Relevance Vector Machine

Evgeny Burnaev

Skoltech, Moscow, Russia

Skolkovo Institute of Science and Technology

Outline

- Linear Basis Function Models

$$f(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w} \cdot \boldsymbol{\phi}(\mathbf{x})^\top$$

  where $\boldsymbol{\phi}(\mathbf{x})$ is a vector of known basis functions $\phi_j(\mathbf{x})$
- Typical basis functions
- We assume that parameters of basis functions are fixed to some known values

- Linear Basis Function Models

$$f(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w} \cdot \boldsymbol{\phi}(\mathbf{x})^\top$$

  where $\boldsymbol{\phi}(\mathbf{x})$ is a vector of known basis functions $\phi_j(\mathbf{x})$

- Typical basis functions

- We assume that parameters of basis functions are fixed to some known values

- Linear Basis Function Models

$$f(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w} \cdot \boldsymbol{\phi}(\mathbf{x})^\top$$

where $\boldsymbol{\phi}(\mathbf{x})$ is a vector of known basis functions $\phi_j(\mathbf{x})$

- Typical basis functions

$$\phi_j(\mathbf{x}) = x_{j_1}^{j_0},$$

- We assume that parameters of basis functions are fixed to some known values

- Linear Basis Function Models

$$f(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w} \cdot \boldsymbol{\phi}(\mathbf{x})^\top$$

where $\boldsymbol{\phi}(\mathbf{x})$ is a vector of known basis functions $\phi_j(\mathbf{x})$

- Typical basis functions

$$\phi_j(\mathbf{x}) = x_{j_1}^{j_0},\ \phi_j(\mathbf{x}) = \exp\left\{-\frac{\|\mathbf{x} - \boldsymbol{\mu}_j\|^2}{2s^2}\right\}$$

- We assume that parameters of basis functions are fixed to some known values

- Linear Basis Function Models

$$f(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w} \cdot \boldsymbol{\phi}(\mathbf{x})^\top$$

where $\boldsymbol{\phi}(\mathbf{x})$ is a vector of known basis functions $\phi_j(\mathbf{x})$

- Typical basis functions

$$\phi_j(\mathbf{x}) = x_{j_1}^{j_0},\ \phi_j(\mathbf{x}) = \exp\left\{-\frac{\|\mathbf{x} - \boldsymbol{\mu}_j\|^2}{2s^2}\right\}$$

$$\phi(\mathbf{x}) = \sigma\left(\boldsymbol{\mu}_{j,1} \cdot \mathbf{x}^\top + \mu_{j,0}\right),$$

- We assume that parameters of basis functions are fixed to some known values

- Linear Basis Function Models

$$f(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w} \cdot \boldsymbol{\phi}(\mathbf{x})^\top$$

where $\boldsymbol{\phi}(\mathbf{x})$ is a vector of known basis functions $\phi_j(\mathbf{x})$

- Typical basis functions

$$\phi_j(\mathbf{x}) = x_{j_1}^{j_0},\ \phi_j(\mathbf{x}) = \exp\left\{-\frac{\|\mathbf{x} - \boldsymbol{\mu}_j\|^2}{2s^2}\right\}$$

$$\phi(\mathbf{x}) = \sigma\left(\boldsymbol{\mu}_{j,1} \cdot \mathbf{x}^\top + \mu_{j,0}\right),\ \sigma(a) = \frac{1}{1 + e^{-a}}$$

- We assume that parameters of basis functions are fixed to some known values

- Linear Basis Function Models

$$f(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w} \cdot \boldsymbol{\phi}(\mathbf{x})^\top$$

where $\boldsymbol{\phi}(\mathbf{x})$ is a vector of known basis functions $\phi_j(\mathbf{x})$

- Typical basis functions

$$\phi_j(\mathbf{x}) = x_{j_1}^{j_0},\ \phi_j(\mathbf{x}) = \exp\left\{-\frac{\|\mathbf{x} - \boldsymbol{\mu}_j\|^2}{2s^2}\right\}$$

$$\phi(\mathbf{x}) = \sigma\left(\boldsymbol{\mu}_{j,1} \cdot \mathbf{x}^\top + \mu_{j,0}\right),\ \sigma(a) = \frac{1}{1 + e^{-a}}$$

- We assume that parameters of basis functions are fixed to some known values

- Data model for y (ε is a Gaussian white noise with variance β^{-1})

$$y = f(\mathbf{x}, \mathbf{w}) + \varepsilon$$

- For $\mathbf{Y}_m = \{y_1, \ldots, y_m\}$ and $\mathbf{X}_m = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ data likelihood

$$p(\mathbf{Y}_m | \mathbf{X}_m, \mathbf{w}, \beta) = \prod_{i=1}^{m} \mathcal{N}(y_i | \mathbf{w} \cdot \boldsymbol{\phi}(\mathbf{x}_i)^\top, \beta^{-1})$$

- Data log-likelihood has the form

- Data model for y (ε is a Gaussian white noise with variance β^{-1})

$$y = f(\mathbf{x}, \mathbf{w}) + \varepsilon$$
$$p(y|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(y|f(\mathbf{x}, \mathbf{w}), \beta^{-1})$$

