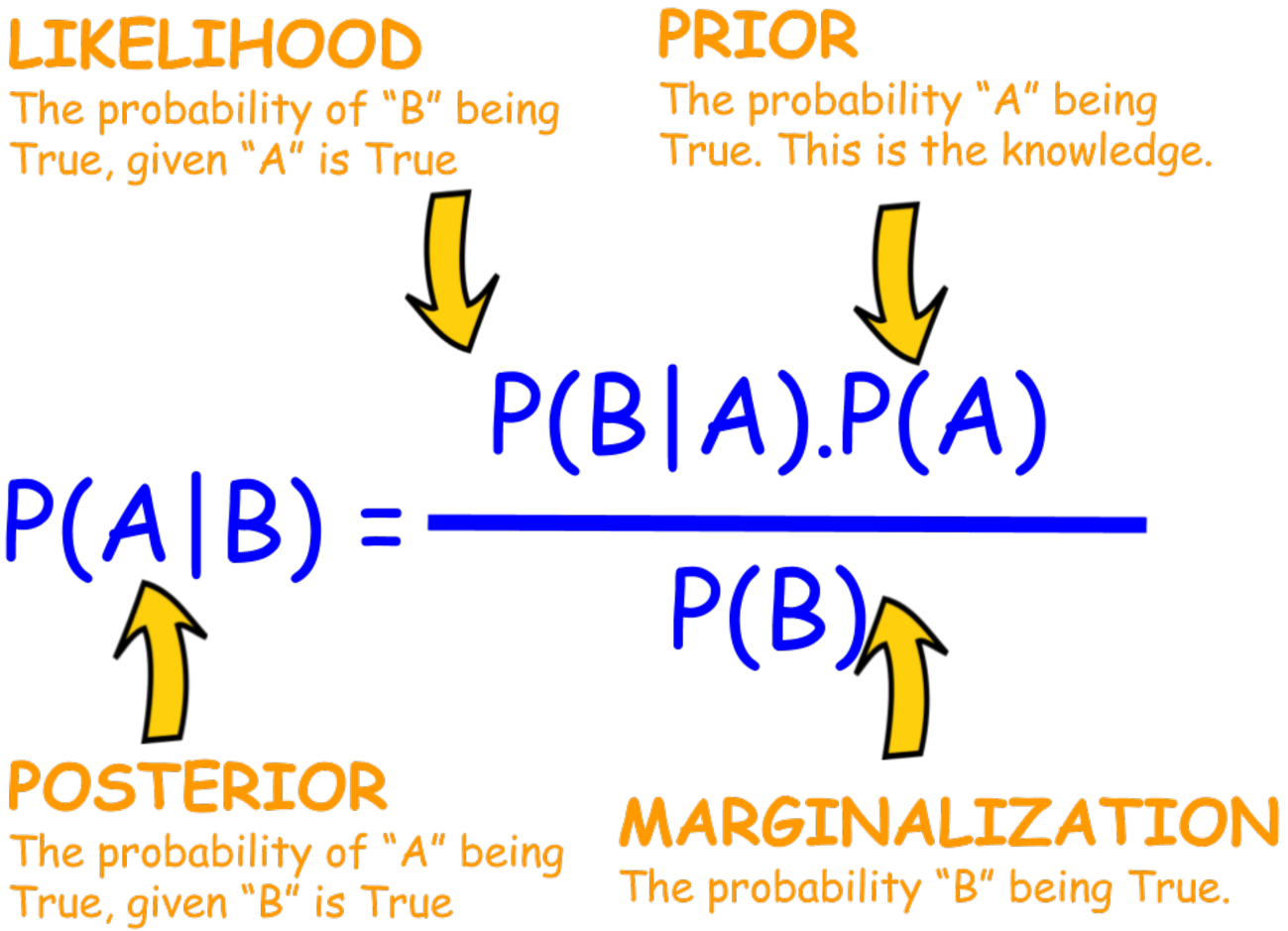
OIST Mini Course

Bayesian Methods in Neuroscience

Yuzhe Li
20240705@OIST

Bayesian Theory

Bayesian Theoram

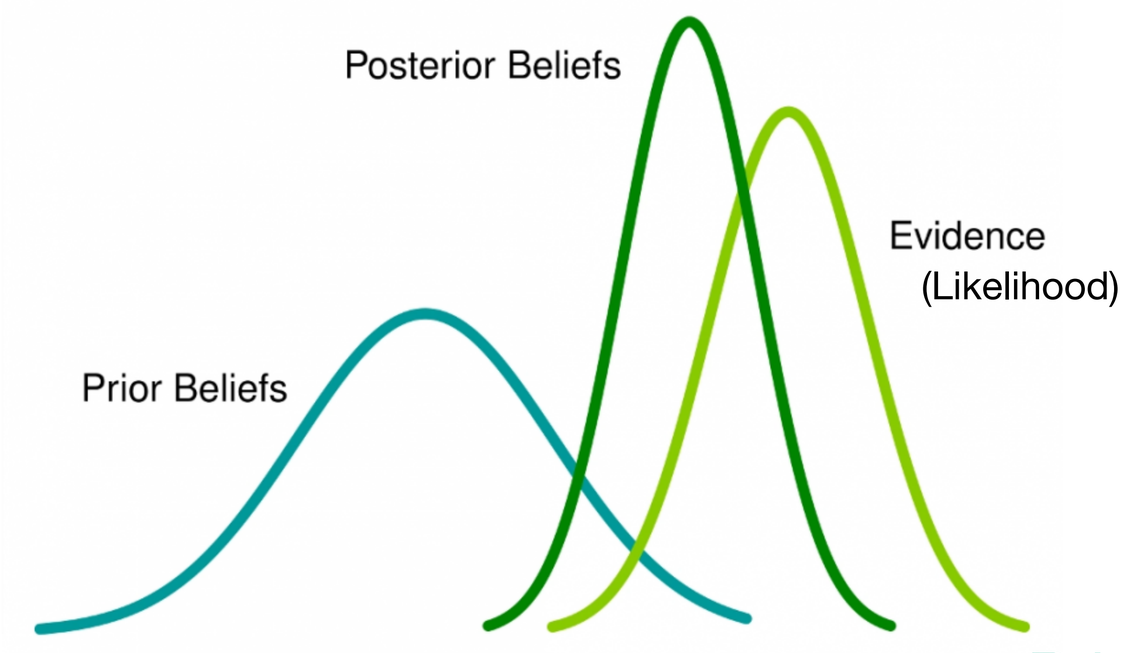


(Thomas Bayes, 1763)

(Pierre-Simon Laplace, 1774)

Bayesian Theory

Bayesian Inference



$$\underbrace{P(\text{Hypothesis}|\text{Data})}_{\text{Posterior}} = \frac{\overbrace{P(\text{Hypothesis})}^{\text{Prior}} \times \overbrace{P(\text{Data}|\text{Hypothesis})}^{\text{Likelihood}}}{\underbrace{P(\text{Data})}_{\text{Marginal}}}$$

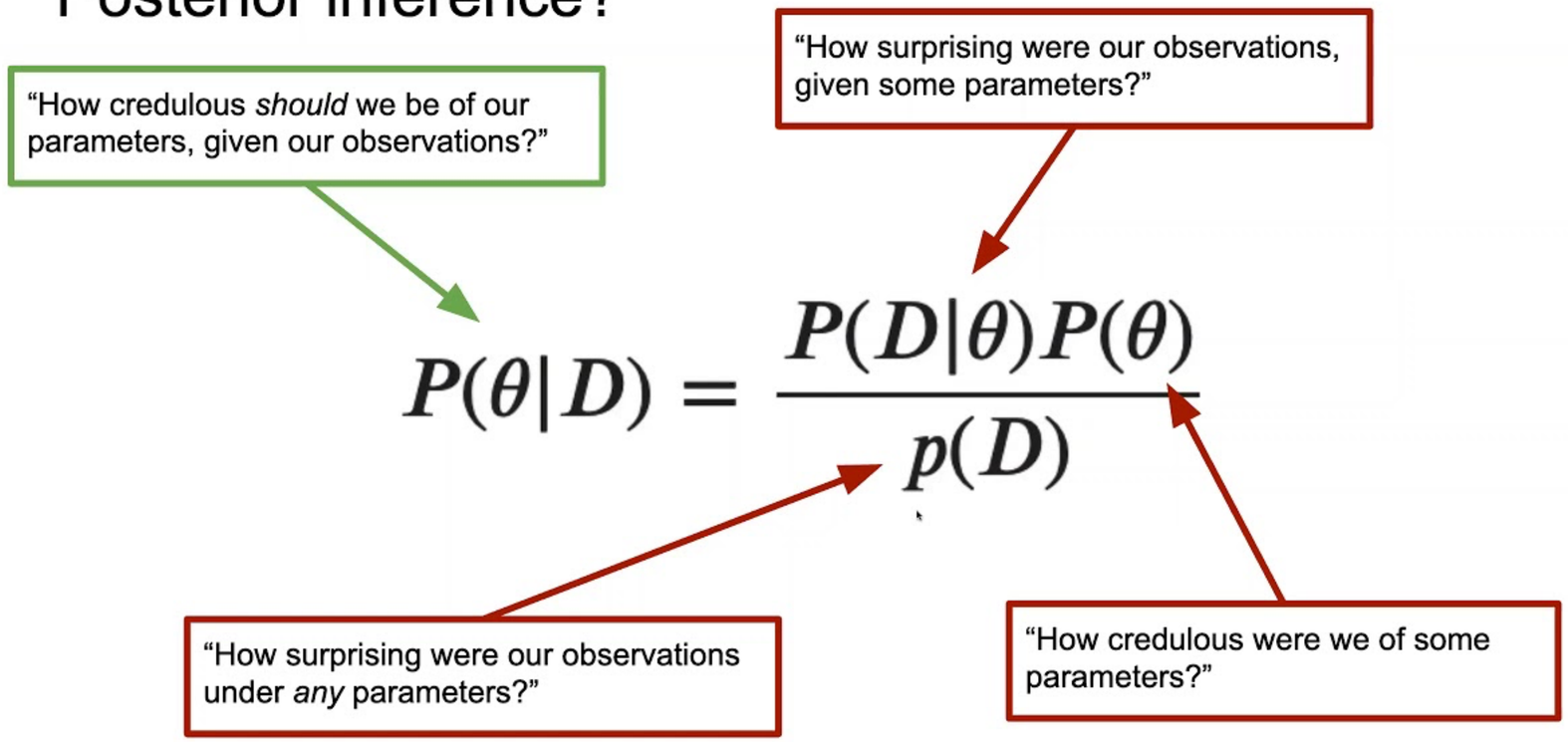
$$P(\text{Hypothesis}|\text{Data}) \propto P(\text{Hypothesis}) \times P(\text{Data}|\text{Hypothesis})$$

Bayesian Theory

Bayesian Parameter estimation

$$y = f(x; \theta)$$

Posterior inference?



