On the sparse estimation

ATR Computational Neuroscience Laboratories

Masa-aki Sato

Contents

Ill-posed problem

Large number of parameters are estimated
from small number of data

- Maximum Likelihood (Minimum Squared Error)
  - Estimate optimal parameter which maximize likelihood
  - Overfitting:

    Complex models with many adjustable parameters tend to fit noise in training data and degrade generalization ability

- Sparse estimation

  Extract relevant features and discard irrelevant features to attain good generalization ability

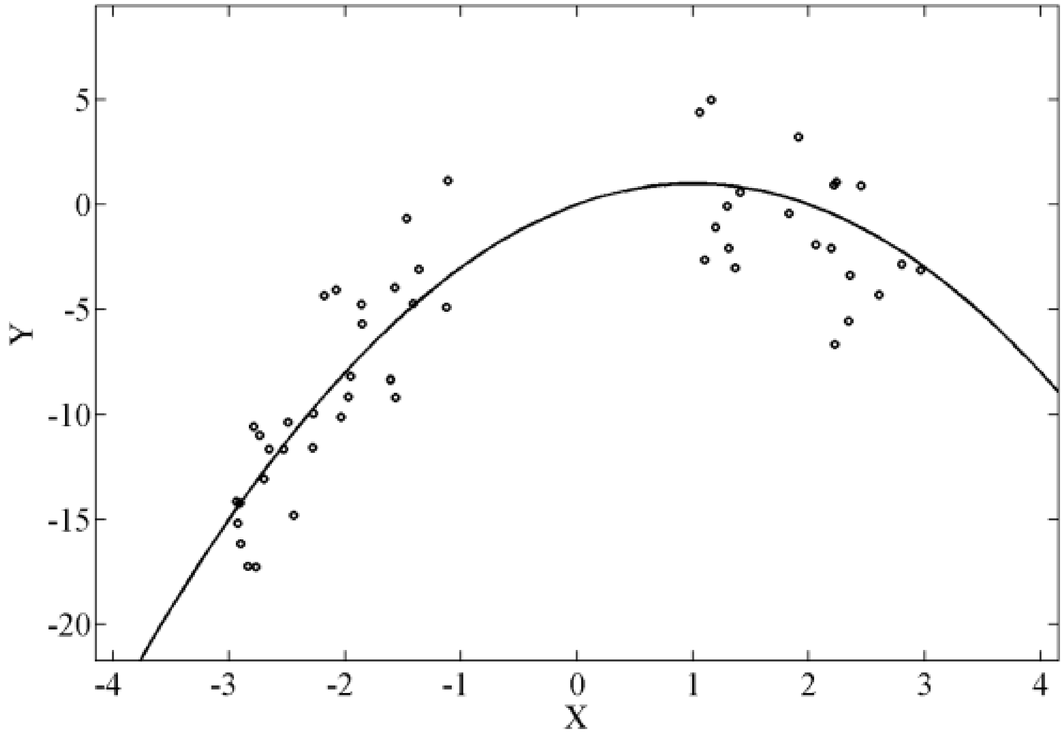
Function approximation by polynomial

$$y = w_0 + w_1 x + w_2 x^2 + \ldots + w_N x^N$$

$$= W \cdot X$$ Linear parameter model

$$X = \left[1,\ x,\ x^2,\ \ldots,\ x^N\right]$$ Input (feature) variable

$$W = \left[w_0,\ w_1,\ w_2,\ \ldots,\ w_N\right]$$ Weight parameter vector



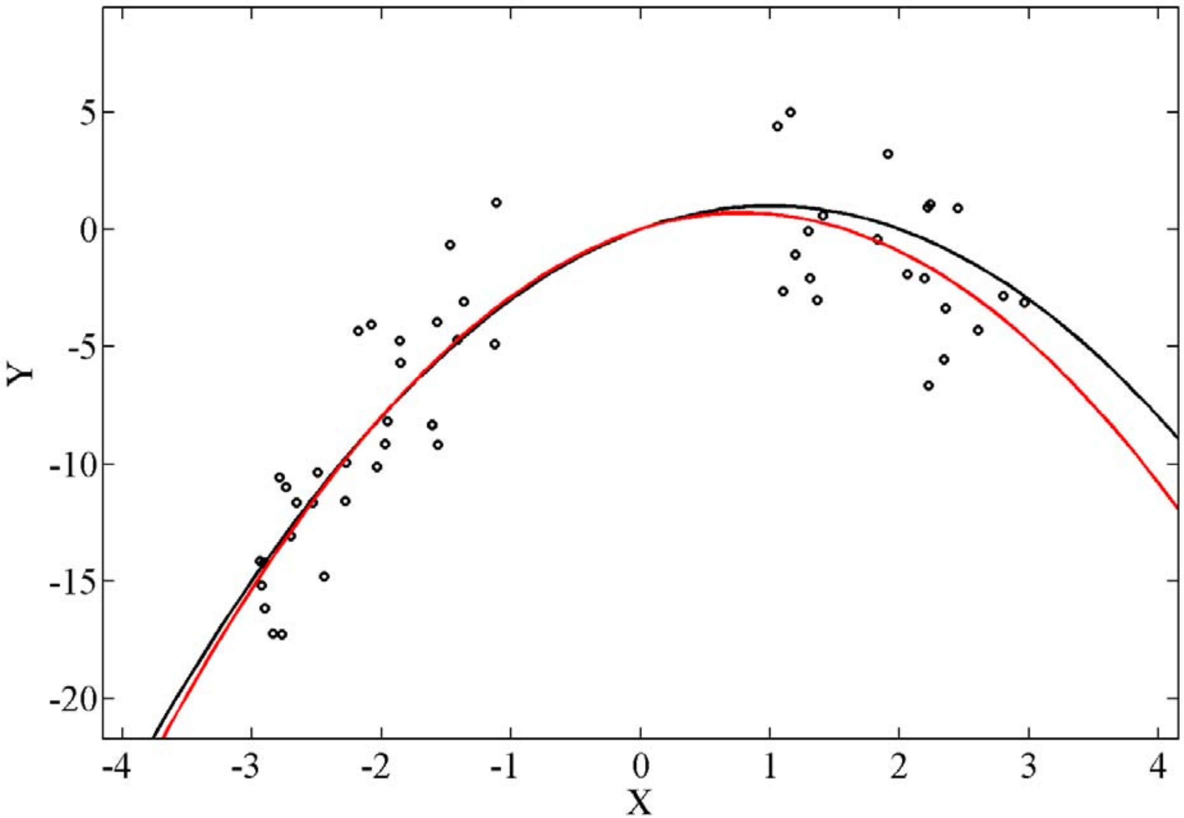
50 data points are randomly generated from quadratic function

$$y = 2x - x^2 + noise$$

Maximum likelihood (Quadratic function)

