Probabilistic models for neural data:
From single neurons to population dynamics

NEUROBIO 316QC

Jan Drugowitsch

jan_drugowitsch@hms.harvard.edu

Session 2: Gaussians & Linear models

Today

Q&A about previous session

Discuss assignment (~20min)

Gaussian distribution, linear models & model comparison (remaining time)

Note
no course on March 16
last session on March 30

Overview

Univariate and multivariate Gaussian distributions

- Means and variances of linearly transformed random variables (general)
- Univariate standard Gaussian and linear transformations thereof
- Multivariate Gaussians through linear constructions
- Probabilistic operations: marginalization and conditioning
- Maximum likelihood parameter estimation
- Priors as added observations / regularizers

Linear models

- Maximum likelihood and least squares estimates
- What is linear in linear models?
- Priors and Bayesian inference

Bayesian model comparison

- Trading off goodness-of-fit with model complexity
- Adjusting model complexity through hyperpriors
- Evidence approximation

Overview

Univariate and multivariate Gaussian distributions

- Means and variances of linearly transformed random variables (general)
- Univariate standard Gaussian and linear transformations thereof
- Multivariate Gaussians through linear constructions
- Probabilistic operations: marginalization and conditioning
- Maximum likelihood parameter estimation
- Priors as added observations / regularizers

Linear models

- Maximum likelihood and least squares estimates
- What is linear in linear models?
- Priors and Bayesian inference

Bayesian model comparison

- Trading off goodness-of-fit with model complexity
- Adjusting model complexity through hyperpriors
- Evidence approximation

Mean and variance of linearly transformed RVs

Linear transformation of arbitrary single random variable z

Scaling

$$x = az$$

$$\mathrm{E}[x] = \int x\, p(x)\, \mathrm{d}x = \int \overbrace{az}^{x}\, \overbrace{p(z) \left|\frac{\mathrm{d}z}{\mathrm{d}x}\right|}^{p(x)}\, \overbrace{\frac{\mathrm{d}x}{\mathrm{d}z}\mathrm{d}z}^{\mathrm{d}x} = a \int z\, p(z) \frac{1}{a} a\, \mathrm{d}z = a\, \mathrm{E}[z]$$

$$\mathrm{var}[x] = a^2 \mathrm{var}[z]$$

Shifting & scaling

$$x = az + b \qquad E[x] = a\, E[z] + b$$

$$\mathrm{var}[x] = a^2 \mathrm{var}[z]$$

Linear transformation of arbitrary random variable pair, z_1 and z_2

$$x = a_1 z_1 + a_2 z_2 + b$$

$$\mathrm{E}[x] = a_1 \mathrm{E}[z_1] + a_2 \mathrm{E}[z_2] + b$$

$$var[x] = a_1^2\, var[z_1] + a_2^2\, var[z_2] + 2 a_1 a_2\, cov[z_1, z_2]$$

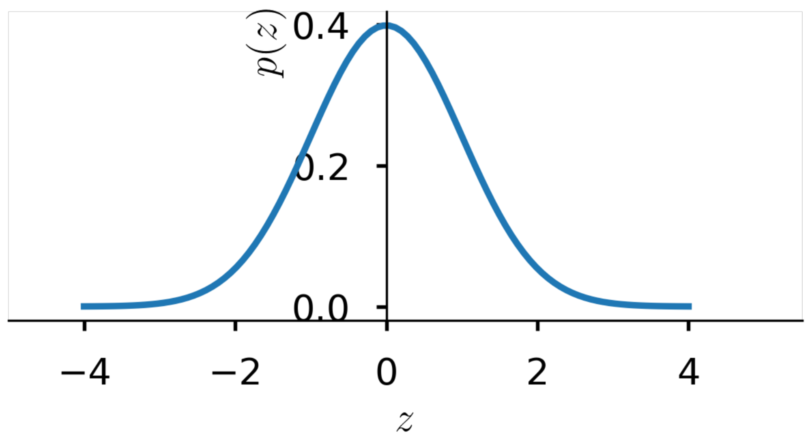
The univariate standard Gaussian

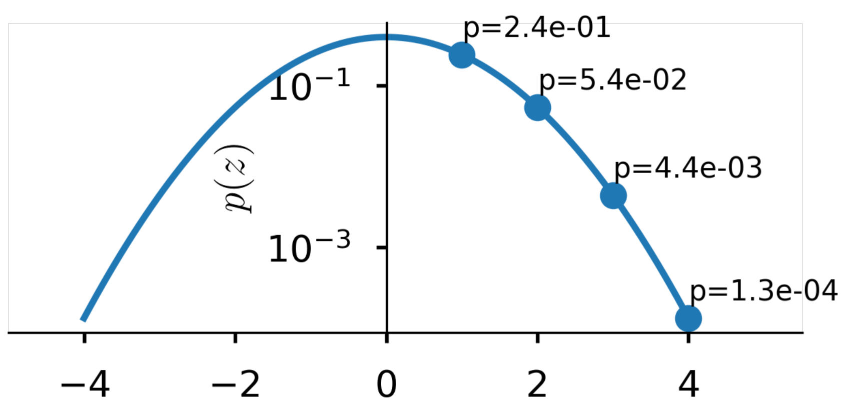
Standard Gaussian: zero mean, unit variance

mean variance

$$p(z) = \mathrm{N}(z|0,1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \propto e^{-\frac{1}{2}z^2}$$

$$\log p(z) = -\frac{1}{2}z^2 - \frac{1}{2}\log 2\pi = -\frac{1}{2}z^2 + \mathrm{const.}$$

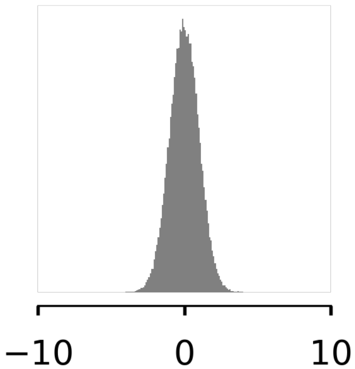




Rapidly dropping probabilities in tails:
considers outliers very unlikely
(sensitivity to outliers)

Linear transformations of univariate Gaussians

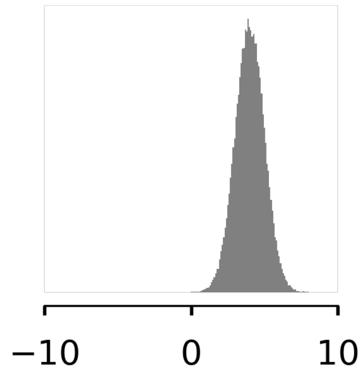
$z \sim \mathrm{N}(0,1)$



$\mathrm{E}[z] = 0$
$\mathrm{var}[z] = 1$

Changing mean

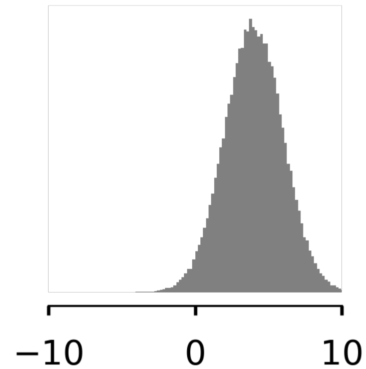
$x = z + \mu$



$\mathrm{E}[x] = \mathrm{E}[z] + \mu = \mu$
$\mathrm{var}[x] = \mathrm{var}[z] = 1$

