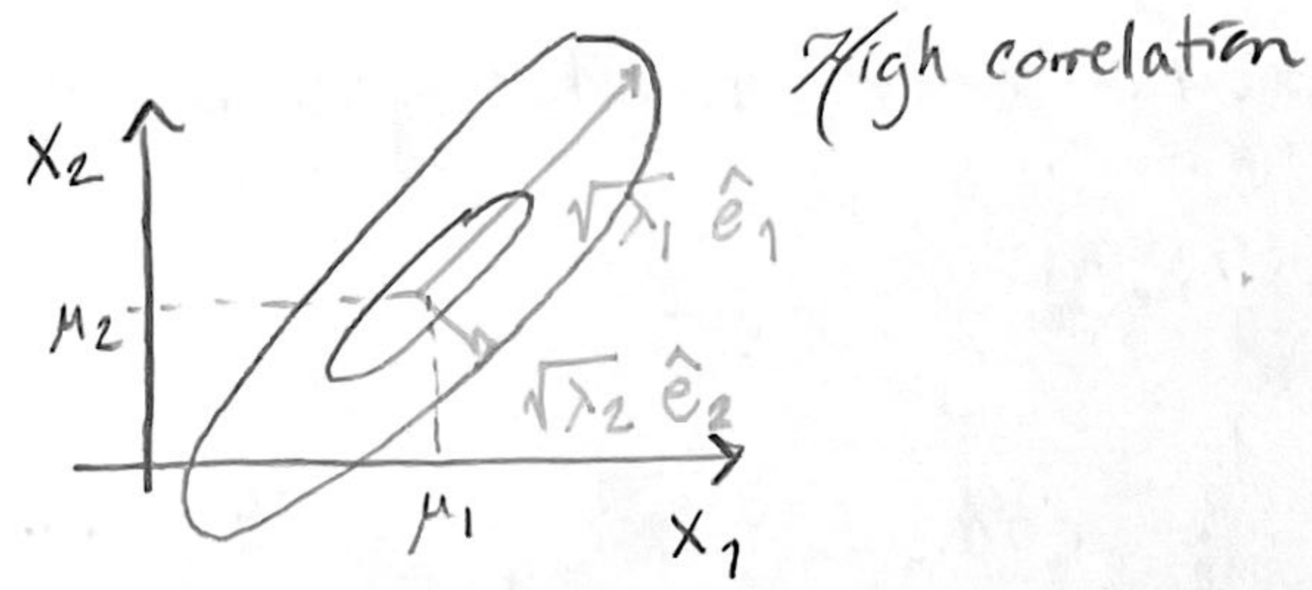
$$P(\vec{x}\,|\,\vec{\mu}, \Sigma) = \mathcal{N}(\vec{x}\,|\,\vec{\mu}, \Sigma) = \frac{1}{|2\pi\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right]$$

$$\mathbb{E}[\vec{x}] = \vec{\mu}$$

$$\text{Cov}(\vec{x}) = \Sigma = \mathbb{E}[(x-\mu)(x-\mu)^T]$$

$$\underbrace{(\,)}_{N\times 1} \underbrace{(\,)}_{1\times N} \rightarrow \text{Cov is an } N\times N \text{ matrix}$$





The eigenvalues of the covariance characterize the variation along the corresponding eigendirections

→ See PCA method from ODSL ML block course!

So back to our modelling ex... where are all the model params backed in?
→ Polynomial coeff $\underline{a}$. Let's solve for $P(a|y)$!

assumption: Gaussian, uncorrelated noise (independent)

$$\mathcal{L}(y|a) = \mathcal{N}(Ra, N)$$

Working thru the same maths we started w/ @ beginning of lecture, now in multi-dim!

Information

$$\hookrightarrow H(y|a) \hat{=} \frac{1}{2}(y - Ra)^T N^{-1} (y - Ra)$$

Prior

$$P(a) = \mathcal{N}(a|0, A) \leftarrow \text{ad hoc assumption}$$

$$H(a) \hat{=} \frac{1}{2} a^T A^{-1} a$$

$$H(a|y) \hat{=} H(y|a) + H(a) = \frac{1}{2}(y - Ra)^T N^{-1}(y - Ra) + \frac{1}{2} a^T a$$

→ Quadratic in a^2

again, solve by completing the square(!)

$$H(a|y) \hat{=} \cancel{\frac{1}{2} y^T N^{-1} y} + \frac{1}{2} a^T R^T N^{-1} R a - a^T \overbrace{R^T N^{-1} y}^{j} + \frac{1}{2} a^T a$$

again, drop constants

Nb: D is a sym matrix, $D = D^T$

$$\hat{=} \frac{1}{2} a^T (\underbrace{R^T N^{-1} R + A^{-1}}_{D^{-1}}) a - a^T R^T N^{-1} y$$

$$= \frac{1}{2} a^T D^{-1} a - a^T \mu$$

$$\hat{=} \frac{1}{2} a^T D^{-1} a - a^T j$$

$$\hat{=} \frac{1}{2} a^T \underline{D^{-1}} a - a^T \underline{D^{-1}} \underbrace{D j}_{\mu} + \frac{1}{2} j^T D D^{-1} D j - \cancel{\frac{1}{2} j^T D D^{-1} D j}$$

$$= \frac{1}{2} a^T D^{-1} a - a^T D^{-1} \mu + \frac{1}{2} \mu^T D^{-1} \mu$$

$$= \frac{1}{2}(a - \mu)^T D^{-1} (a - \mu)$$

We found the posterior over model params

$$P(a|y) \propto \exp -\frac{1}{2}$$

$$D = \left(R^T \underset{N\times N}{N^{-1}} \underset{N\times M}{R} + A^{-1}\right)^{-1}$$ posterior variance

$$\vec{\mu} = D R^T N^{-1} \vec{y}$$ posterior mean

Predictive Distribution

Given: dataset (x,y) & posterior $P(a|y)$

What's the prob of y' @ NEW eval point x'?

$$P(y'|y) = \int da \, P(y'|a) \, P(a|y)$$