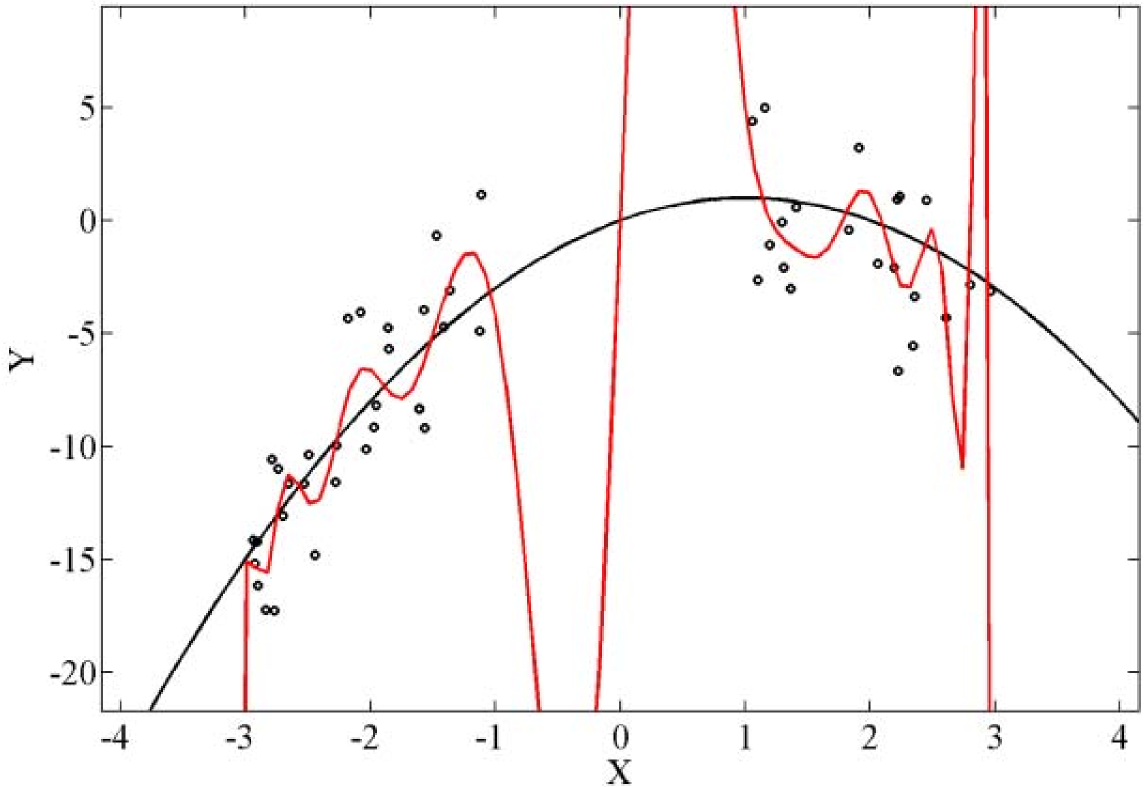
⟺ Minimize Squared Error

$$error = \sum_{t=1}^{T} (y(t) - W \cdot X(t))^2$$ Find optimal W



Maximum likelihood (20 degree polynomial)

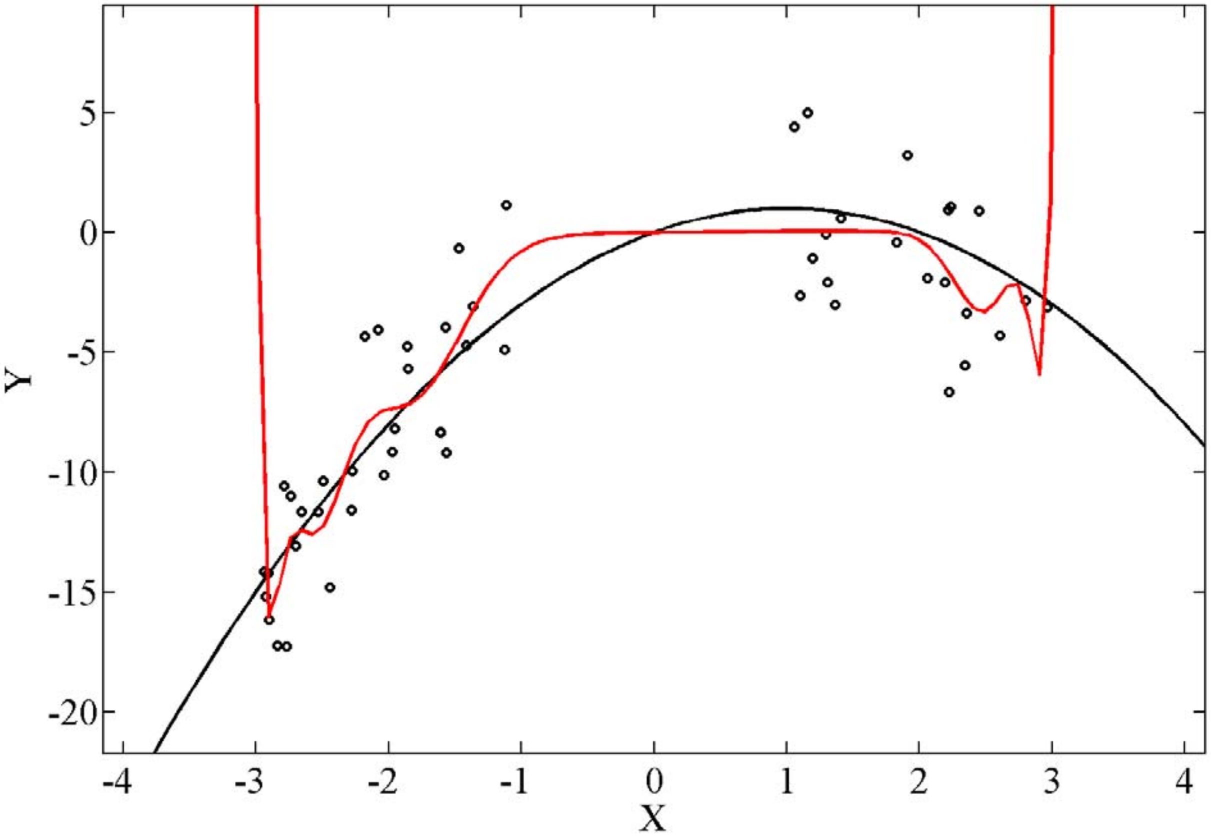
Overfitting : Optimal W fits noise in training data
Generalization is not good