Changing mean & variance

$x = az + \mu$



$\mathrm{E}[x] = \mathrm{aE}[z] + \mu = \mu$
$\mathrm{var}[x] = \mathrm{a}^2\mathrm{var}[z] = a^2$

Linear transformation of Gaussian remains Gaussian

$x = f(z) = az + \mu$

$f'(z) = a \qquad z = \frac{x-\mu}{a}$

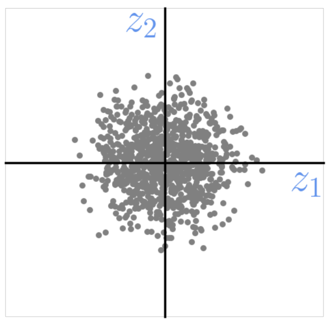
$$p(x) = p(z)\left|\frac{1}{f'(z)}\right| = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}\frac{1}{a} = \frac{1}{\sqrt{2\pi a^2}}e^{-\frac{(x-\mu)^2}{2a^2}} = \mathrm{N}(x|\mu, a^2)$$

In general

$$\left.\begin{array}{l} z \sim \mathrm{N}(\mu_z, \sigma_z^2) \\ x = az + b \end{array}\right\} \left.\begin{array}{l} \mathrm{E}[x] = a\mathrm{E}[z] + b = a\mu_z + b \\ \mathrm{var}[x] = a^2\mathrm{var}[z] = a^2\sigma_z^2 \end{array}\right\} \quad x \sim N(a\mu_z + b, a^2\sigma_z^2)$$

Constructing multivariate Gaussians

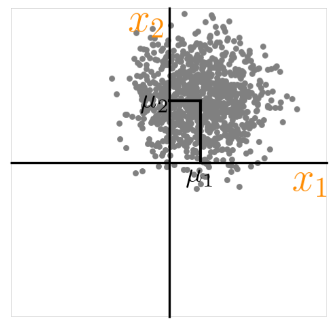
$$\left.\begin{array}{l} z_1 \sim \mathrm{N}(0,1) \\ z_2 \sim \mathrm{N}(0,1) \end{array}\right\} \boldsymbol{z} \sim \mathrm{N}(0, \boldsymbol{I})$$



$$\mathrm{E}[\boldsymbol{z}] = \mathbf{0}$$
$$\mathrm{cov}[\boldsymbol{z}] = \boldsymbol{I}$$

Changing mean

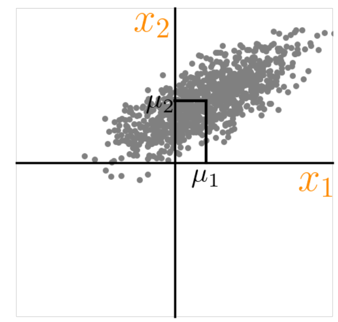
$$\left.\begin{array}{l} x_1 = z_1 + \mu_1 \\ x_2 = z_2 + \mu_2 \end{array}\right\} \boldsymbol{x} = \boldsymbol{z} + \boldsymbol{\mu}$$



$$\mathrm{E}[\boldsymbol{x}] = \mathrm{E}[\boldsymbol{z}] + \boldsymbol{\mu} = \boldsymbol{\mu}$$
$$\mathrm{cov}[\boldsymbol{x}] = \mathrm{cov}[\boldsymbol{z}] = \boldsymbol{I}$$

Changing mean & variance

$$\left.\begin{array}{l} x_1 = a_{11} z_1 + a_{12} z_2 + \mu_1 \\ x_2 = a_{21} z_1 + a_{22} z_2 + \mu_2 \end{array}\right\} \boldsymbol{x} = \boldsymbol{A}\boldsymbol{z} + \boldsymbol{\mu}$$



$$\mathrm{E}[\boldsymbol{x}] = \boldsymbol{A}\mathrm{E}[\boldsymbol{z}] + \boldsymbol{\mu} = \boldsymbol{\mu}$$
$$\mathrm{cov}[\boldsymbol{x}] = \text{see below} = \mathbf{A}\mathbf{A}^{\mathrm{T}}$$

cov$[\boldsymbol{x}]$ when changing mean & variance

$$\boldsymbol{x} - E[\boldsymbol{x}] = \boldsymbol{A}\boldsymbol{z} \qquad \mathrm{E}[\boldsymbol{z}\boldsymbol{z}^T] = \mathrm{cov}[\boldsymbol{z}]$$

$$\mathrm{cov}[\boldsymbol{x}] = \mathrm{E}[(\boldsymbol{x} - \mathrm{E}[\boldsymbol{x}])(\boldsymbol{x} - \mathrm{E}[\boldsymbol{x}])^T] = \mathrm{E}[\boldsymbol{A}\boldsymbol{z}\boldsymbol{z}^T\boldsymbol{A}^{\boldsymbol{T}}] = \boldsymbol{A}\mathrm{E}[\boldsymbol{z}\boldsymbol{z}^T]\boldsymbol{A}^T = \boldsymbol{A}\mathrm{cov}[\boldsymbol{z}]\boldsymbol{A}^T = \boldsymbol{A}\boldsymbol{A}^T$$

Also, **linear transformation of multivariate Gaussian remains Gaussian** (not shown)

In general

$$\left.\begin{array}{l} \boldsymbol{z} \sim \mathrm{N}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z) \\ \boldsymbol{x} = \boldsymbol{A}\boldsymbol{z} + \boldsymbol{\mu} \end{array}\right\} \quad \left.\begin{array}{l} \mathrm{E}[\boldsymbol{x}] = \boldsymbol{A}\mathrm{E}[\boldsymbol{z}] + \boldsymbol{\mu} = \boldsymbol{A}\boldsymbol{\mu}_z + \boldsymbol{\mu} \\ \mathrm{cov}[\boldsymbol{x}] = \boldsymbol{A}\mathrm{cov}[\boldsymbol{z}]\boldsymbol{A}^T = \boldsymbol{A}\boldsymbol{\Sigma}_z\boldsymbol{A}^T \end{array}\right\} \quad \boldsymbol{x} \sim N(\boldsymbol{A}\boldsymbol{\mu}_z + \boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}_z\boldsymbol{A}^T)$$

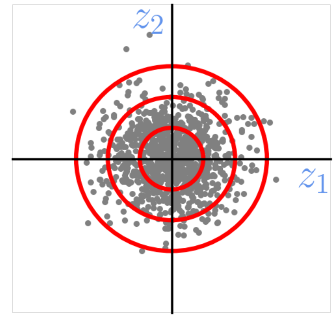
Exploring covariance structure through isoprobability contours