Applications of Bayesian Methods

- **Parameter estimation**
  - Regression
  - Classification
  - Dimensionality reduction
  - …
- **Direct modeling on behaviour and neural activities**
  - Bayesian inference
  - Bayesian Population coding
  - …

Bayesian Parameter Estimation:

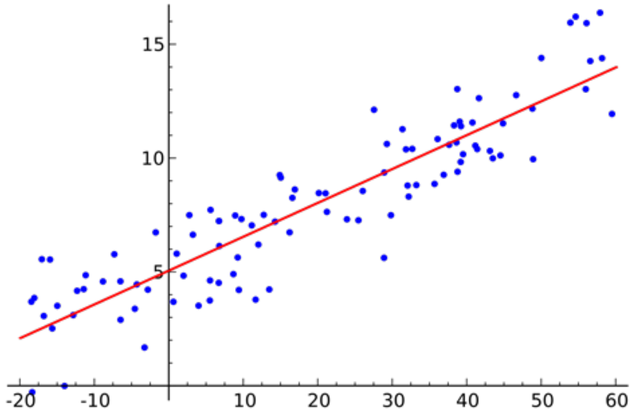
Two-class Classification Problem

Recap: Linear Regression

Linear Regression model

$$y = \sum_{i=0} w_i x_i + b \quad \text{or} \quad y = w^T x + b$$

How to solve w?



Recap: Linear Regression

Minimize Loss Function (Error function)

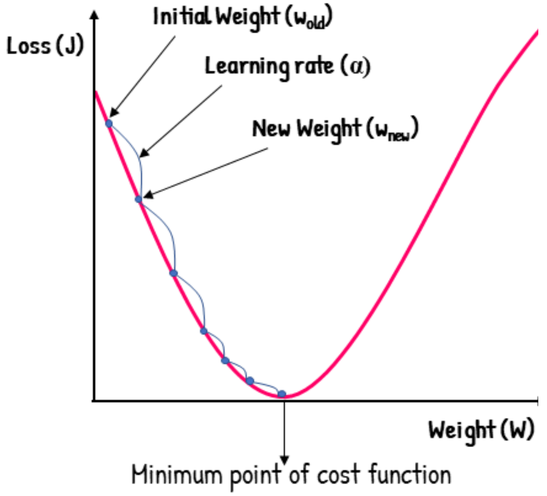
- Error function:
  - Example: Least Square
    $$E(w) = \sum_i (t_i - w_i x_i - b)^2$$

How to Minimize $E(w)$?

- Batch Method:
$$\frac{\partial E(w)}{\partial w} = 0$$

- Sequential Method:
$$w_{new} = w_{old} - \alpha \frac{\partial E(w)}{\partial w}$$

Gradient Descent



Recap: Linear Regression

Probabilistic Point of View

$$\boldsymbol{y} = \boldsymbol{w}^T\boldsymbol{x} + \boldsymbol{b} + \epsilon, \text{ where } \varepsilon \sim N(0,\sigma^2)$$

$$\text{or } \boldsymbol{y} \sim N(\boldsymbol{w}^T\boldsymbol{x} + \boldsymbol{b}, \sigma^2 I)$$

How to solve w?

Likelihood: $P(t \mid w) = N(\boldsymbol{w}^T\boldsymbol{x} + \boldsymbol{b}, \sigma^2 I)$

- **Maximize log-likelihood:**

$$\frac{\partial \log P(t \mid w)}{\partial w} = 0$$

Recap: Linear Regression

Bayesian Point of View

- Prior: $P(w) = N(w \mid m_0, S_0)$

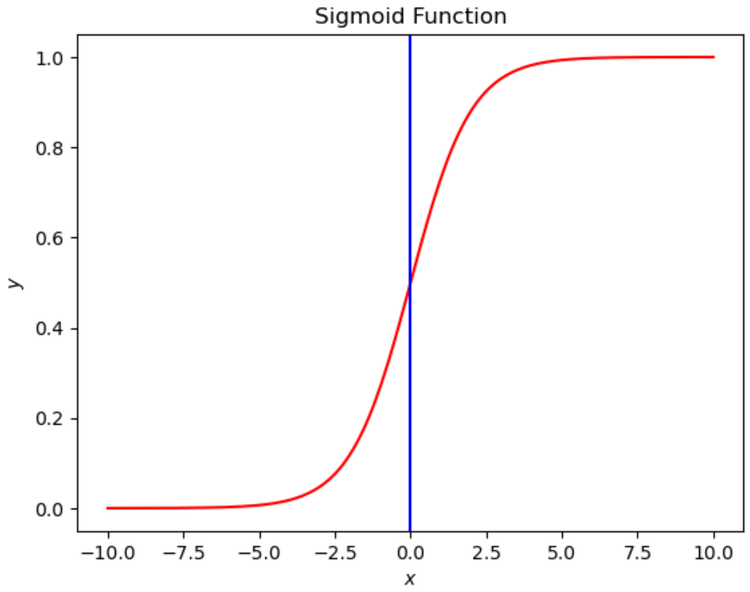
  Likelihood: $P(t \mid w) = N(\boldsymbol{w}^T \boldsymbol{x} + \boldsymbol{b}, \sigma^2 I)$

  Posterior: $P(w \mid t) \propto P(w) \times P(t \mid w)$

  Find w : Maximize Posterior (MAP)

Logistic Regression

Linear Logistic Regression Model



$$y = \sigma(w^T x + b)$$

where $\sigma(x) = \dfrac{1}{1 + e^{-x}}$

- Goal:
  - Find w
- Solution:
  - Minimize $E(w)$
  - Maximize likelihood

Example in python code

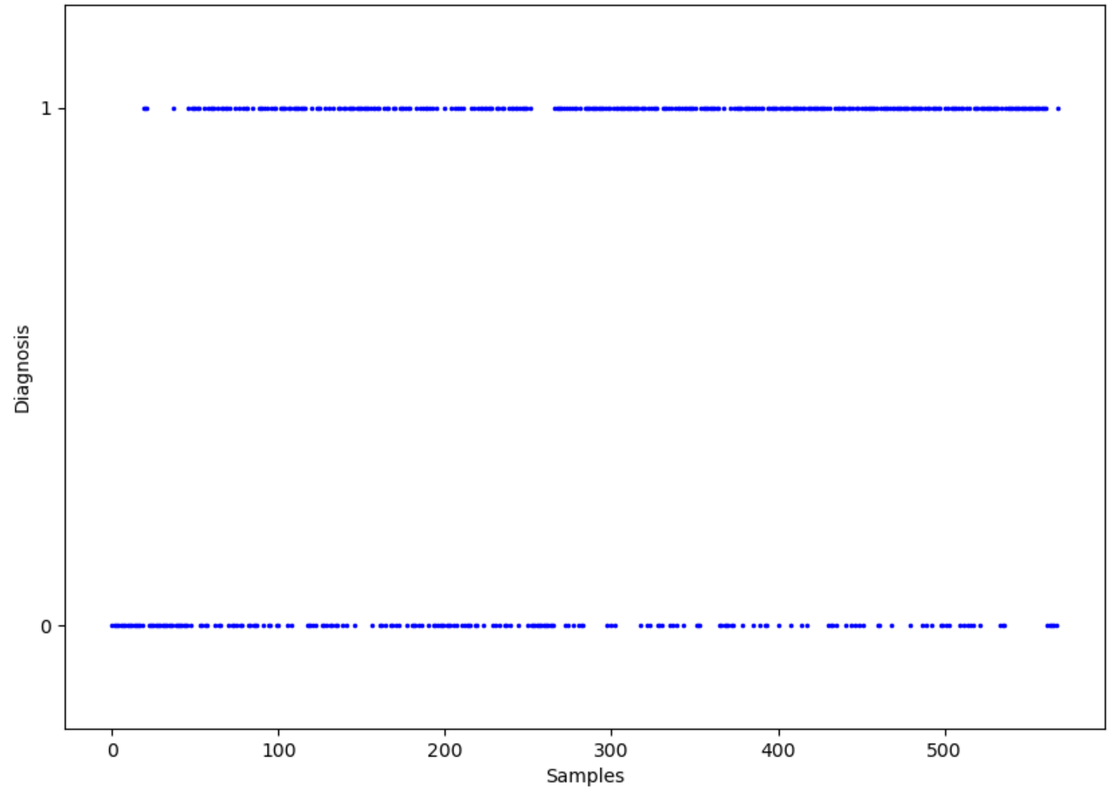
Breast cancer wisconsin (diagnostic) dataset

Features

mean radius	mean texture	mean perimeter	mean area	mean smoothness	mean compactness	mean concavity	mean concave points	mean symmetry	mean fractal dimension	...
0	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.30010	0.14710	0.2419	...
1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.08690	0.07017	0.1812	...
2	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.19740	0.12790	0.2069	...
3	11.42	20.38	77.58	386.1	0.14250	0.28390	0.24140	0.10520	0.2597	...
4	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.19800	0.10430	0.1809	...
...	...	...	...	...	...	...	...	...	...	...
564	21.56	22.39	142.00	1479.0	0.11100	0.11590	0.24390	0.13890	0.1726	...
565	20.13	28.25	131.20	1261.0	0.09780	0.10340	0.14400	0.09791	0.1752	...
566	16.60	28.08	108.30	858.1	0.08455	0.10230	0.09251	0.05302	0.1590	...
567	20.60	29.33	140.10	1265.0	0.11780	0.27700	0.35140	0.15200	0.2397	...
568	7.76	24.54	47.92	181.0	0.05263	0.04362	0.00000	0.00000	0.1587	...