Regularization method (20 degree polynomial)

$$error = \sum_{t=1}^{T} (y(t) - W \cdot X(t))^2 + \alpha \cdot W^2 \quad \text{Minimization}$$

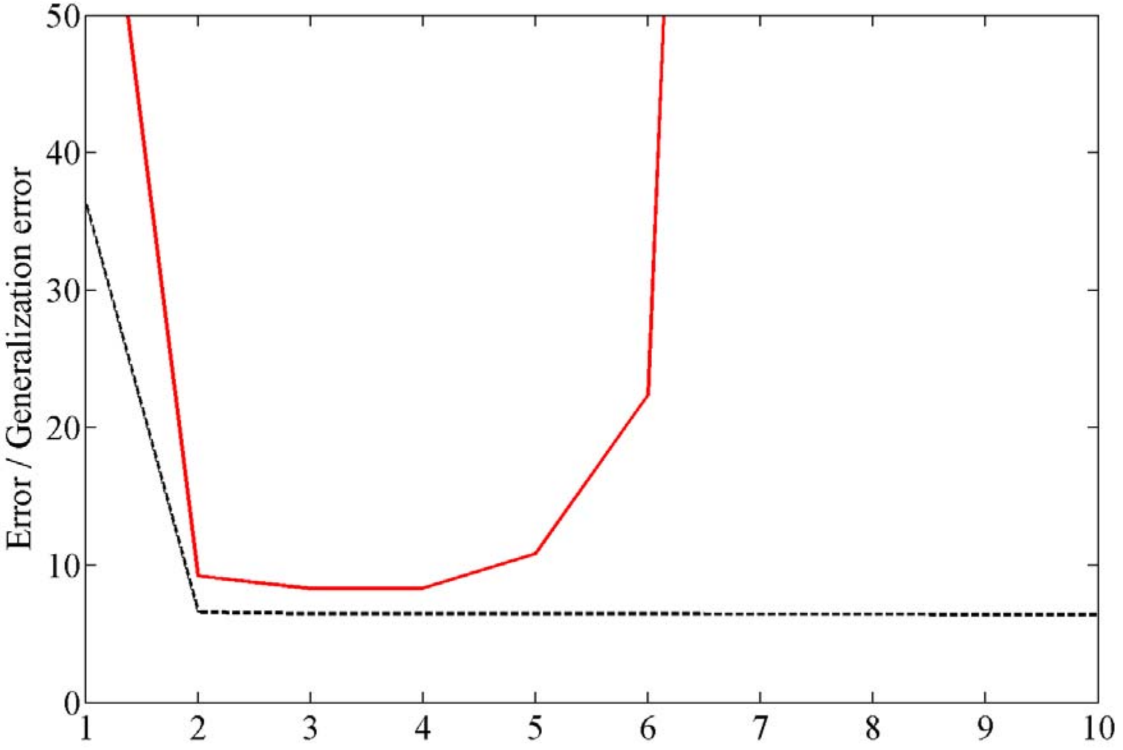
$$\Leftrightarrow \quad \text{Prior :} \quad P(W) \propto \exp\left(-\frac{1}{2}\alpha \cdot W^2\right)$$



Model selection

- Search best model which gives best generalization error by changing number pf parameters (polynomial degree)
- Combinatorial serach is almost impossible for large degree of freedom

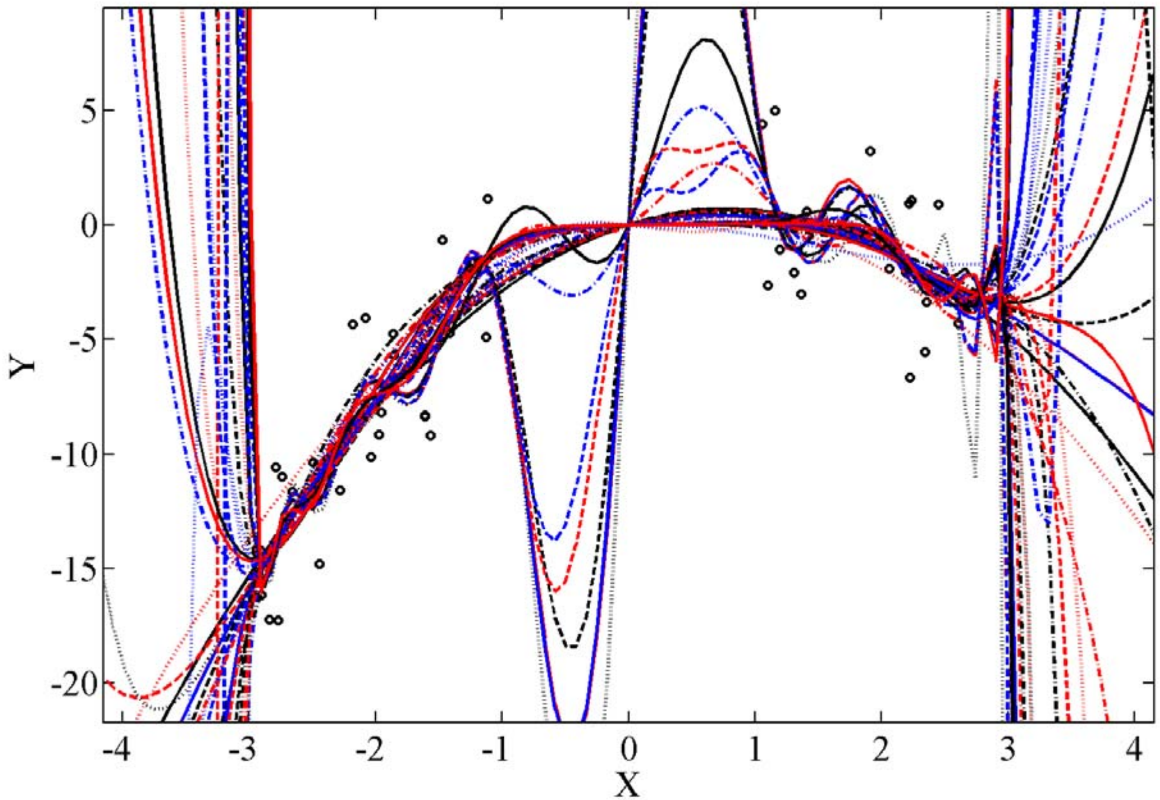
Training error & Generalization error



Number pf parameters (Polynomial degree)