Standard Gaussian

$$\boldsymbol{z} \sim N(\mathbf{0}, \boldsymbol{I})$$

$$p(\boldsymbol{z}) \propto e^{-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{z}} = \text{const.} \longrightarrow \boldsymbol{z}^T\boldsymbol{z} = \text{const.}$$

in 2D: $z_1^2 + z_2^2 = \text{const.}$

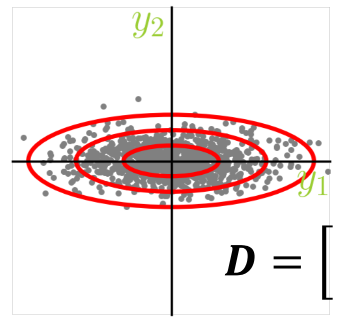


Introducing scaling $\boldsymbol{D}$ (diagonal)

$$\boldsymbol{y} = \boldsymbol{D}\boldsymbol{z} \qquad \text{cov}[\boldsymbol{y}] = \boldsymbol{\Sigma}_{\text{y}} = \boldsymbol{D}\boldsymbol{D}^T = \boldsymbol{D}^2$$

$$\boldsymbol{y}^T\boldsymbol{\Sigma}_y^{-1}\boldsymbol{y} = \boldsymbol{y}^T\boldsymbol{D}^{-2}\boldsymbol{y} = \boldsymbol{z}^T\boldsymbol{z} = \text{const.}$$

in 2D: $\left(\frac{y_1}{d_1}\right)^2 + \left(\frac{y_2}{d_2}\right)^2 = \text{const.}$



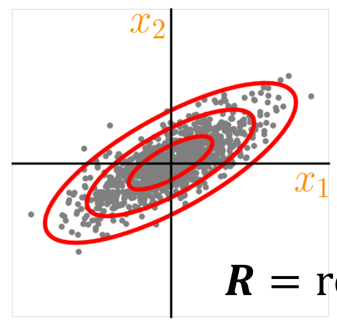
$$\boldsymbol{D} = \begin{bmatrix} 1.5 & 0 \\ 0 & 0.5 \end{bmatrix}$$

Introducing rotation $\boldsymbol{R}$ (orthonormal)

$$\boldsymbol{x} = \boldsymbol{R}\boldsymbol{y} = \boldsymbol{R}\boldsymbol{D}\boldsymbol{z} \qquad \text{cov}[\boldsymbol{x}] = \boldsymbol{\Sigma}_{\text{x}} = \boldsymbol{R}\boldsymbol{D}\boldsymbol{D}^T\boldsymbol{R}^T = \boldsymbol{R}\boldsymbol{D}^2\boldsymbol{R}^T$$

$$\boldsymbol{x}^T\boldsymbol{\Sigma}_x^{-1}\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{R}\boldsymbol{D}^{-2}\boldsymbol{R}^T\boldsymbol{x} = \boldsymbol{z}^T\boldsymbol{z} = \text{const.}$$

in 2D: $\left(\frac{(\boldsymbol{R}^T\boldsymbol{x})_1}{d_1}\right)^2 + \left(\frac{(\boldsymbol{R}^T\boldsymbol{x})_2}{d_2}\right)^2 = \text{const.}$



$$\boldsymbol{R} = \text{rot}(30°)$$

columns = $\boldsymbol{\Sigma}_x$ eigenvectors diagonal = (positive) $\boldsymbol{\Sigma}_x$ eigenvalues

Spectral decomposition: $\boldsymbol{\Sigma}_x = \boldsymbol{R}\boldsymbol{D}^2\boldsymbol{R}^T$

Rank-deficient covariance matrices

So far: (implicitly) assumed $\dim(\boldsymbol{x}) \equiv N_x = N_z \equiv \dim(\boldsymbol{z})$ (i.e., a square $\boldsymbol{A}$)

$$\boldsymbol{\Sigma}_x = \underbrace{\boldsymbol{R}}_{N_x \times N_z}\boldsymbol{D}^2\underbrace{\boldsymbol{R}^T}_{N_z \times N_z}$$

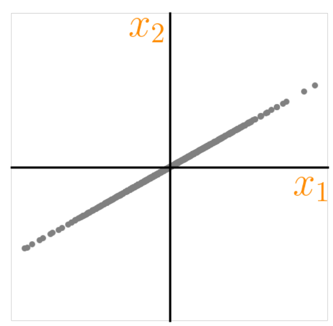
Case 1: Mapping from higher-dimensional $\boldsymbol{z}$ to lower-dimensional $\boldsymbol{x}$, $\mathrm{N}_x < N_z$ (more "inputs" than "outputs")

$\boldsymbol{\Sigma}_x$ is full rank (i.e., all eigenvalues are non-zero)

Case 2: Mapping from lower-dimensional $\boldsymbol{z}$ to higher-dimensional $\boldsymbol{x}$, $\mathrm{N}_x > N_z$ (fewer "inputs" than "outputs")

$\boldsymbol{\Sigma}_x$ is full rank-deficient (i.e., some eigenvalues are zero)

With $N_z = 1$ and $N_x = 2$: $\boldsymbol{x} = \boldsymbol{a}z + \boldsymbol{\mu}$



Interim summary: the Gaussian distribution

Fully specified by mean (vector) $\boldsymbol{\mu}$ and (co)variance (matrix) $\boldsymbol{\Sigma}$

Probability drops off quickly with distance from mean (“light” tail)
- sensitive to outliers

Linear transformation of (univariate/multivariate) Gaussian = Gaussian

Multivariate Gaussian = linear transformation of standard Gaussians
- Linear transformation = shift + scaling & rotation
- Scaling/rotation = eigenvalues/vectors of covariance matrix

Linear transformation from low-d to high-d: rank-deficient covariance (perfect correlations)

Probabilistic operations: marginalization

Linear construction of multivariate Gaussian

$$\boldsymbol{z} \sim \mathrm{N}(0, \boldsymbol{I}) \qquad \boldsymbol{x} = \boldsymbol{A}\boldsymbol{z} + \boldsymbol{\mu} \qquad \boldsymbol{x} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad \boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^T$$

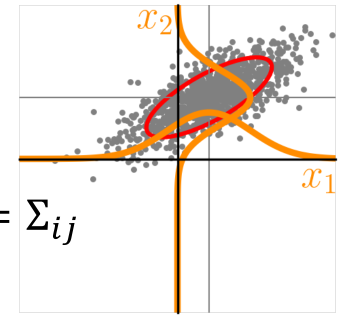
Moments of $\boldsymbol{x}$ components

ith row of $\boldsymbol{A}$

$$x_i = \boldsymbol{a}_{i:}^T \boldsymbol{z} + \boldsymbol{\mu}_i$$

$$\mathrm{E}[x_i] = \mu_i$$

$$\mathrm{cov}[x_i, x_j] = \mathrm{E}\big[\underbrace{\boldsymbol{a}_{i:}^T \boldsymbol{z}}_{x_i - \mathrm{E}[x_i]}\, \boldsymbol{z}^T \boldsymbol{a}_{j:}\big] = \boldsymbol{a}_{i:}^T \mathrm{E}[\boldsymbol{z}\boldsymbol{z}^T] \boldsymbol{a}_{j:} = \boldsymbol{a}_{i:}^T \boldsymbol{a}_{j:} = \Sigma_{ij}$$