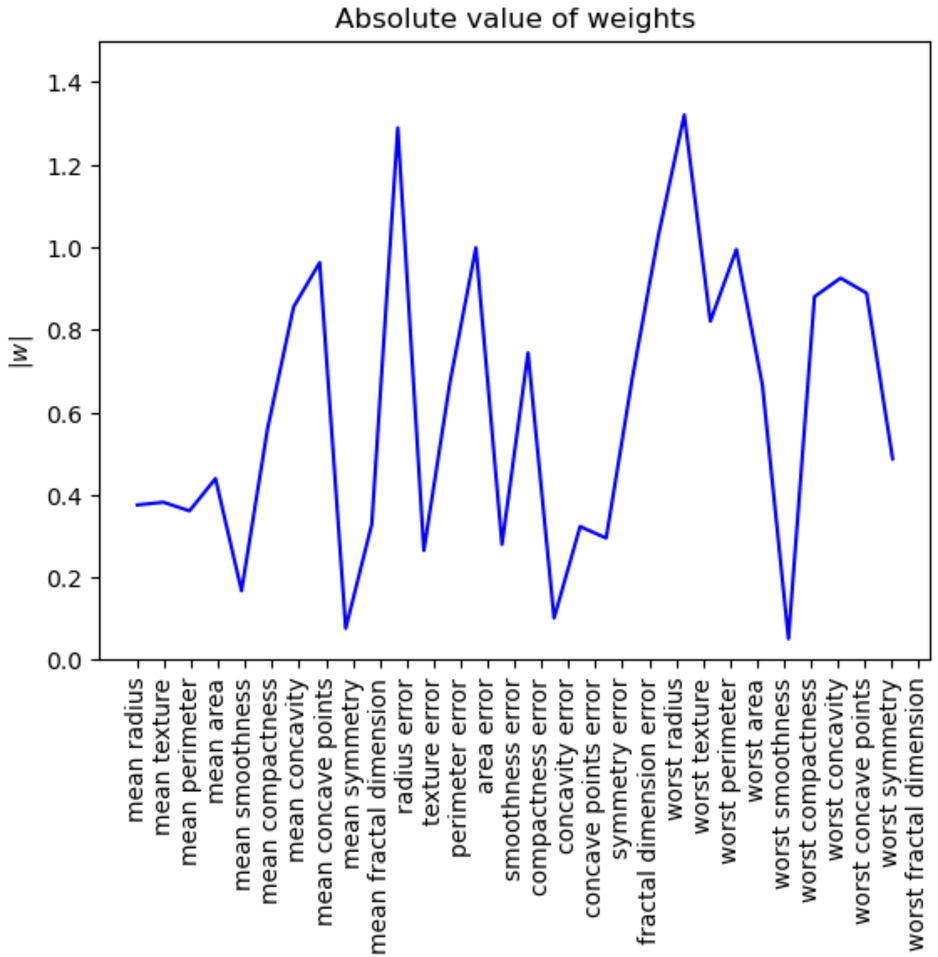
Number of Instances: 569
Number of Attributes: 30

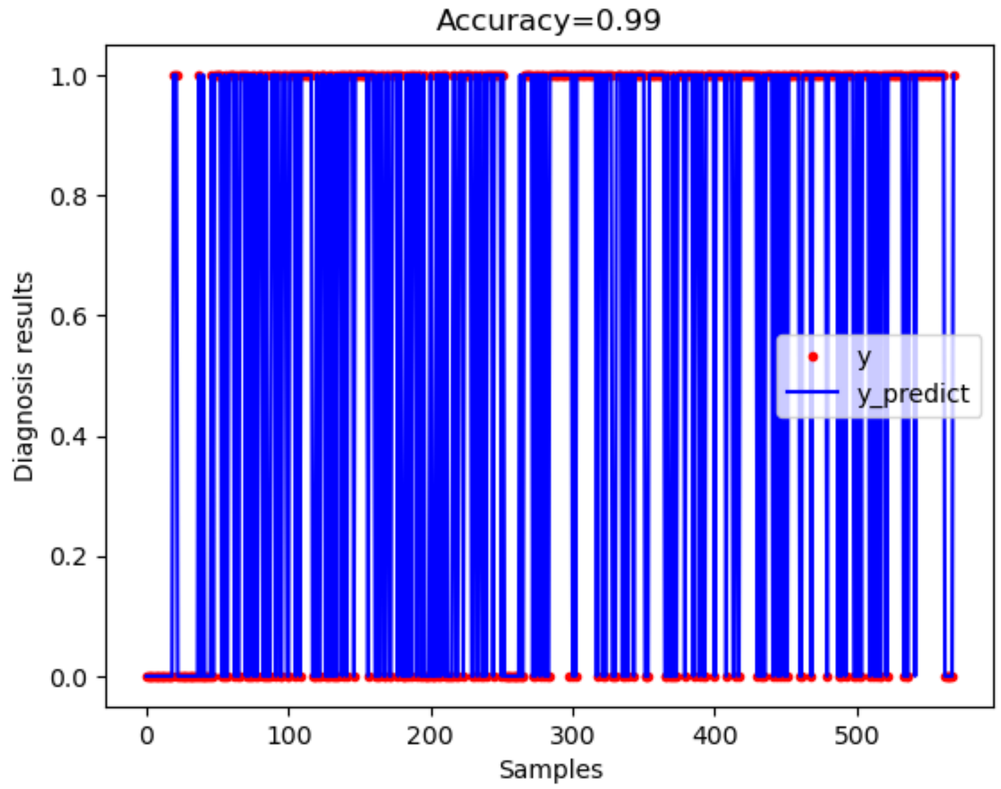
Diagnosis results



Results using Logistic regression

Solve using Gradient Descent method





Bayesian Logistic regression

Maximum Posterior

Prior: $P(w) = N(w \mid m_0, S_0)$

Likelihood: $P(t \mid w, x) = Bernulli(t \mid y) = y^t(1-y)^{1-t}$

where $y = P(t = 1 \mid w, x) = \sigma(w^T x)$

Posterior: $P(w \mid t) \propto P(t \mid w)P(w)$

$$\log P(w \mid t) = \log P(w) + \log P(t \mid w)$$

$$= -\frac{1}{2}(w - m_0)^T S_0^{-1}(w - m_0) + \sum_n \{t_n \log y_n + (1 - t_n)\log(1 - y_n)\} + const$$

Find w using MAP

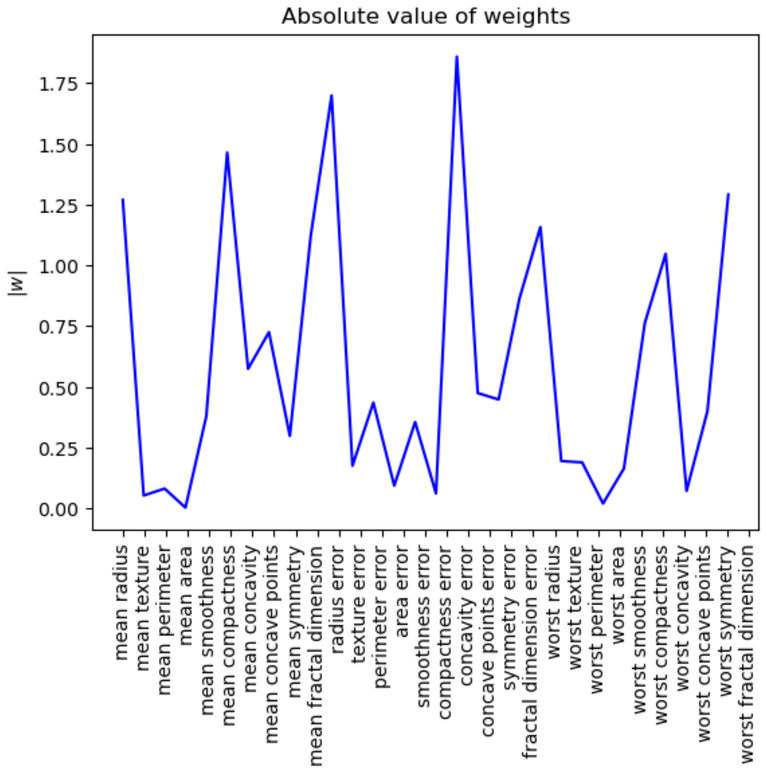
Bayesian Logistic regression

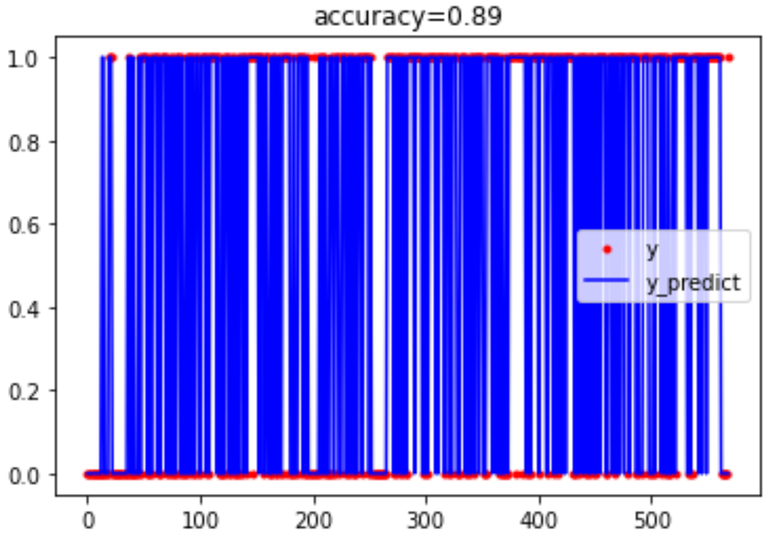
Predictive distribution

Direct Prediction (Semi-Bayesian): $y = \sigma(w_{MAP}^T x)$

Full Bayesian: $P(C_1 \mid t, x) = \int P(C_1 \mid t, x, w) P(w \mid t) dw \approx \int \sigma(w^T x) q(w) dw$