Sparse estimation by Bayesian method

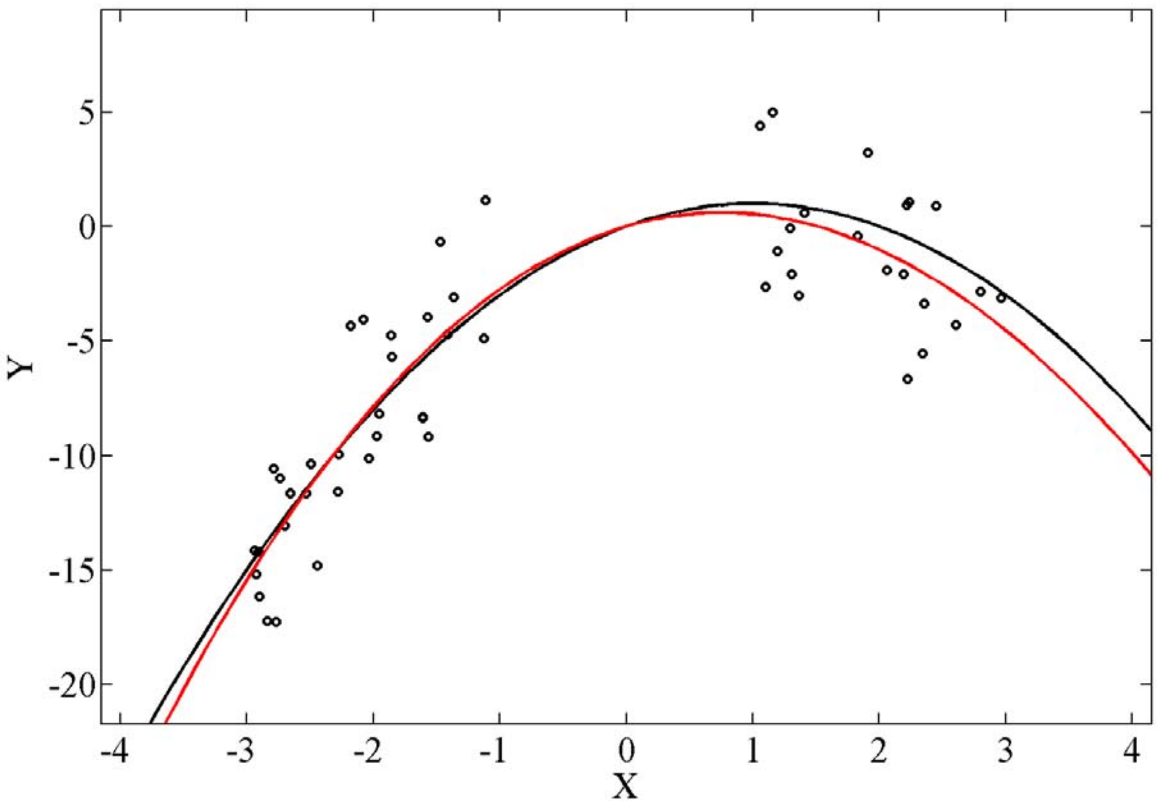
- Parameters are considered as random variable
- Posterior probability is calculated for possible parameter value
- Estimation is done by integrated over possible value according to posterior probability distribution



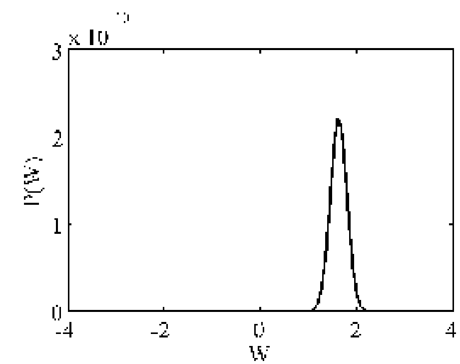
Sparse estimation (20 degree polynomial)

- Precision parameter α_n is introduced for each weight component w_n and estimated from observed data

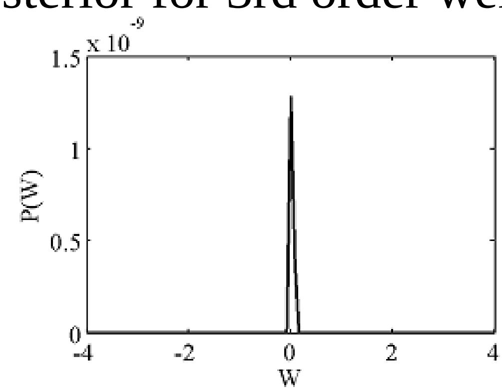
Prior $\quad P(w_n) \propto \exp\left(-\frac{1}{2}\alpha_n \cdot {w_n}^2\right), \;\; \alpha_n = 不定$



Posterior for 2nd order weight



Posterior for 3rd order weight



Sparse Estimation

Prunes irrelevant features in the model and increase generalization ability

MAP / ML Estimation

(Maximum a Posteriori / Maximum Likelihood)

- Posterior $$P(\boldsymbol{\theta} \mid \boldsymbol{X}) = \frac{P(\boldsymbol{X} \mid \boldsymbol{\theta}) P_0(\boldsymbol{\theta})}{P(\boldsymbol{X})}$$

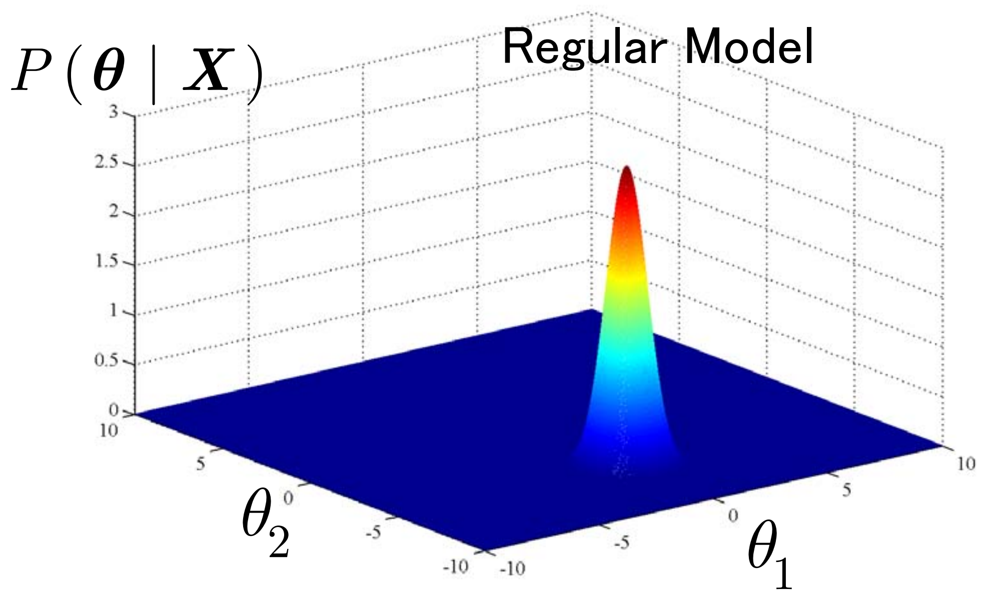
Liklihood: $P(\boldsymbol{X} \mid \boldsymbol{\theta})$

prior: $P_0(\boldsymbol{\theta})$

Marginal likelihood: $P(\boldsymbol{X})$

- Estimate optimal parameter

$$\boldsymbol{\theta}_{\mathrm{MAP}} = \arg\max \log\left(P(\boldsymbol{X} \mid \boldsymbol{\theta}) P_0(\boldsymbol{\theta})\right)$$