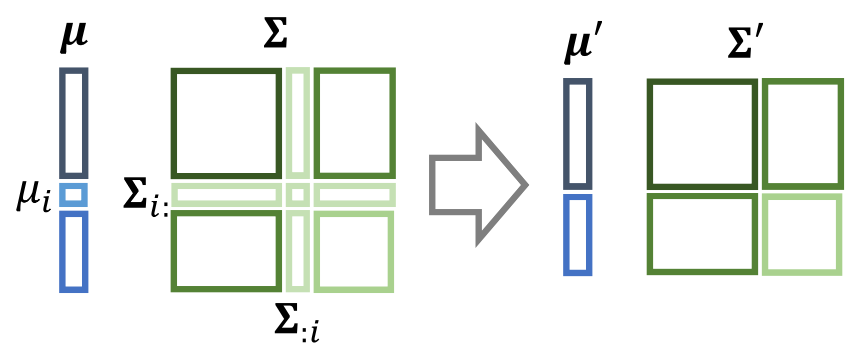


Marginalization: ‘removing’ x_i from $\boldsymbol{x}$ by

$$p(\boldsymbol{x}') = p(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_N) = \int p(\boldsymbol{x}) dx_i$$

does not change moments of $x_{j \neq i}$
(as moments of x_j only depend on μ_j and $\boldsymbol{a}_{j:}$)

$$\boldsymbol{x} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad \boldsymbol{x}' \sim \mathrm{N}(\boldsymbol{\mu}', \boldsymbol{\Sigma}')$$



(Alternative approach: define $\boldsymbol{x}' = \boldsymbol{B}\boldsymbol{x}$, where $\boldsymbol{B}$ ‘picks’ a subset of $\boldsymbol{x}$ components)

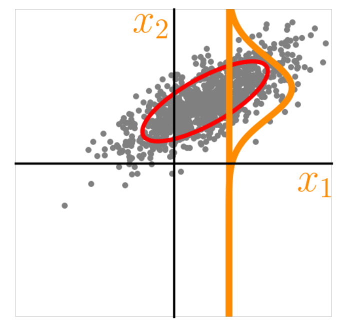
Probabilistic operations: conditioning

Linear construction of multivariate Gaussian

$$\boldsymbol{z} \sim \mathrm{N}(0, \boldsymbol{I}) \qquad \boldsymbol{x} = \boldsymbol{A}\boldsymbol{z} + \boldsymbol{\mu} \qquad \boldsymbol{x} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad \boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}^T$$

Conditioning: find $p(\boldsymbol{x}'|x_i) = p(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_N|x_i)$ from $p(\boldsymbol{x})$

To demonstrate: assuming 2D $\boldsymbol{x}$



$$x_1 = a_{11}z_1 + a_{12}z_2 + \mu_1$$
$$x_2 = a_{21}z_1 + a_{22}z_2 + \mu_2$$

If we 'know' (i.e., condition on) x_1:

$$z_2 = \frac{x_1 - a_{11}z_1 - \mu_1}{a_{12}}$$

(the value of z_2 is fully determined by z_1 and x_1)

$$x_2 = a_{21}z_1 + \frac{a_{22}(x_1 - a_{11}z_1 - \mu_1)}{a_{12}} + \mu_2$$
$$= \left(a_{21} - \frac{a_{22}a_{11}}{a_{12}}\right)z_1 + \frac{a_{22}(x_1 - \mu_1)}{a_{12}} + \mu_2$$

(x_1 is a linear function of $z_1 \rightarrow$ Gaussian)

More generally: conditionals of multivariate Gaussians are again Gaussian

Estimating parameters by maximum likelihood

Assume N 1D observations $x_{1:N}$ drawn i.i.d. from Gaussian $\mathrm{N}(x_n|\mu,\sigma^2)$

We find **maximum likelihood** parameters by solving

$$\hat{\mu}_{ML}, \hat{\sigma}^2_{ML} = \mathrm{argmax}_{\mu,\sigma^2} \prod_{n=1}^{N} \mathrm{N}(x_n|\mu,\sigma^2) = \mathrm{argmax}_{\mu,\sigma^2} \sum_{n=1}^{N} \log N(x_n|\mu,\sigma^2)$$

$$= \mathrm{argmax}_{\mu,\sigma^2} - \frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2$$

Setting the derivative with respect to μ and σ^2 to zero yields

$$\hat{\mu}_{ML} = \frac{1}{N}\sum_n x_n \qquad \hat{\sigma}^2_{ML} = \frac{1}{N}\sum_n (x_n - \hat{\mu}_{ML})^2$$

(note: ML variance estimate is biased; see PRML Sec. 1.2.4)

ML estimates are simply sample mean & sample variance
(this also holds for multivariate Gaussian)

Priors as implicitly added observations

Gaussian prior $p(\mu) = N\left(\mu \middle| 0, \frac{\sigma^2}{\lambda}\right)$ **on mean** μ results in MAP estimate

$$\hat{\mu}_{MAP} = \text{argmax}_{\mu} p(\mu) \prod_{n=1}^{N} p(x_n|\mu, \sigma^2) = \frac{1}{N + \lambda}\left(\sum_{n=1}^{N} x_n + \lambda 0\right)$$

Prior implicitly adds λ observations with value zero;
$\lambda \rightarrow 0$ recovers ML estimate, $\hat{\mu}_{MAP} \rightarrow \hat{\mu}_{ML}$; reveals implicit assumption of ML estimate

(Conjugate) **Normal Inverse Gamma prior** $p(\mu, \sigma^2) = \mathrm{N}\left(\mu \middle| 0, \frac{\sigma^2}{\lambda}\right)\mathrm{IG}(\sigma^2|\alpha, \beta)$ results in

$$\hat{\mu}_{MAP} = \frac{1}{N + \lambda} \sum_{n=1}^{N} x_n$$

$$\hat{\sigma}^2_{MAP} = \frac{2\alpha + 3}{N + 2\alpha + 3} \frac{\lambda \hat{\mu}_{ML} + 2\beta}{2\alpha + 3} + \frac{N}{N + 2\alpha + 3} \hat{\sigma}^2_{ML}$$

Prior implicitly adds

λ observations with value zero

$2\alpha + 3$ observations
with value $\frac{\lambda \hat{\mu}_{ML} + 2\beta}{2\alpha + 3}$