Classification using Bayesian logistic regression



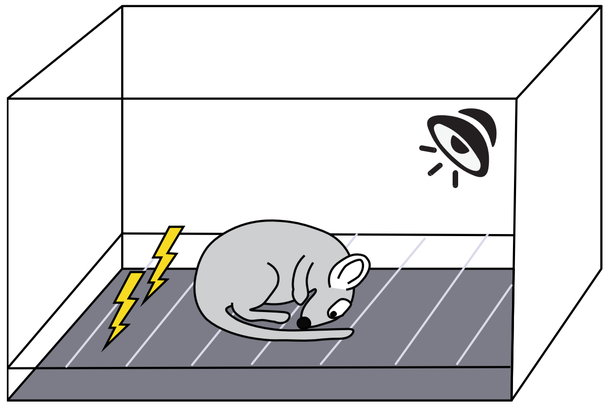


Bayesian model in Associative learning

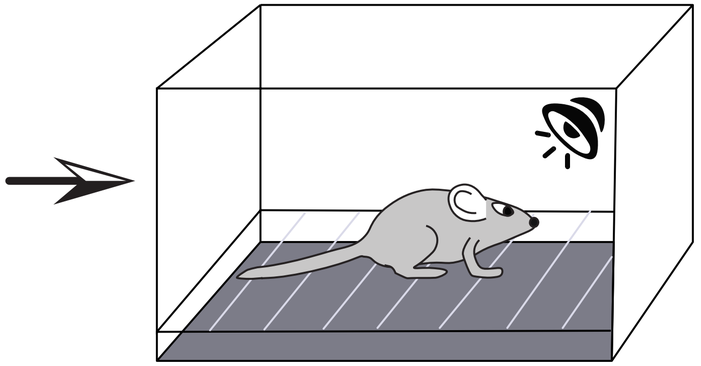
Bayesian model on behavior

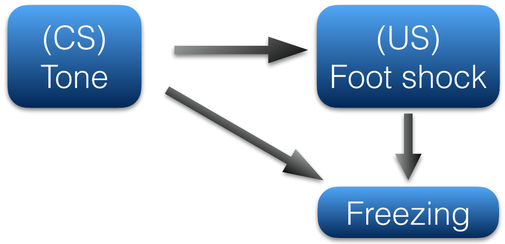
Fear conditioning

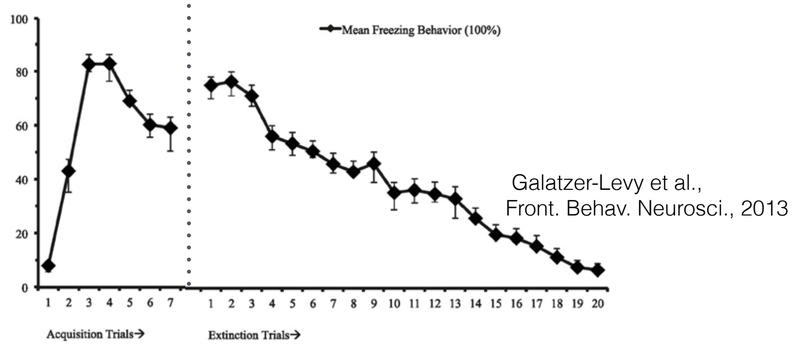
Conditioning



Extinction







Galatzer-Levy et al.,
Front. Behav. Neurosci., 2013

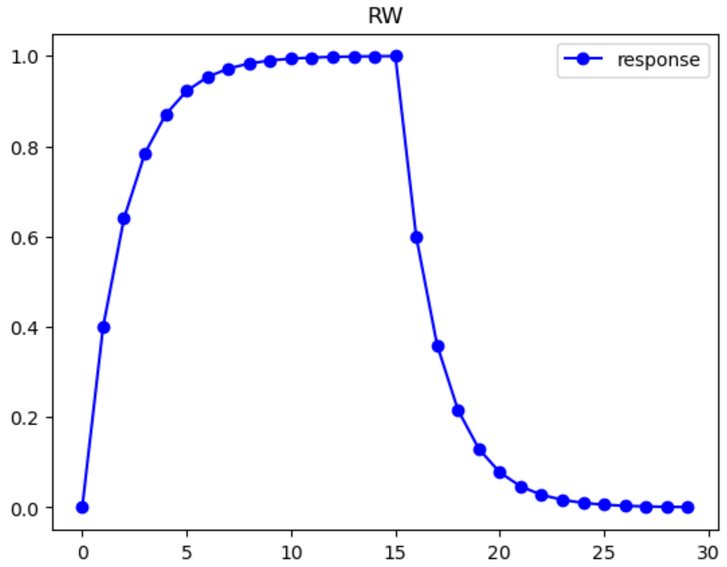
Associative learning

Rescorla-Wagner model

$$W_{n+1} = W_n + \Delta W$$

$$\Delta W = \alpha \mathrm{CS}(\mathrm{US}_n - \hat{\mathrm{US}}_n)$$

$$\hat{\mathrm{US}}_n = W_n \mathrm{CS}_n$$



Associative learning

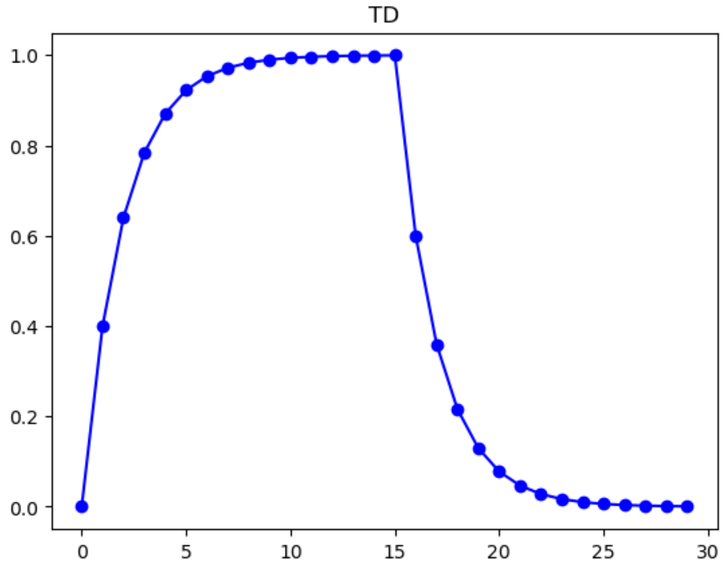
Temporal Difference (TD) error model

$$W_{n+1} = W_n + \Delta W$$

$$\delta_n = \mathrm{US}_n + \gamma W_n \mathrm{CS}_{n+1} - W_n \mathrm{CS}_n$$

$$\Delta W = \alpha \mathrm{CS} \delta_n$$

$$\hat{\mathrm{US}}_n = W_n \mathrm{CS}_n$$



Bayesian inference and Kalman Filter

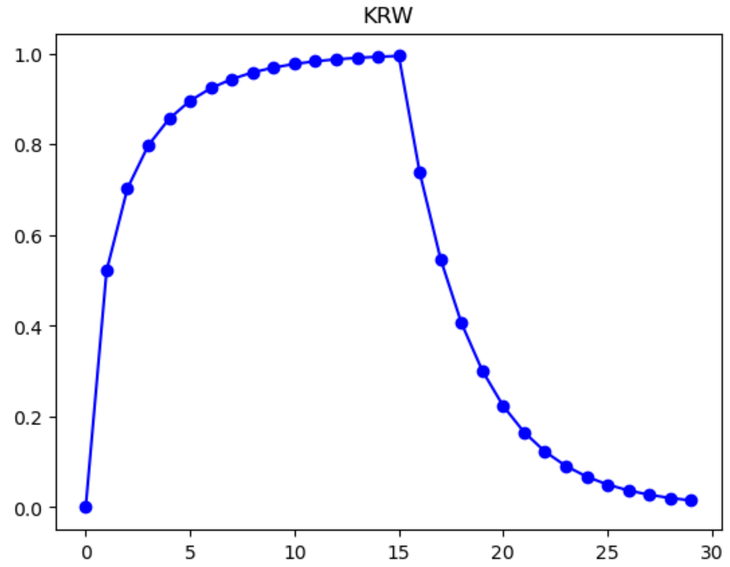
Rescorla-Wagner model with Kalman Filter

$$P(\mathbf{w}_n|x_{1:n}) \propto P(x_{1:n}|\mathbf{w}_n)P(\mathbf{w}_n)$$

The posterior Gaussian distribution of $\mathbf{w}_n$ with mean $\hat{\mathbf{w}}_n$ and covariance Σ_n, can be updated by Kalman Filter:

$$\hat{\mathbf{w}}_{n+1} = \hat{\mathbf{w}}_n + k_n\delta_n$$
$$\Sigma_{n+1} = \Sigma_n + \tau^2\mathbf{I} - k_n x_n^T(\Sigma_n + \tau^2\mathbf{I})$$
$$\mathbf{v}_n = \mathbf{w}_n^T x_n$$
$$\delta_n = r_n - \mathbf{v}_n$$

where, $k_n = \frac{(\Sigma_n+\tau^2\mathbf{I})x_n}{x_n^T(\Sigma_n+\tau^2\mathbf{I})x_n+\sigma_r^2}$,
$\hat{\mathbf{w}}_0 = 0, \Sigma_0 = \sigma_w^2\mathbf{I}$



TD model with Kalman Filter

$$P(\mathbf{w}_t|x_{1:t}) \propto P(x_{1:t}|\mathbf{w}_t)P(\mathbf{w}_t)$$

The posterior Gaussian distribution of $\mathbf{w}_t$ with mean $\hat{\mathbf{w}}_t$ and covariance Σ_t, can be updated by Kalman Filter:

$$\hat{\mathbf{w}}_{t+1} = \hat{\mathbf{w}}_t + k_t\delta_t$$

TD error:

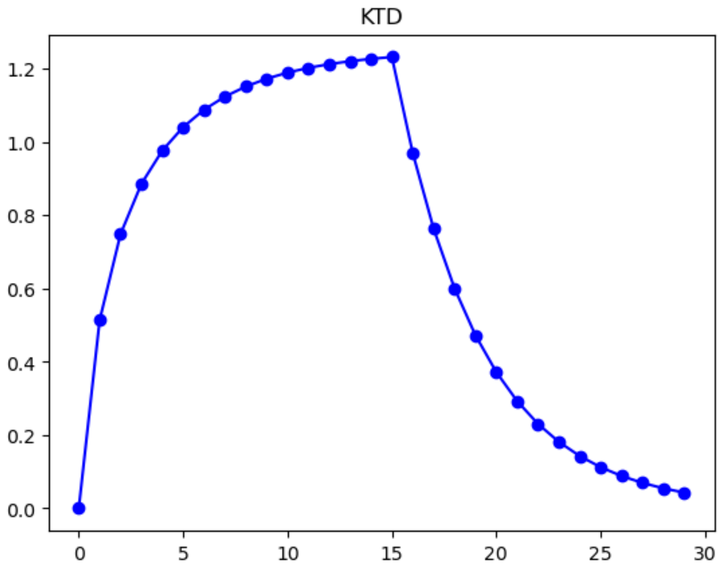
$$\delta_t = r_t + \gamma\hat{\mathbf{w}}_t^T x_{t+1} - \hat{\mathbf{w}}_t^T x_t$$

$$k_t = \frac{(\Sigma_t+\tau^2\mathbf{I})h_t}{h_t^T(\Sigma_n+\tau^2\mathbf{I})h_t+\sigma_r^2},$$

where $h_t = \gamma x_{t+1} - x_t$

$$\Sigma_{t+1} = \Sigma_t + \tau^2\mathbf{I} - k_t h_t^T(\Sigma_t + \tau^2\mathbf{I})$$

$$\hat{\mathbf{w}}_0 = 0, \Sigma_0 = \sigma_w^2\mathbf{I}$$



Bayesian logistic regression model

- The probability that the US would occur : P(US_t)

$$P(US_t = 1) = \sigma(\mathbf{w}_{t,0} + \mathbf{w}_{t,1} US_{t-1})$$

- Sequential Bayesian updating of $\mathbf{w}_t$

$$P(\mathbf{w}_t \mid US_{1:t}) \propto P(US_t \mid \mathbf{w}_t, US_{1:t-1}) P(\mathbf{w}_t \mid US_{1:t-1})$$

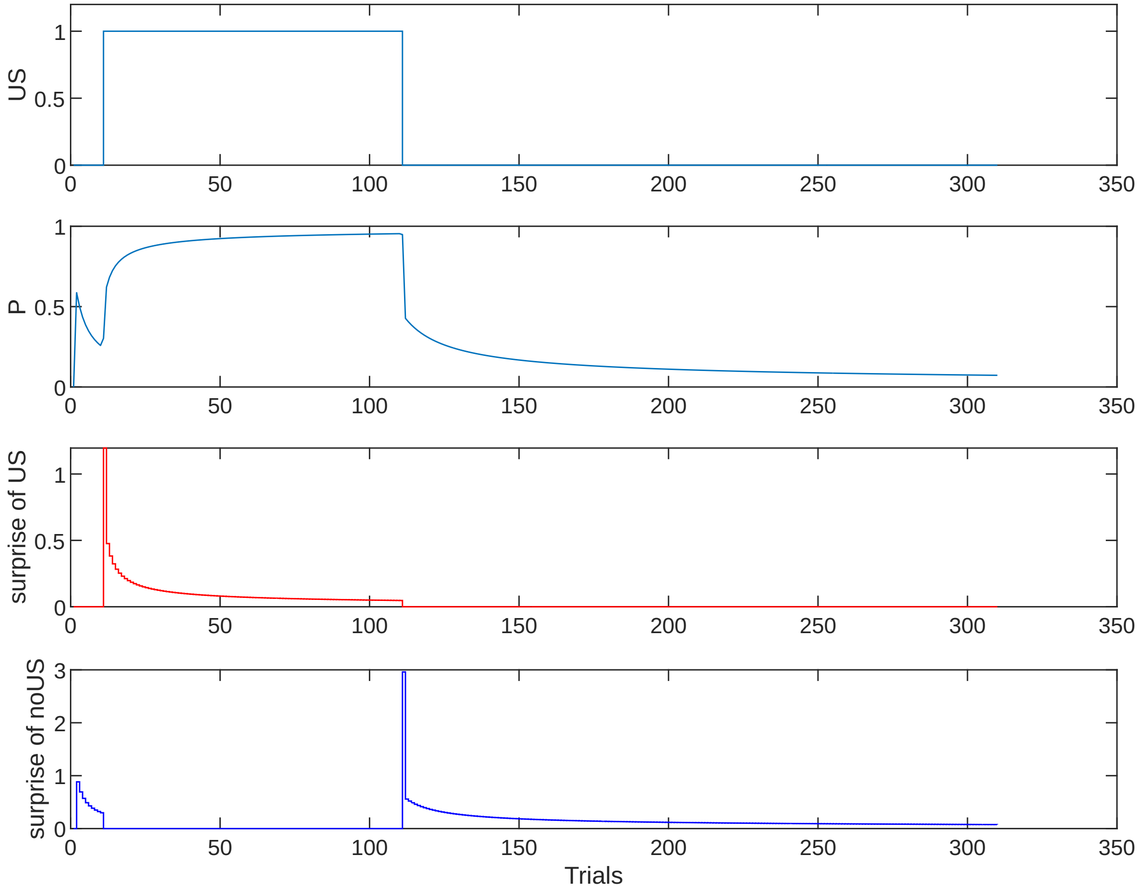
$$= P(US_t \mid \mathbf{w}_t, US_{t-1}) \int P(\mathbf{w}_t \mid \mathbf{w}_{t-1}) P(\mathbf{w}_{t-1} \mid US_{1:t-1}) dw_{t-1}$$

$$\mathbf{w}_t = \arg\max_{w_t} P(\mathbf{w}_t \mid US_{1:t})$$

- **Surprise: information amount**

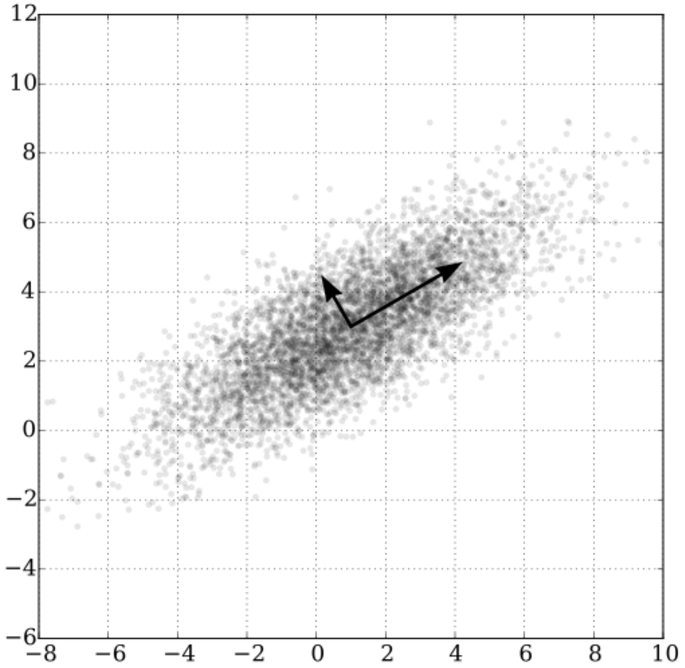
$$S_t(\text{US}) = -\log P(US_t = 1)$$

$$S_t(\text{no-US}) = -\log\left[1 - P(US_t = 1)\right]$$



Bayesian method in Dimensionality reduction

Principle Component Analysis (PCA)



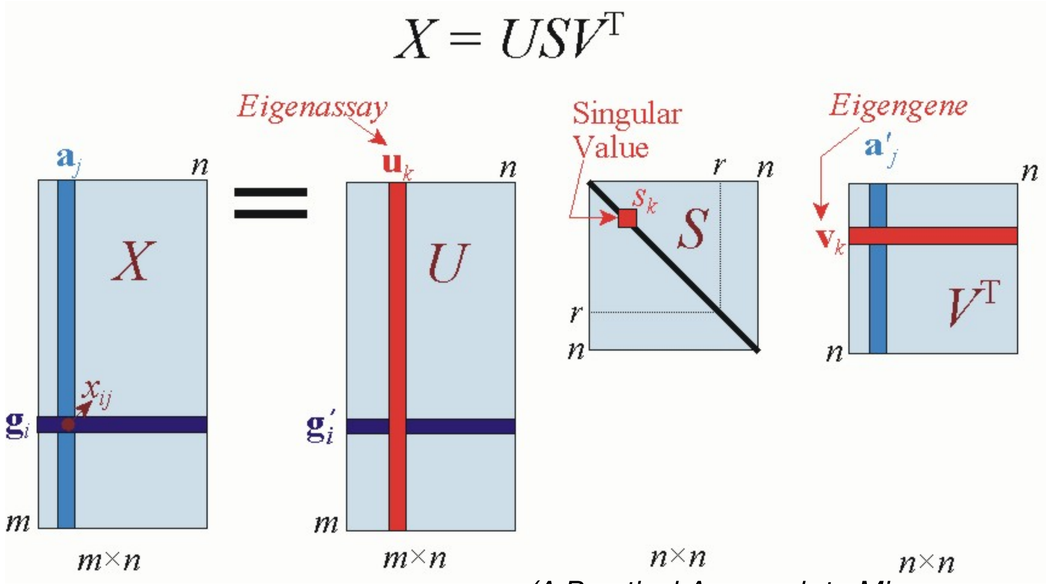
Eigen decomposition on Covariance Matrix

$$M = X^T X$$

$$Mv = \lambda v$$

Singular Vector Decomposition (SVD)

$$X = USV^{\mathrm{T}}$$



(A Practical Approach to Microarray Data Analysis, 2003)

Origin of PCA

(Pearson, 1901; Hotelling, 1933)