Full Bayesian Estimation

- Posterior $$P(\boldsymbol{\theta} \mid \boldsymbol{X}) = \frac{P(\boldsymbol{X} \mid \boldsymbol{\theta}) P_0(\boldsymbol{\theta})}{P(\boldsymbol{X})}$$

Liklihood: $P(\boldsymbol{X} \mid \boldsymbol{\theta})$

prior: $P_0(\boldsymbol{\theta})$

Marginal likelihood: $P(\boldsymbol{X})$

- Estimate posterior parameter distribution and integrate over parameters according to the posterior.

Marginal likelihood $$P(\boldsymbol{X}) = \int d\boldsymbol{\theta} \; P(\boldsymbol{X} \mid \boldsymbol{\theta}) P_0(\boldsymbol{\theta})$$

Estimated parameter $$\overline{\boldsymbol{\theta}} = \int d\boldsymbol{\theta} \; P(\boldsymbol{\theta} \mid \boldsymbol{X}) \; \boldsymbol{\theta}$$

Model reduction
by Bayesian method

Mixture of Gaussian example

Redundant Model (ill-posed problem)

Estimation model is a Mixture of two Gaussian units

$$P(\boldsymbol{x} \mid \boldsymbol{\theta}) = g_0 \cdot N(\boldsymbol{x} \mid \theta_1) + (1 - g_0) \cdot N(\boldsymbol{x} \mid \theta_2)$$

Assume data is generated by a single Gaussian

Single unit model correspond to three cases :

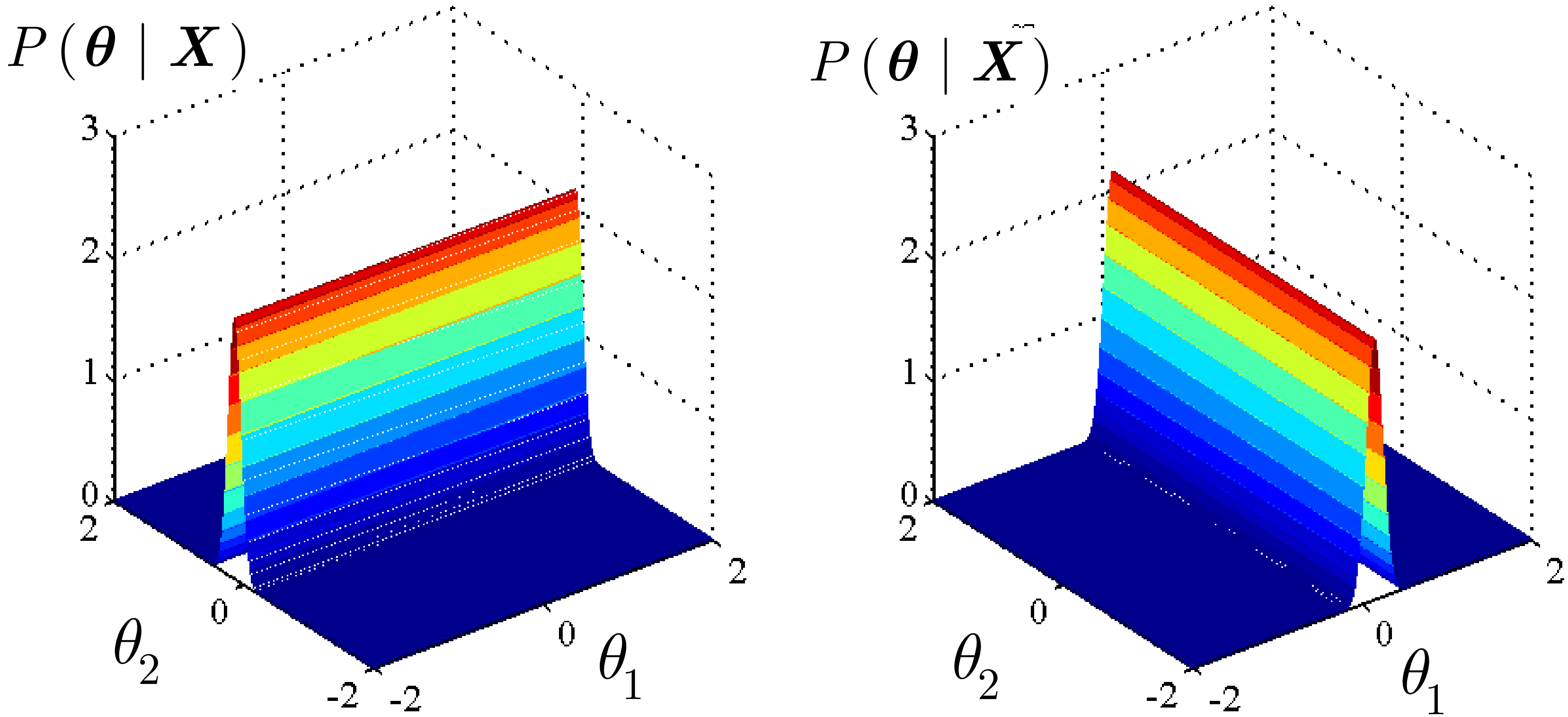
$$P(\boldsymbol{x} \mid \boldsymbol{\theta}) = N(\boldsymbol{x} \mid \theta_1), \quad g_0 = 1 \ , \quad \theta_2 = \text{arbitrary}$$

$$P(\boldsymbol{x} \mid \boldsymbol{\theta}) = N(\boldsymbol{x} \mid \theta_2), \quad g_0 = 0 \ , \quad \theta_1 = \text{arbitrary}$$

$$P(\boldsymbol{x} \mid \boldsymbol{\theta}) = N(\boldsymbol{x} \mid \theta_1), \quad \theta_1 = \theta_2 \ , \quad g_0 = \text{arbitrary}$$