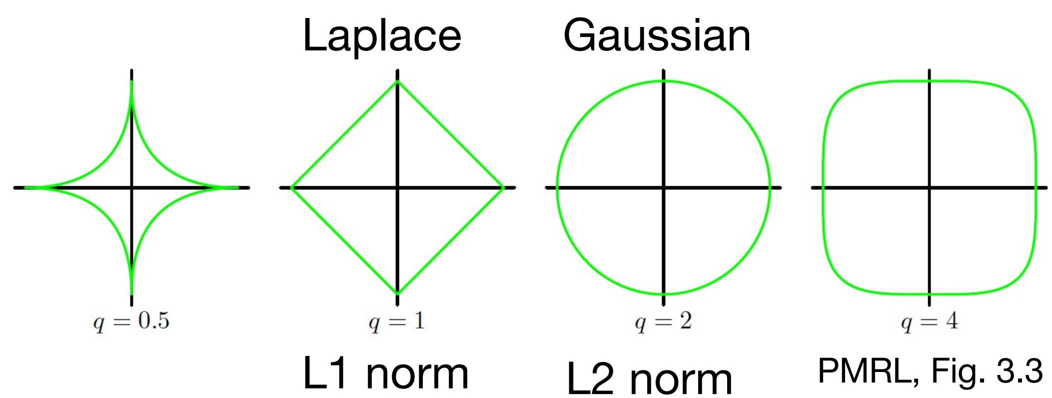
(Note: for the above, one can analytically find the whole posterior $p(\mu, \sigma^2|x_{1:N})$)

Priors as regularizers

MAP estimate balances prior $p(\boldsymbol{\mu})$ and likelihood $p(\boldsymbol{x}_{1:N}|\boldsymbol{\mu},\boldsymbol{\Sigma})$

$$\widehat{\boldsymbol{\mu}}_{MAP} = \mathrm{argmax}_{\mu}(\log p(\boldsymbol{\mu}) + \log p(\boldsymbol{x}_{1:N}|\boldsymbol{\mu},\boldsymbol{\Sigma}))$$

Assume parametric prior,
$\log p(\boldsymbol{\mu}) = \sum_{i=1}^{N_x} |\mu_i|^q + \text{const.}$

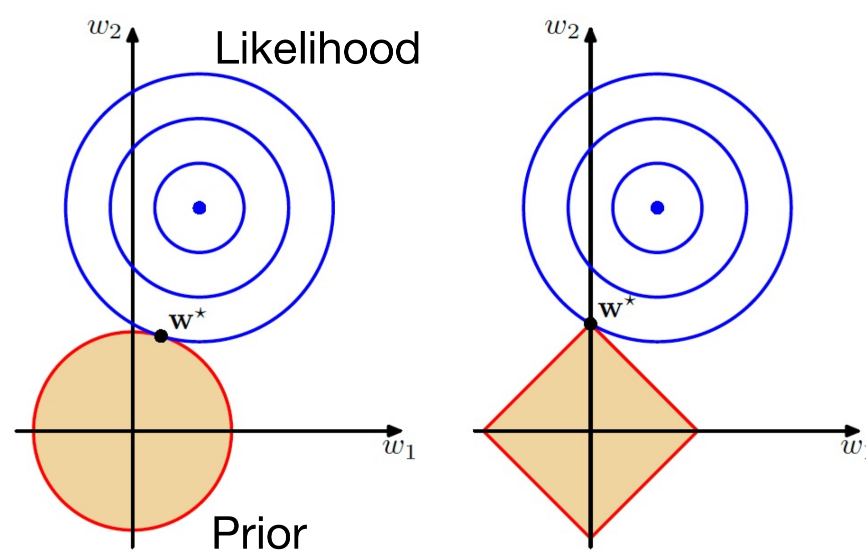


PMRL, Fig. 3.3

Laplace prior pushes MAP estimate components to zero

LASSO:
Laplace prior = "LASSO"

Elastic net:
Gaussian + Laplace prior



PMRL, Fig. 3.4

Interim summary: inference & priors

Both marginalization & conditioning of multivariate Gaussian yields Gaussian

Marginalization: removing rows / columns of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$

Summary of inference on Gaussians: PRML Appendix B, Sec. *Gaussian*

Maximum likelihood = matching sample moments (biased)

Priors can be interpreted as additional observations contributing to posterior / MAP estimate

Priors regularize / Laplace prior (L1 regularization) pushes (some) MAP components to zero

Overview

Univariate and multivariate Gaussian distributions

- Means and variances of linearly transformed random variables (general)
- Univariate standard Gaussian and linear transformations thereof
- Multivariate Gaussians through linear constructions
- Probabilistic operations: marginalization and conditioning
- Maximum likelihood parameter estimation
- Priors as added observations / regularizers

Linear models

- Maximum likelihood and least squares estimates
- What is linear in linear models?
- Priors and Bayesian inference

Bayesian model comparison

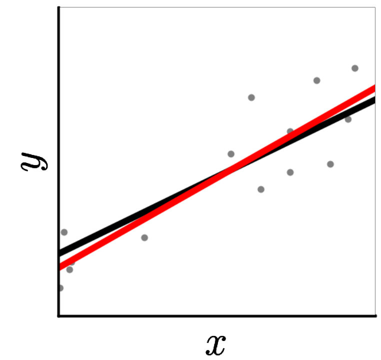
- Trading off goodness-of-fit with model complexity
- Adjusting model complexity through hyperpriors
- Evidence approximation

Linear models: maximum likelihood = least squares

Assume 'output' y is a linear function of 'input' $\boldsymbol{x}$ + Gaussian noise

$$y = \boldsymbol{w}^T\boldsymbol{x} + w_0 + \eta \qquad \eta \sim \mathrm{N}(0, \sigma^2)$$

$$p(y|\boldsymbol{x}, \boldsymbol{w}) = \mathrm{N}(y|\boldsymbol{w}^T\boldsymbol{x} + w_0, \sigma^2)$$



Linear regression

Observe $\{\boldsymbol{x}_n, y_n\}_{n=1}^N$, what is $\boldsymbol{w}$ and w_0 that makes data most likely?

$$\mathrm{argmax}_{\boldsymbol{w},w_0} \log \prod_{n=1}^{N} p(y_n|\boldsymbol{x}_n, \boldsymbol{w}) = \mathrm{argmax}_{\boldsymbol{w},w_0} -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(\underbrace{y_n}_{\text{target}} - \underbrace{\boldsymbol{w}^T\boldsymbol{x}_n - w_0}_{\text{model}})^2$$

Maximizing data likelihood = minimizing squared error
(sensitive to outliers)

Solution (assuming $\mathrm{w}_0 = 0$, we'll get to this)

Setting gradient of LLH with respect to $\boldsymbol{w}$ to zero

$$\widehat{\boldsymbol{w}}_{ML} = \left(\sum\nolimits_n \boldsymbol{x}_n\boldsymbol{x}_n^T\right)^{-1}\sum\nolimits_n \boldsymbol{x}_n y_n = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

design matrix

$$\boldsymbol{X} = \begin{pmatrix} -\boldsymbol{x}_1^T - \\ -\boldsymbol{x}_2^T - \\ \vdots \end{pmatrix} \quad \boldsymbol{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \end{pmatrix}$$

Linear models: linear in parameters, not 'inputs'

'Inputs' are assumed known/given – can change them while keeping solution structure

$$y = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}) + \eta \quad \eta \sim \mathrm{N}(0, \sigma^2)$$

$$p(y|\boldsymbol{x}, \boldsymbol{w}) = \mathrm{N}(y|\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}), \sigma^2)$$