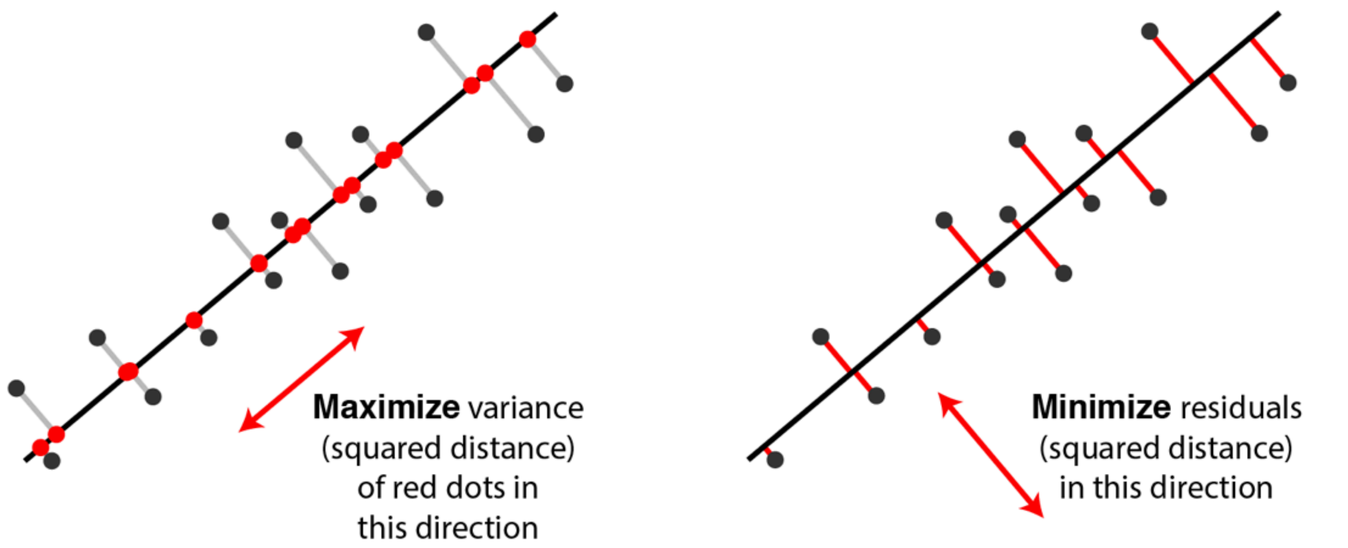
Latent variable model

$$y = Wx + \mu$$

Observation Latent variable

Least-squares error solution

- the same direction of the maximum data variation
- perpendicular to the direction of the maximum variation of the residuals (orthogonal PCs)



(Udell, 2015)

Two equivalent views of principal component analysis.

Maximize variance v.s. Least-squares error

Statistical versions of PCA

Probabilistic PCA

(Tipping and Bishop,1999)

$$\boldsymbol{y} \,|\, \boldsymbol{x} \sim N(\boldsymbol{y} \,|\, Wx + \mu, \tau^{-1} I)$$
$$\boldsymbol{x} \sim N(\boldsymbol{x} \,|\, 0, I)$$

Maximum likelihood solution

$$W_{ML} = U_q(\Lambda_q - \tau^{-1} I)^{1/2}$$

Factor analysis

(Joreskog, 1983)

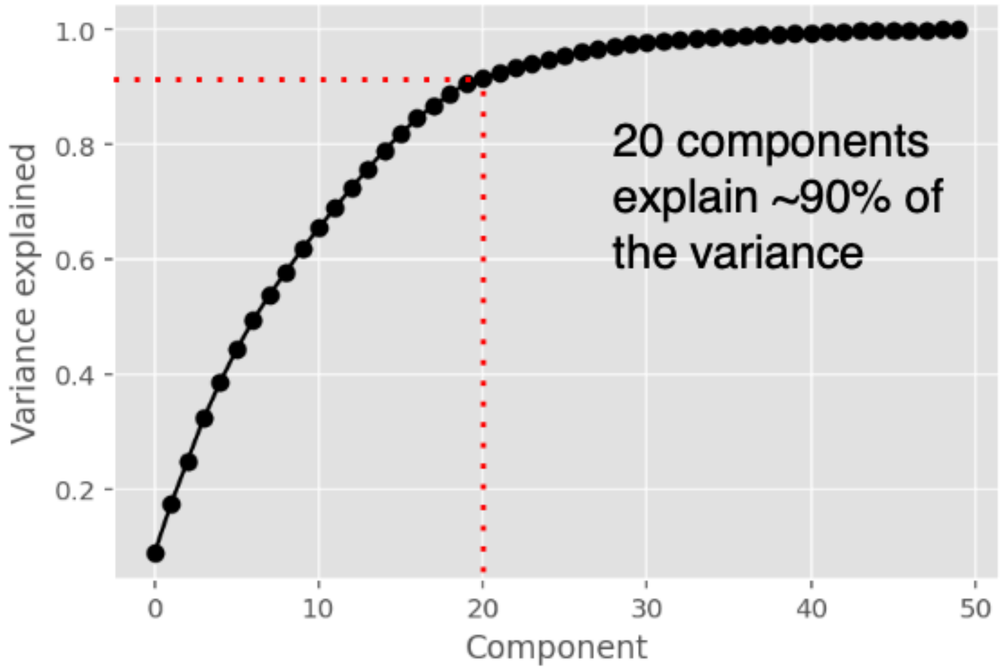
$$\boldsymbol{y} \,|\, \boldsymbol{x} \sim N(\boldsymbol{y} \,|\, Wx + \mu, diag(\boldsymbol{\tau}_D^{-1}))$$
$$\boldsymbol{x} \sim N(\boldsymbol{x} \,|\, 0, I)$$

No close form solution for Maximum likelihood

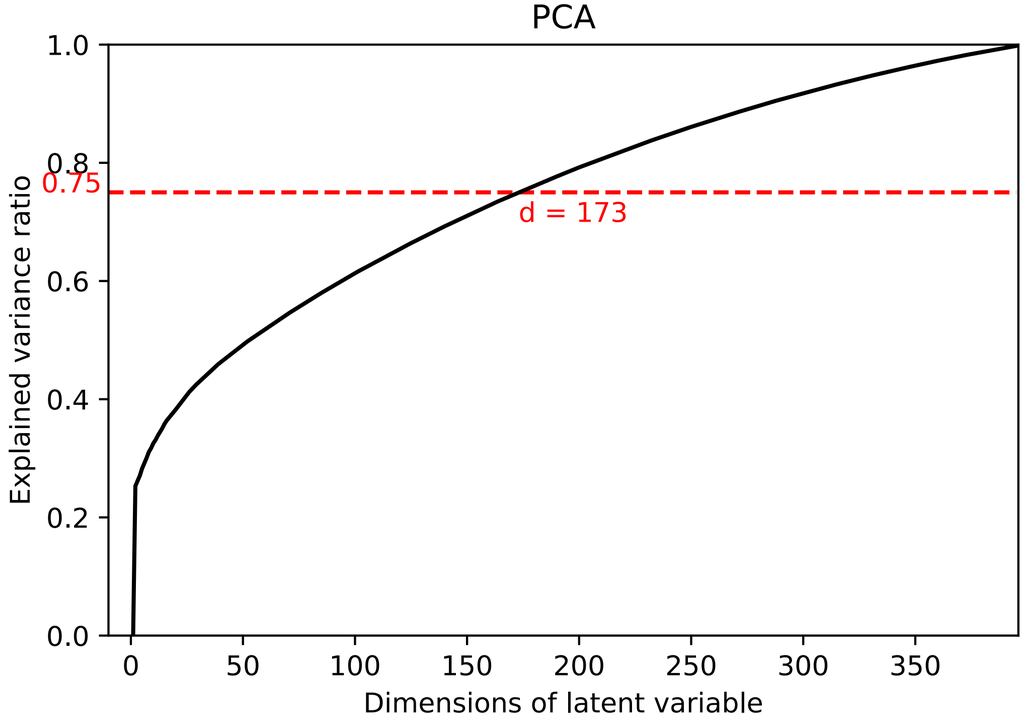
EM algorithm

Practical difficulty: How many PCs?

Explained variance ratio



Explained variance ratio in calcium imaging data



How to choose a proper threshold?

Automatic dimensionality reduction: Bayesian PCA

Linear latent variable model

$$P(y \mid x, W) = N(y \mid Wx + \mu, \tau^{-1} I)$$

Loading matrix (weights) Latent variable

Using Bayesian method to solve the model

ARD Prior of W

(automatic relevance determination)

$$P(W \mid \alpha) = \prod_i N(w_i \mid 0, \alpha_i^{-1} I)$$ ← **ARD on W**

$$P(x) = N(x \mid 0, I)$$

automatically cause Sparsity in W

Regularization effects of ARD

Regularization effects:

$$P(y \mid x, W) = N(y \mid Wx + \mu, \tau^{-1} I)$$

$$P(W \mid \alpha) = \prod_i N(w_i \mid 0, \alpha_i^{-1} I)$$

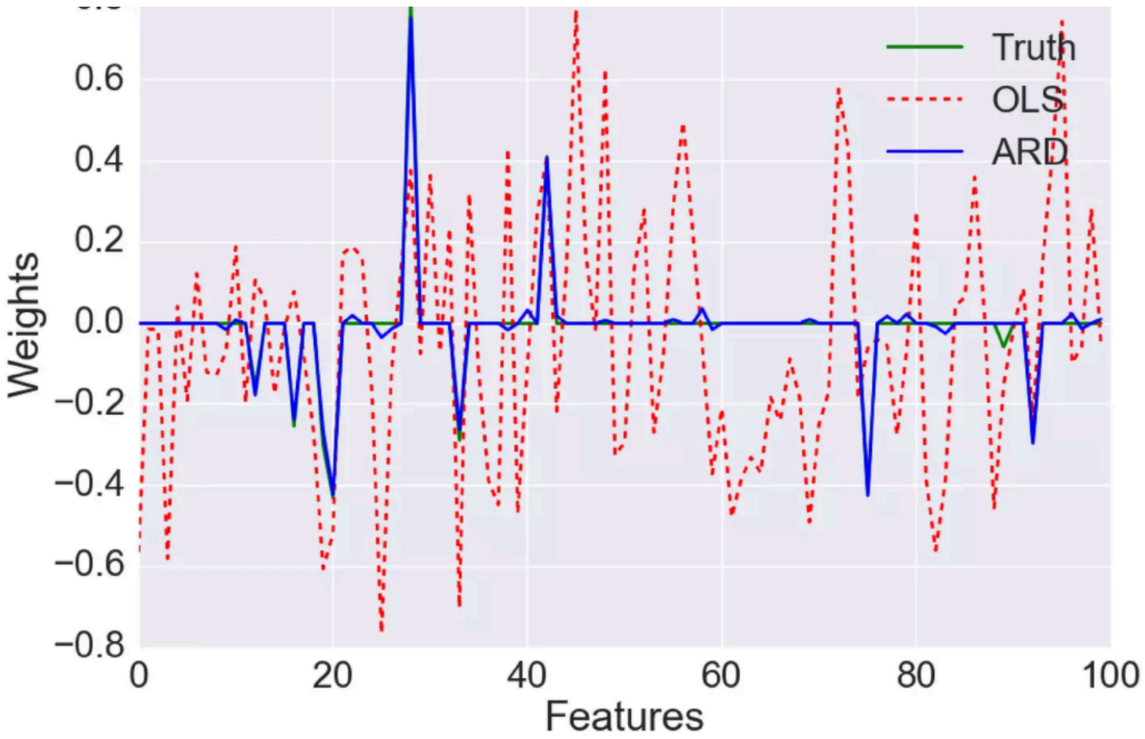
Solution:

Give τ, α**,**

argmin $. \tau \mid\mid Wx + \mu - y \mid\mid^2 + \alpha \mid\mid w \mid\mid^2$

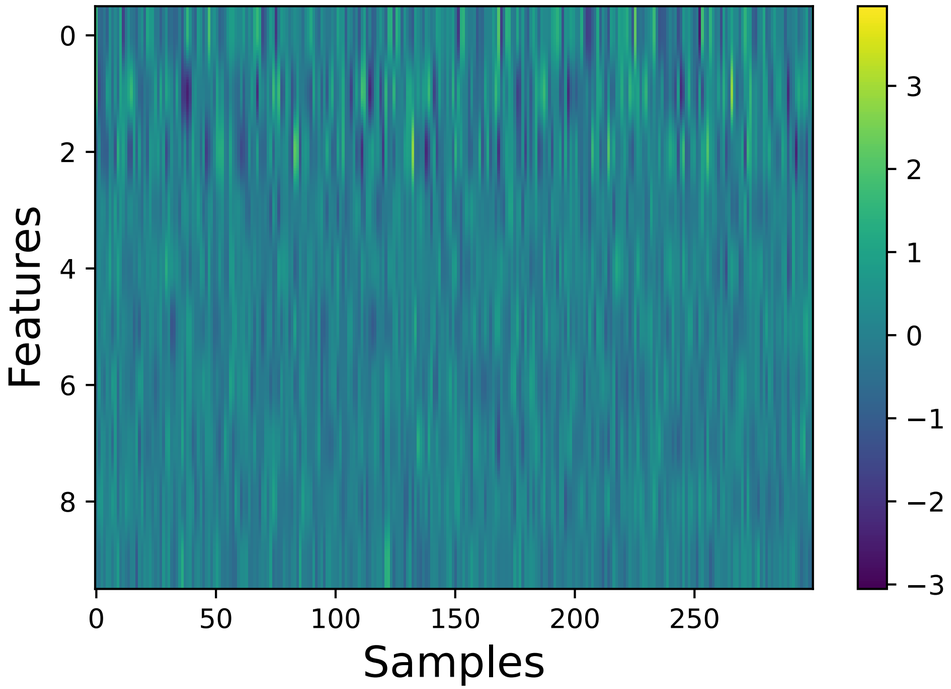
Bayesian Ridge

Sparsity in estimated weights



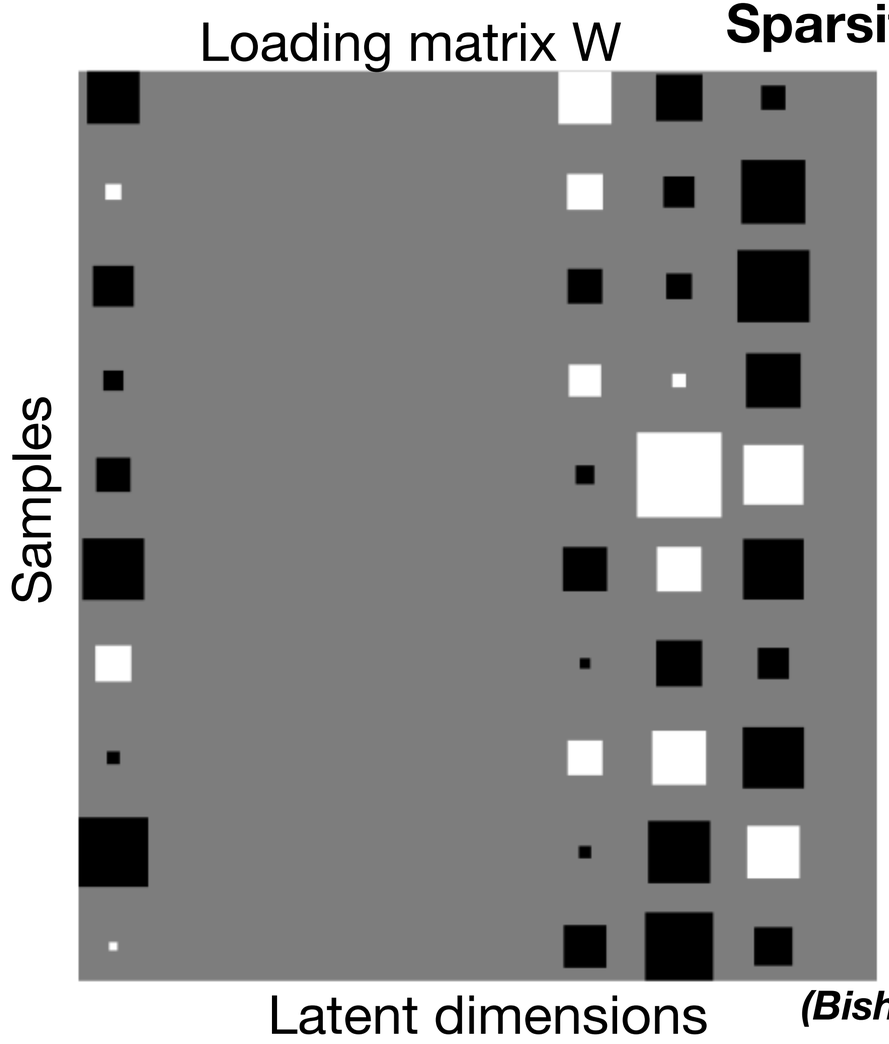
Bayesian PCA works on Gaussian data

Simulated data with Gaussian variance



10 dimensions
the first 3 dimensions: x ~ N(0, 1)
the rest 7 dimensions: x ~ N(0,0.5)

Sparsity in W



(Bishop, 1999)

Sparsity analysis on W

Marginal likelihood

$$L(\boldsymbol{\alpha}) : \log P(\boldsymbol{y}\,|\,\boldsymbol{x},\boldsymbol{\alpha}) = \log \int P(\boldsymbol{t}\,|\,\boldsymbol{x},W)P(W\,|\,\boldsymbol{\alpha})dW$$

$$= -\frac{1}{2}\log|C'| - \frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T C'^{-1}(\boldsymbol{y}-\boldsymbol{\mu}) + \textbf{const.}$$

$$= L(\alpha_{-i}) + l(\alpha_i)$$

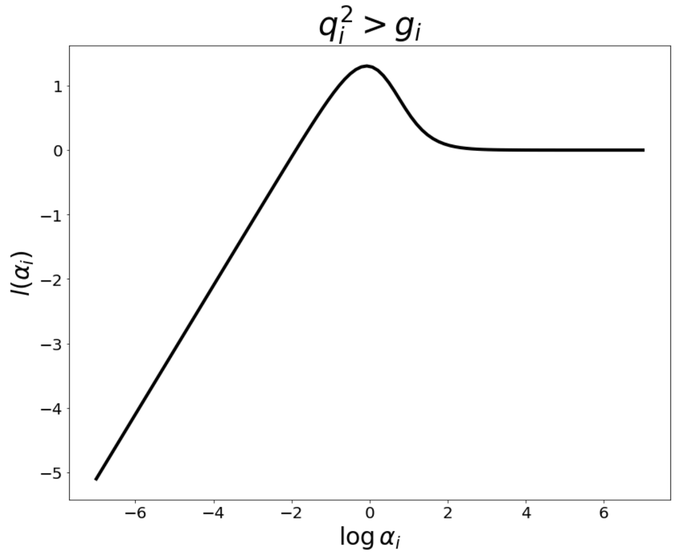
Separate $\alpha_i, \alpha - i$

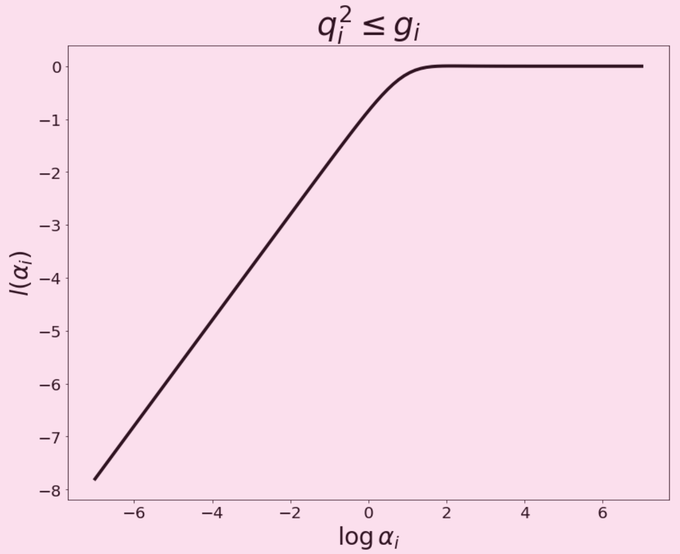
Two stationary points:

$$\alpha_i = \begin{cases} +\infty, \ \textbf{subject to} \ \ q_i'^2 \le g_i' \\ \dfrac{g'^2}{q_i'^2 - g_i'}, \ \textbf{subject to} \ \ q_i'^2 > g_i' \end{cases}$$