Posterior distribution for redundant model

$g_0 = 1$ $\qquad$ $g_0 = 0$



Fisher Information matrix becomes singular

Pruning of redundant parameters

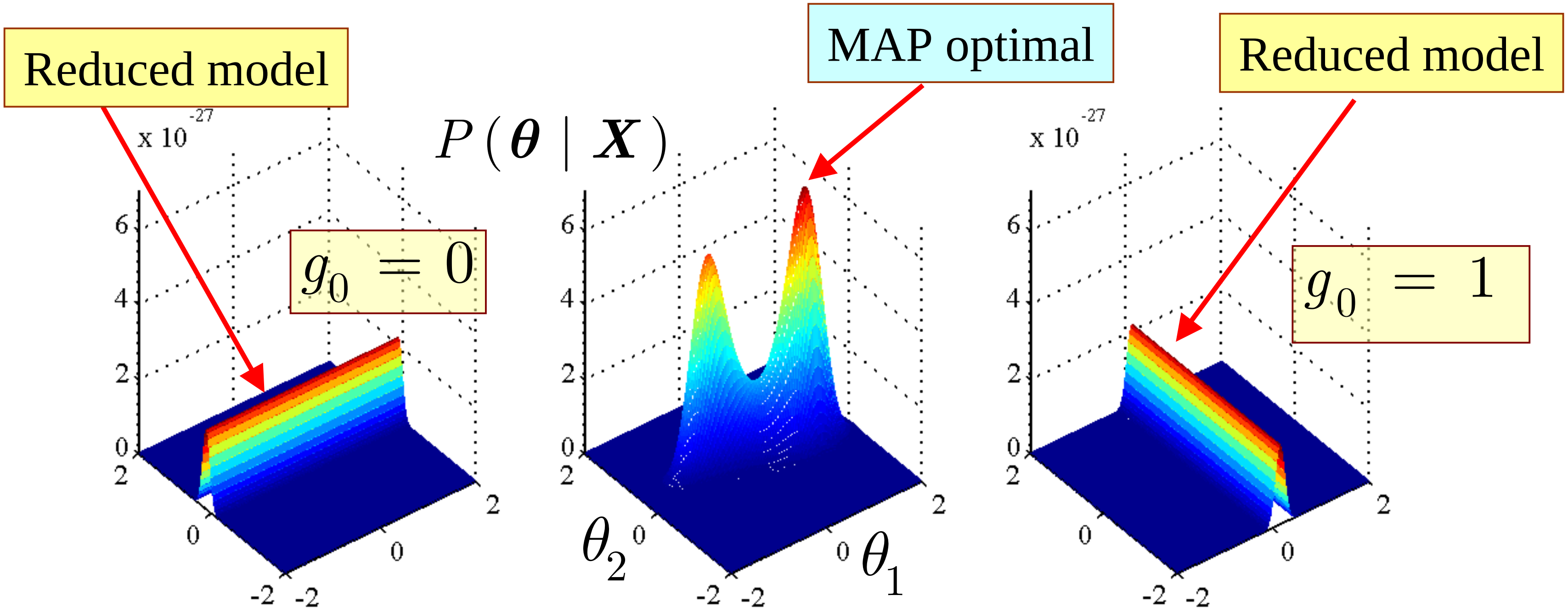
MAP

- Complex models explain a given data better than simpler models and give higher posterior value
- Then all parameters are used for prediction

Full Bayesian

- Reduced simpler model dominates by integration over parameters

Posterior distribution for 100 sample data generated by a single Gaussian model



Parameter pruning
in Sparse Estimation

Prior in Sparse Estimation Model

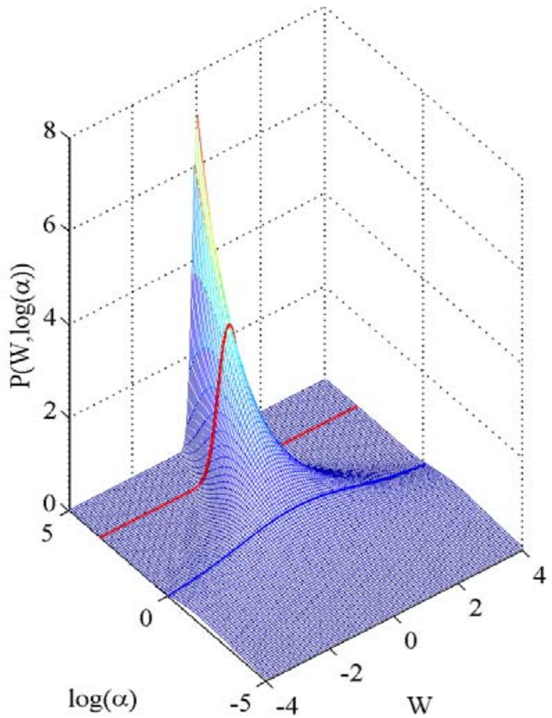
α_n controles a precision (width) of weight parameter distribution

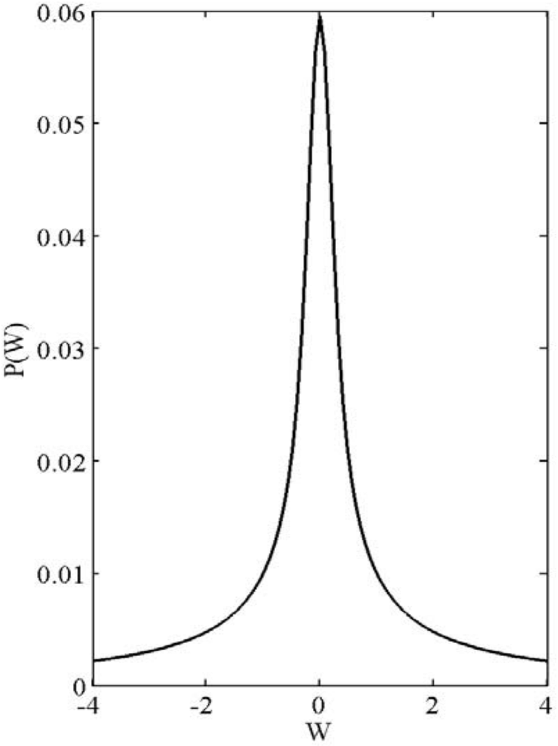
Prior

$$P(w_n \mid \alpha_n) \propto \sqrt{\alpha_n} \cdot \exp\left(-\frac{1}{2}\alpha_n \cdot {w_n}^2\right)$$