Solution keep structure: $\hat{\boldsymbol{w}}_{ML} = (\boldsymbol{\Phi}^T \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T \boldsymbol{y}$

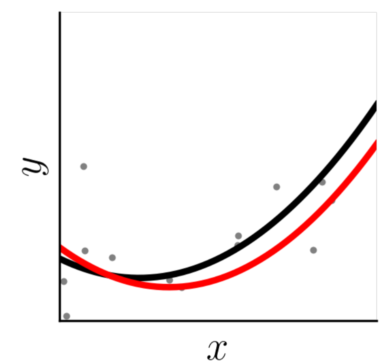
$$\boldsymbol{\Phi} = \begin{pmatrix} -\boldsymbol{\phi}(\boldsymbol{x}_1)^T - \\ -\boldsymbol{\phi}(\boldsymbol{x}_2)^T - \\ \vdots \end{pmatrix}$$

Example: include bias term (assumed from now on)

$$\boldsymbol{\phi}(\boldsymbol{x}) = \begin{pmatrix} 1 \\ \boldsymbol{x} \end{pmatrix} \qquad \boldsymbol{w} = \begin{pmatrix} \boldsymbol{w}_0 \\ \tilde{\boldsymbol{w}} \end{pmatrix} \longrightarrow \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}) = \tilde{\boldsymbol{w}}^T x + w_0$$

Example: polynomial regression

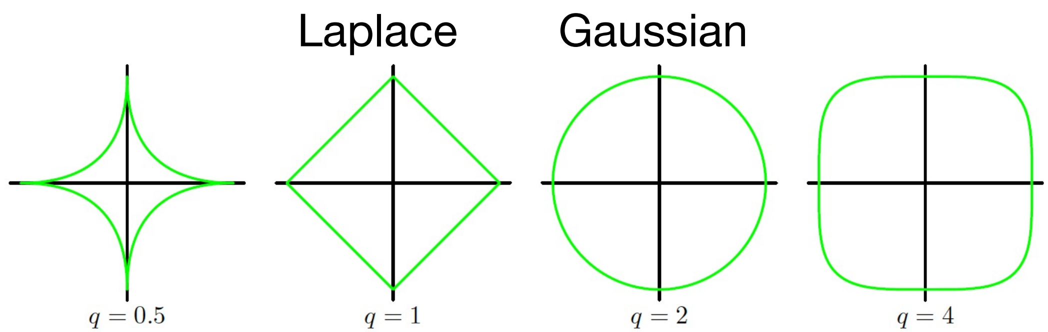
$$\boldsymbol{\phi}(x) = \begin{pmatrix} 1 \\ x \\ x^2 \\ \vdots \end{pmatrix} \longrightarrow \boldsymbol{w}^T \boldsymbol{\phi}(x) = w_0 + w_1 x + w_2 x^2 + \cdots$$



Linear models: revisiting (parameter) priors

$$\hat{w}_{MAP} = \text{argmax}_w p(\boldsymbol{y}|\boldsymbol{w},\boldsymbol{X})p(\boldsymbol{w}) = \text{argmax}_w(\log p(\boldsymbol{y}|\boldsymbol{w},\boldsymbol{X}) + \log p(\boldsymbol{w}))$$

Assume parametric prior,
$\log p(\boldsymbol{w}) = \sum_{i=1}^{N_x} |w_i|^q + \text{const.}$

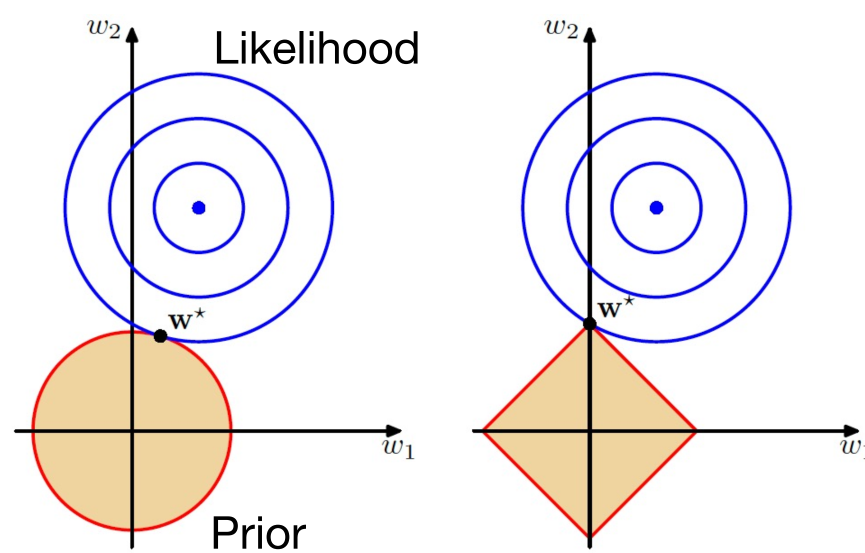


PMRL, Fig. 3.3

Laplace prior pushes MAP estimate components to zero – some 'inputs' get ignored / don't matter

LASSO:
Laplace prior = "LASSO"