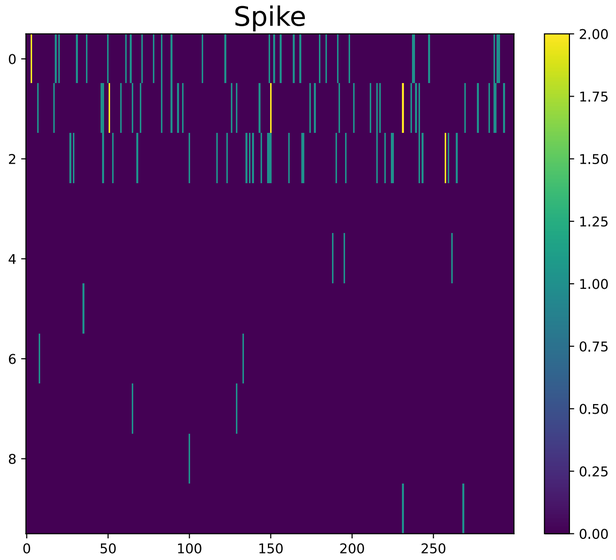
where $g_i' = x_i^2 tr(C_{-i}'^{-1}), q_i'^2 = x_i^2(\boldsymbol{y}-\boldsymbol{\mu})^T C_{-i}'^{-2}(\boldsymbol{y}-\boldsymbol{\mu})$

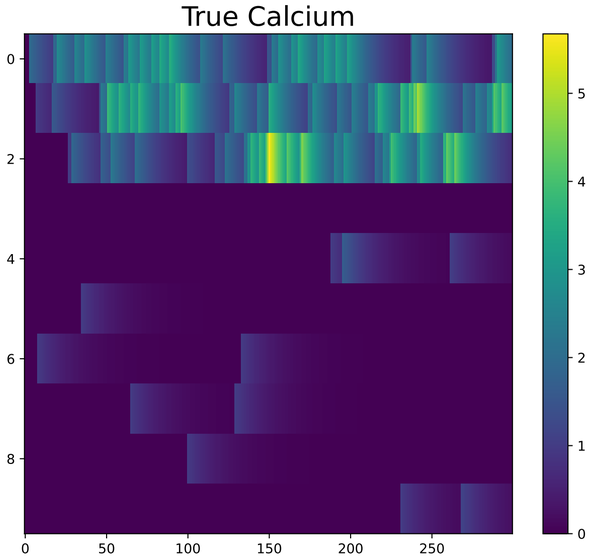
$l(\alpha_i)$

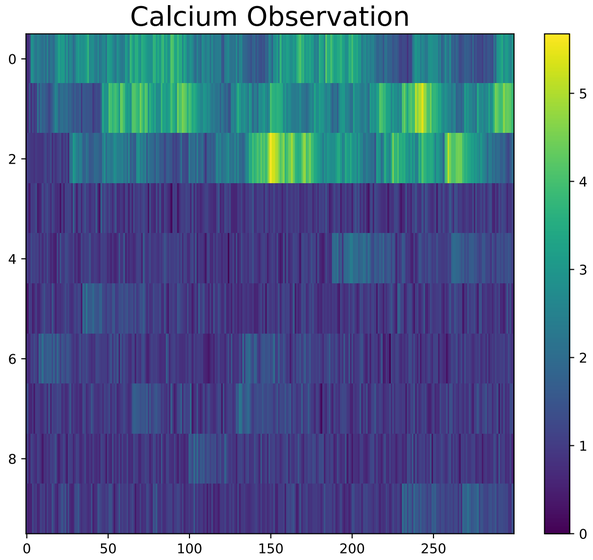




Artificial calcium imaging data







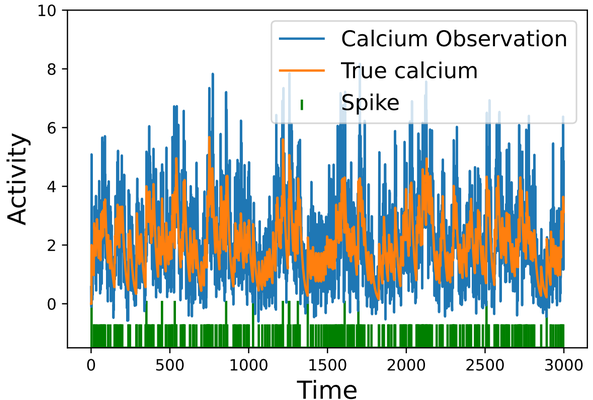
$$s_t \sim Poisson(\lambda)$$

$$c_t = \gamma c_{t-1} + s_t$$

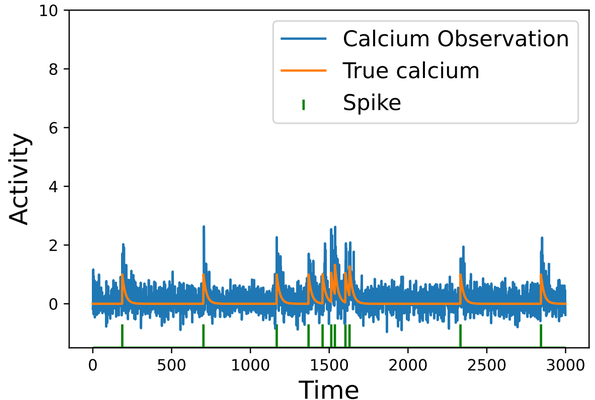
$$y_{i,t} = ac_{i,t} + b + \varepsilon_g$$

$$\varepsilon_g \sim \mathcal{N}(0,\sigma^2)$$

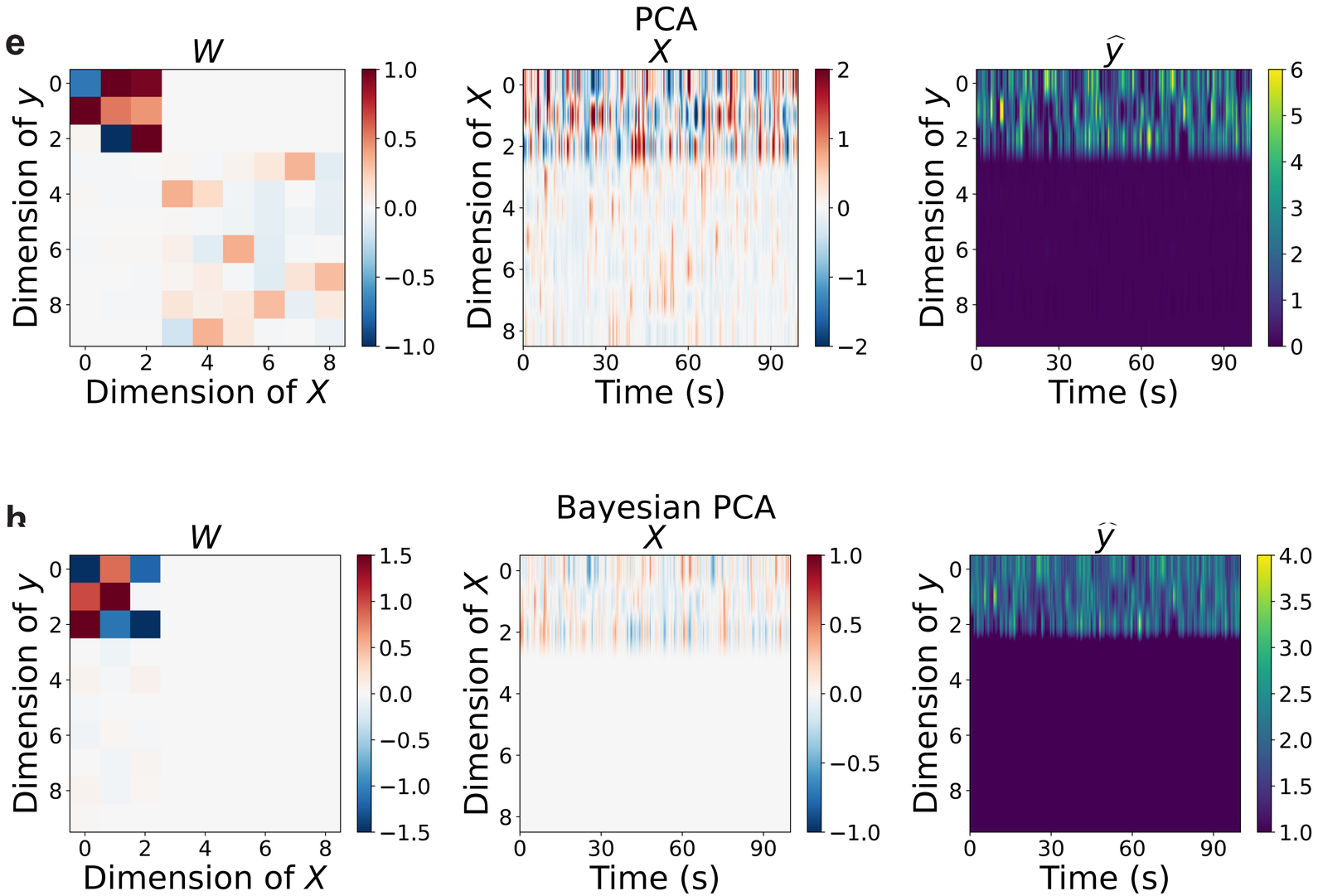
Example neuron with high firing rate



Example neuron with low firing rate



Results


e
W
Dimension of y
Dimension of X
PCA
X
Dimension of X
Time (s)
ŷ
Dimension of y
Time (s)
h
W
Dimension of y
Dimension of X
Bayesian PCA
X
Dimension of X
Time (s)
ŷ
Dimension of y
Time (s)

Summary

- **Bayesian Parameter estimation**
  - Classification: Bayesian Logistic regression
  - Dimensionality reduction: Bayesian PCA
- **Bayesian Logistic regression on behavior in associative learning**
  - Bayesian inference