$$P(\log(\alpha_n)) = const.$$ (Non-informative prior)

$$\langle w_n \rangle = 0 \;,\; \alpha_n = \text{Arbitrary}$$

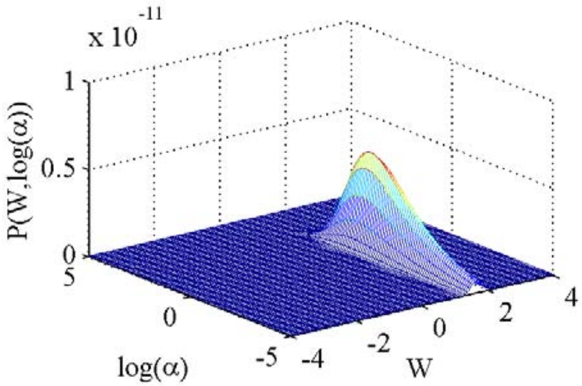


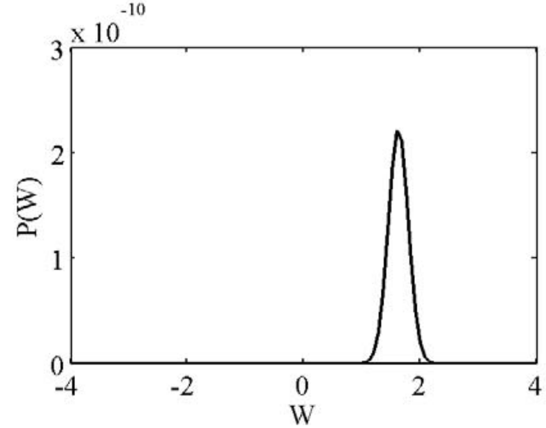


Posterior parameter distribution for (w_n, α_n)

Other parameters are integrated out
And their effects are taken into account

Posterior for relevant parameter

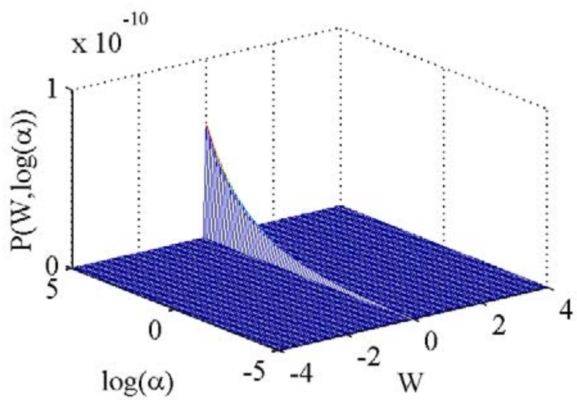


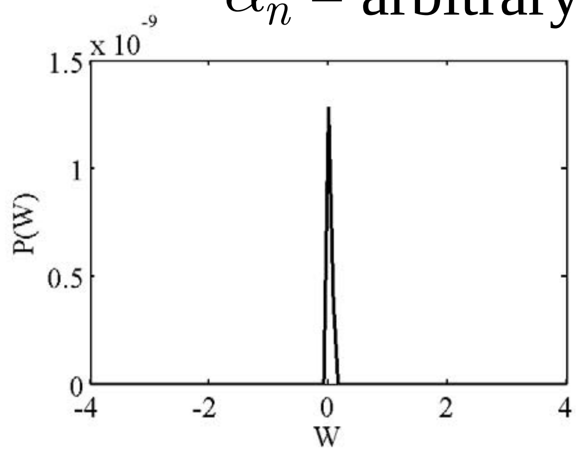


Posterior for irrelevant parameter

$w_n = 0$

α_n = arbitrary





Calculation of posterior distribution

Liklihood (Hierarhical) prior

- Posterior $P(\boldsymbol{J},\boldsymbol{\alpha}|\boldsymbol{B}) = \dfrac{P(\boldsymbol{B}|\boldsymbol{J})P_0(\boldsymbol{J}|\boldsymbol{\alpha})P_0(\boldsymbol{\alpha})}{P(\boldsymbol{B})}$

$$\langle \boldsymbol{J} \rangle = \int \boldsymbol{J} P(\boldsymbol{J},\boldsymbol{\alpha}|\boldsymbol{B}) d\boldsymbol{J} d\boldsymbol{\alpha}$$

Marginal likelihood

Free energy maximization ⇔ Posterior calculation

Distance between trial posterior $Q(\boldsymbol{J},\boldsymbol{\alpha})$ and true posterior $P(\boldsymbol{J},\boldsymbol{\alpha}|\boldsymbol{B})$

$$F[Q(\boldsymbol{J},\boldsymbol{\alpha})] = \ln P(\boldsymbol{B}) - KL[Q(\boldsymbol{J},\boldsymbol{\alpha}) \| P(\boldsymbol{J},\boldsymbol{\alpha}|\boldsymbol{B})]$$

Log marginal likelihood (Evidence)

Variational Bayesian (VB) method

Factorization assumption : $Q(\boldsymbol{J}, \boldsymbol{\alpha}) = Q_{\boldsymbol{J}}(\boldsymbol{J}) Q_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})$

Maximization of $F(Q)$

- Maximization of $F(Q)$ w.r.t. $Q_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})$
- Maximization of $F(Q)$ w.r.t. $Q_J(\boldsymbol{J})$

Repeated until convergence

Posterior distribution $P(\boldsymbol{J} \mid \boldsymbol{B}) \approx Q_{\boldsymbol{J}}(\boldsymbol{J})$ at the maximum

Log marginal likelihood $\log(P(\boldsymbol{B})) \approx$ maximized $F(Q)$

Free energy

Free energy after integration of current distribution

$$F = -\frac{1}{2}\left[\, Tr\left(\boldsymbol{\Sigma}_{VB}^{-1} \cdot \langle \boldsymbol{B} \cdot \boldsymbol{B}' \rangle\right) + \log\left|\boldsymbol{\Sigma}_{VB}\right| \,\right]$$

MEG covariance matrix $\langle \boldsymbol{B} \cdot \boldsymbol{B}' \rangle = \boldsymbol{G} \cdot \left\langle \boldsymbol{J}_0 \cdot \boldsymbol{J}_0{}' \right\rangle \cdot \boldsymbol{G}' + \sigma_0 \boldsymbol{I}$

Estimated MEG covariance $\boldsymbol{\Sigma}_{VB} = \boldsymbol{G} \cdot \boldsymbol{W} \cdot \boldsymbol{\alpha} \cdot \boldsymbol{W}' \cdot \boldsymbol{G}' + \sigma \boldsymbol{I}$