Elastic net:
Gaussian + Laplace prior



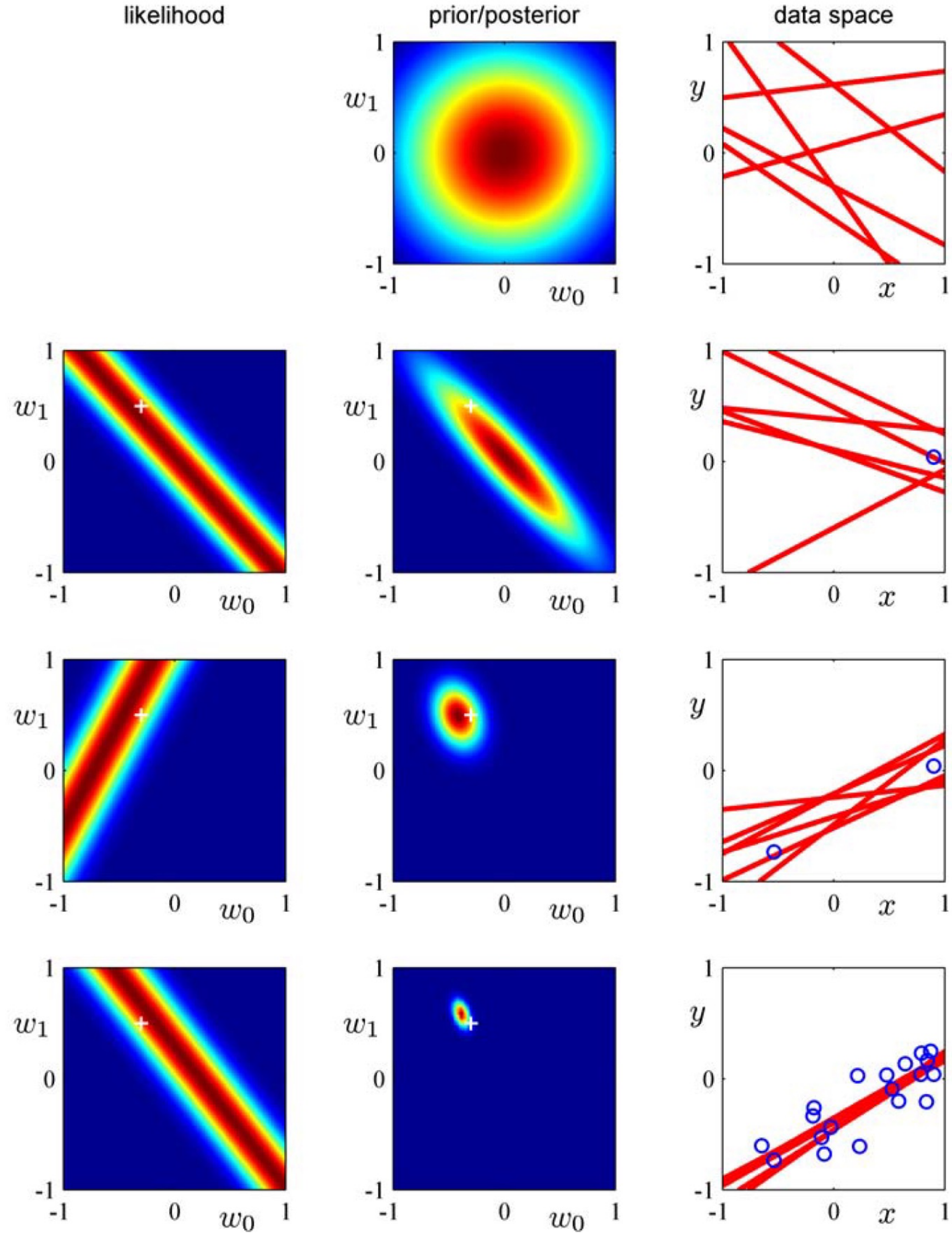
PMRL, Fig. 3.4

Bayesian linear regression

posterior ∝ likelihood x prior

$$p(\boldsymbol{w}|\boldsymbol{X},\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{w},\boldsymbol{X})p(\boldsymbol{w})$$

Gaussian in $\boldsymbol{w}$



PRML Fig 3.7

Interim summary: linear models

Linear models are linear in the parameters

Using non-linear functions of ‘inputs’ increases linear model applicability

Simple analytical solutions for ML estimates,
and MAP estimates / full posterior with Gaussian priors

ML solution = least squares solution (due to Gaussian residuals/noise)

L1 / Laplace parameter prior introduces sparsity

Overview

Univariate and multivariate Gaussian distributions

- Means and variances of linearly transformed random variables (general)
- Univariate standard Gaussian and linear transformations thereof
- Multivariate Gaussians through linear constructions
- Probabilistic operations: marginalization and conditioning
- Maximum likelihood parameter estimation
- Priors as added observations / regularizers

Linear models

- Maximum likelihood and least squares estimates
- What is linear in linear models?
- Priors and Bayesian inference

Bayesian model comparison

- Trading off goodness-of-fit with model complexity
- Adjusting model complexity through hyperpriors
- Evidence approximation

Bayesian model comparison

So far implicit: models as yet another variable to condition on

e.g., linear function   e.g., quadratic function

Assume models M_1 and M_2 with priors $p(\boldsymbol{w}|M_j)$ and data likelihood $p(\boldsymbol{X}|\boldsymbol{w}, M_j)$

Observe $\boldsymbol{X}$: probability of model M_j is

$$p(M_j|\boldsymbol{X}) \propto p(\boldsymbol{X}|M_j)p(M_j)$$

Model evidence $p(\boldsymbol{X}|M_j) = \int p(\boldsymbol{X}|\boldsymbol{w}, M_j)p(\boldsymbol{w}|M_j)dw$

Bayes Factor: comparing models

$p(M_1|\boldsymbol{X}) > p(M_2|\boldsymbol{X})$, and $p(M_1) = p(M_2)$, implies

$$\text{Bayes Factor } K \equiv \frac{p(\boldsymbol{X}|M_1)}{p(\boldsymbol{X}|M_2)} = \frac{p(M_1|\boldsymbol{X})}{p(M_2|\boldsymbol{X})} > 1$$

K	dHart	bits	Strength of evidence
$< 10^0$	< 0	< 0	Negative (supports M_2)
10^0 to $10^{1/2}$	0 to 5	0 to 1.6	Barely worth mentioning
$10^{1/2}$ to 10^1	5 to 10	1.6 to 3.3	Substantial
10^1 to $10^{3/2}$	10 to 15	3.3 to 5.0	Strong
$10^{3/2}$ to 10^2	15 to 20	5.0 to 6.6	Very strong
$> 10^2$	> 20	> 6.6	Decisive