Optimal condition (Free energy maximum)

$$\alpha(n) = \frac{\left\langle \bar{\boldsymbol{J}} \cdot \bar{\boldsymbol{J}}' \right\rangle}{\mathcal{G}_{VB}(n)}$$

Estimation gain

Variational Bayesian method

Posterior calculation is converted to free energy maximization

Free energy = (Likelihood)+ (Model complexity)

Likelihood

$$L = -\frac{1}{2}\left[(Y - W \cdot X)^2 + W' \cdot \alpha \cdot W\right]$$

$$= -\frac{1}{2}\left[(Y)^2 - (Y \cdot X)^2 \cdot \left(X \cdot X' + \alpha\right)^{-1}\right]$$

Error

Decreasing function of α

Model complexity

$$H = -\frac{1}{2}\log\left|X \cdot X' \cdot \alpha^{-1} + 1\right|$$

$$\to -\frac{1}{2} N \cdot \log(T) \quad \text{for finite } \alpha, T \gg 1$$

Increasing function of α

BIC

One dimensional case

$$H = -\frac{1}{2}\log\left|X \cdot X' \cdot \alpha^{-1} + 1\right|$$

$$\rightarrow 0 \qquad as \quad \alpha \rightarrow \infty$$

Genaral dimension

$$H = -\frac{1}{2}\log\left|X \cdot X' \cdot \alpha^{-1} + 1\right|$$

$$\rightarrow -\frac{1}{2}N_{eff} \cdot \log(T)$$

N_{eff} Number of finite α , T $\gg$ 1

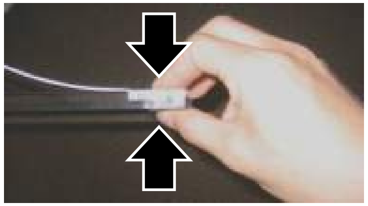
Example of sparse estimation for real problem

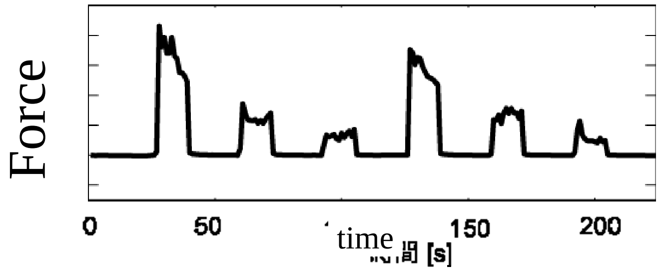
I Nambu, R Osu, M Sato, S Ando, M Kawato, E Naito

Single-trial reconstruction of !nger-pinch forces from human motor-cortical activation measured by near-infrared spectroscopy (NIRS)

NeuroImage 47 (2009) 628.637

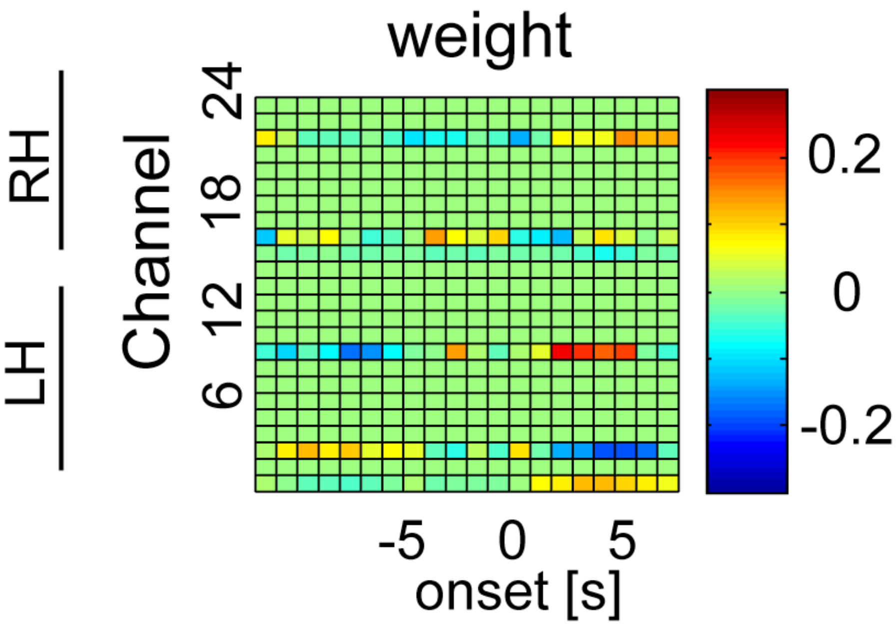
Sparse estimation



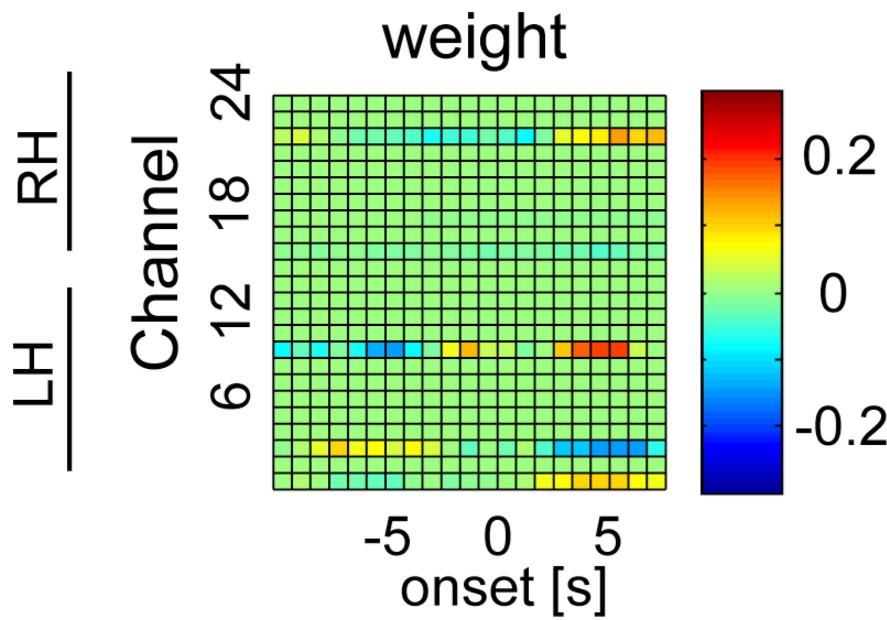


Estimate pinching force
from 24ch x 21 (sec) NIRS data
(Nambu et al)

Estimated weight
by brute force model search



Estimated weight
by sparse estimation



Estimated force from NIRS data