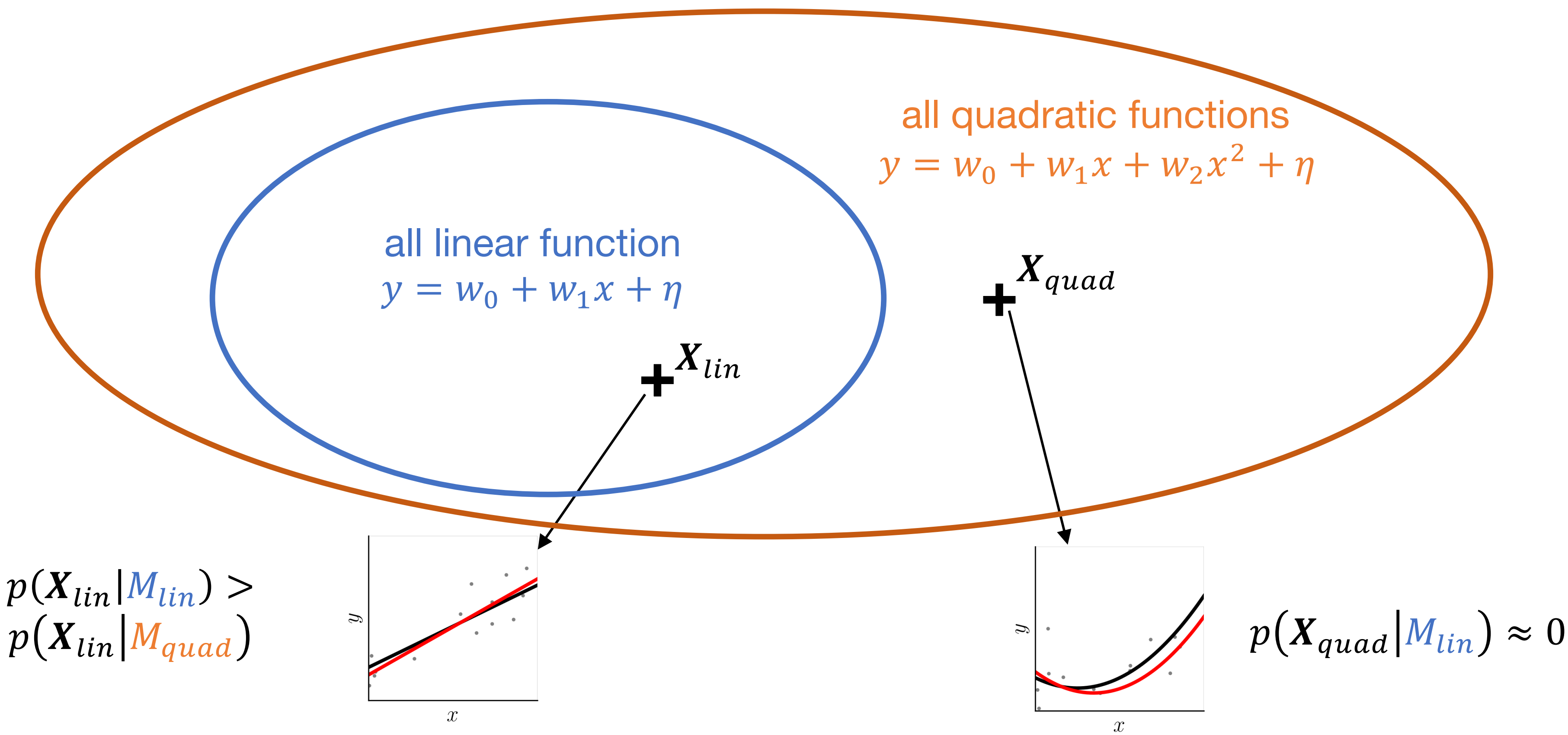
Jeffreys (1998); Wikipedia

Balancing model fit and complexity

Bayes Factor relies on $p(\boldsymbol{X}|M_j)$: how likely is it to observe data $\boldsymbol{X}$ under given model M_j?

$\int p(\boldsymbol{X}|M_j)d\boldsymbol{X} = 1$ $\longrightarrow$ the more different $\boldsymbol{X}$'s a model can 'explain' (i.e., $p(\boldsymbol{X}|M_j) \gg 0$) , the lower $p(\boldsymbol{X}|M_j)$ for each $\boldsymbol{X}$

Example: linear vs. quadratic function



Adjusting model complexity through hyperpriors

Hyperpriors: priors $p(\boldsymbol{\alpha})$ on parameters $\boldsymbol{\alpha}$ of the prior $p(\boldsymbol{w}|\boldsymbol{\alpha})$

Example: hyperprior on quadratic coefficient

$$p(w_2|\,\alpha) = N(w_2|0, \alpha^{-1})$$

$$y = w_0 + w_1 x + w_2 x^2 + \eta$$

$\alpha \to 0$ w_2 can take arbitrary values

$\alpha \to \infty$ $w_2 \to 0$ (linear function)

Inferring α from data: are quadratic or linear models more adequate?

$\longrightarrow$ determines model complexity from data
(a form of Automated Relevance Determination; ARD)

Full inference (usually intractable):

posterior predictive density

data likelihood

$$p(\boldsymbol{x}|\boldsymbol{X}) = \iint p(\boldsymbol{x}|\boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{X})p(\boldsymbol{\alpha}|\boldsymbol{X})\mathrm{d}\boldsymbol{w}\mathrm{d}\boldsymbol{\alpha}$$

posterior

hyperposterior

$$\begin{aligned} p(\boldsymbol{\alpha}|\boldsymbol{X}) &\propto p(\boldsymbol{X}|\boldsymbol{\alpha})p(\boldsymbol{\alpha}) \\ &= \int p(\boldsymbol{X}|\boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{\alpha})\mathrm{d}\boldsymbol{w} \end{aligned}$$

Evidence approximation

Tractable hyperparameter inference

$$p(\boldsymbol{x}|\boldsymbol{X}) = \iint p(\boldsymbol{x}|\boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{\alpha},\boldsymbol{X})p(\boldsymbol{\alpha}|\boldsymbol{X})\mathrm{d}\boldsymbol{w}\mathrm{d}\boldsymbol{\alpha}$$

hyperposterior

approximate by MAP, $\widehat{\boldsymbol{\alpha}}_{MAP}$

$$p(\boldsymbol{x}|\boldsymbol{X}) \approx p(\boldsymbol{x}|\boldsymbol{X},\widehat{\boldsymbol{\alpha}}_{MAP}) = \int p(\boldsymbol{x}|\boldsymbol{w})p(\boldsymbol{w}|\widehat{\boldsymbol{\alpha}}_{MAP},\boldsymbol{X})\mathrm{d}\boldsymbol{w}$$

Estimating MAP hyperparameters

$$\widehat{\boldsymbol{\alpha}}_{MAP} = \mathrm{argmax}_{\boldsymbol{\alpha}} \log p(\boldsymbol{\alpha}|\boldsymbol{X}) = \mathrm{argmax}_{\boldsymbol{\alpha}}(\log p(\boldsymbol{X}|\boldsymbol{\alpha}) + \log p(\boldsymbol{\alpha}))$$

$$p(\boldsymbol{X}|\boldsymbol{\alpha}) = \int p(\boldsymbol{X}|\boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{\alpha})\mathrm{d}\boldsymbol{w}$$

In PRML, Ch. 3.5:

- Assume uninformative $p(\boldsymbol{\alpha})$: $\widehat{\boldsymbol{\alpha}}_{MAP} \approx \widehat{\boldsymbol{\alpha}}_{ML}$
- $\widehat{\boldsymbol{\alpha}}_{MAP}$ depends on mean estimate, which depends on $\widehat{\boldsymbol{\alpha}}_{MAP}$: estimate recursively

Interim summary: Bayesian model comparison

Comparing models by marginalizing over the parameters / hyperparameters

Implicit complexity penalty by lower model evidence for models that can fit more data

Hyperpriors can provide automatic inference of desired model complexity

Evidence approximation for tractable hyperparameter marginalization

Overview

Univariate and multivariate Gaussian distributions

- Means and variances of linearly transformed random variables (general)
- Univariate standard Gaussian and linear transformations thereof
- Multivariate Gaussians through linear constructions
- Probabilistic operations: marginalization and conditioning
- Maximum likelihood parameter estimation
- Priors as added observations / regularizers

Linear models

- Maximum likelihood and least squares estimates
- What is linear in linear models?
- Priors and Bayesian inference

Bayesian model comparison

- Trading off goodness-of-fit with model complexity
- Adjusting model complexity through hyperpriors
- Evidence approximation

Summary

Gaussians are fully described by their mean (vector) and (co)variance matrix

A linear transformation of a Gaussian remains a Gaussian

Marginalization and conditioning Gaussians (on Gaussians) yields Gaussians

Priors can act as additional observations / as regularizers

Linear models: linear in parameters, not ‘inputs’

Maximum likelihood in linear models = least squares, with analytic solution

Bayesian model comparison trades off goodness-of-fit with model complexity

Hyperparameters can continuously modulate model complexity (evidence approx.)

Until next week

Read paper and prepare presentation (see notes for Session 3)

Read statistical methods sections (see notes for Session 3)

Next session

Q&A for previous session (~15min)

Paper discussions (~1h)

Introducing generalized linear models (~30min)