- For $\mathbf{Y}_m = \{y_1, \ldots, y_m\}$ and $\mathbf{X}_m = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ data likelihood

$$p(\mathbf{Y}_m|\mathbf{X}_m, \mathbf{w}, \beta) = \prod_{i=1}^{m} \mathcal{N}(y_i|\mathbf{w} \cdot \boldsymbol{\phi}(\mathbf{x}_i)^\top, \beta^{-1})$$

- Data log-likelihood has the form

- Data model for y (ε is a Gaussian white noise with variance β^{-1})

$$y = f(\mathbf{x}, \mathbf{w}) + \varepsilon$$
$$p(y|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(y|f(\mathbf{x}, \mathbf{w}), \beta^{-1})$$

- For $\mathbf{Y}_m = \{y_1, \ldots, y_m\}$ and $\mathbf{X}_m = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ data likelihood

$$p(\mathbf{Y}_m|\mathbf{X}_m, \mathbf{w}, \beta) = \prod_{i=1}^{m} \mathcal{N}(y_i|\mathbf{w} \cdot \boldsymbol{\phi}(\mathbf{x}_i)^\top, \beta^{-1})$$

- Data log-likelihood has the form

- Data model for y (ε is a Gaussian white noise with variance β^{-1})

$$
\begin{aligned}
y &= f(\mathbf{x}, \mathbf{w}) + \varepsilon \\
p(y|\mathbf{x}, \mathbf{w}, \beta) &= \mathcal{N}(y|f(\mathbf{x}, \mathbf{w}), \beta^{-1})
\end{aligned}
$$

- For $\mathbf{Y}_m = \{y_1, \ldots, y_m\}$ and $\mathbf{X}_m = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ data likelihood

$$
p(\mathbf{Y}_m|\mathbf{X}_m, \mathbf{w}, \beta) = \prod_{i=1}^{m} \mathcal{N}(y_i|\mathbf{w} \cdot \boldsymbol{\phi}(\mathbf{x}_i)^\top, \beta^{-1})
$$

- Data log-likelihood has the form

- Data model for y (ε is a Gaussian white noise with variance β^{-1})

$$
\begin{aligned}
&y = f(\mathbf{x}, \mathbf{w}) + \varepsilon \\
&p(y|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(y|f(\mathbf{x}, \mathbf{w}), \beta^{-1})
\end{aligned}
$$

- For $\mathbf{Y}_m = \{y_1, \dots, y_m\}$ and $\mathbf{X}_m = \{\mathbf{x}_1, \dots, \mathbf{x}_m\}$ data likelihood

$$
p(\mathbf{Y}_m|\mathbf{X}_m, \mathbf{w}, \beta) = \prod_{i=1}^{m} \mathcal{N}(y_i|\mathbf{w} \cdot \boldsymbol{\phi}(\mathbf{x}_i)^\top, \beta^{-1})
$$

- Data log-likelihood has the form

$$
\log p(\mathbf{Y}_m|\mathbf{X}_m, \mathbf{w}, \beta) = \sum_{i=1}^{m} \log \mathcal{N}(y_i|\mathbf{w} \cdot \boldsymbol{\phi}(\mathbf{x}_i), \beta^{-1})
$$

- Data model for y (ε is a Gaussian white noise with variance β^{-1})

$$y = f(\mathbf{x}, \mathbf{w}) + \varepsilon$$
$$p(y|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(y|f(\mathbf{x}, \mathbf{w}), \beta^{-1})$$

- For $\mathbf{Y}_m = \{y_1, \ldots, y_m\}$ and $\mathbf{X}_m = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ data likelihood

$$p(\mathbf{Y}_m|\mathbf{X}_m, \mathbf{w}, \beta) = \prod_{i=1}^{m} \mathcal{N}(y_i|\mathbf{w} \cdot \boldsymbol{\phi}(\mathbf{x}_i)^\top, \beta^{-1})$$

- Data log-likelihood has the form

$$\log p(\mathbf{Y}_m|\mathbf{X}_m, \mathbf{w}, \beta) = \sum_{i=1}^{m} \log \mathcal{N}(y_i|\mathbf{w} \cdot \boldsymbol{\phi}(\mathbf{x}_i), \beta^{-1})$$
$$= \frac{m}{2} \log \beta - \frac{m}{2} \log(2\pi) - \beta E_D(\mathbf{w})$$

- Data model for y (ε is a Gaussian white noise with variance β^{-1})

$$y = f(\mathbf{x}, \mathbf{w}) + \varepsilon$$
$$p(y|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(y|f(\mathbf{x}, \mathbf{w}), \beta^{-1})$$

- For $\mathbf{Y}_m = \{y_1, \ldots, y_m\}$ and $\mathbf{X}_m = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ data likelihood

$$p(\mathbf{Y}_m|\mathbf{X}_m, \mathbf{w}, \beta) = \prod_{i=1}^{m} \mathcal{N}(y_i|\mathbf{w} \cdot \boldsymbol{\phi}(\mathbf{x}_i)^\top, \beta^{-1})$$

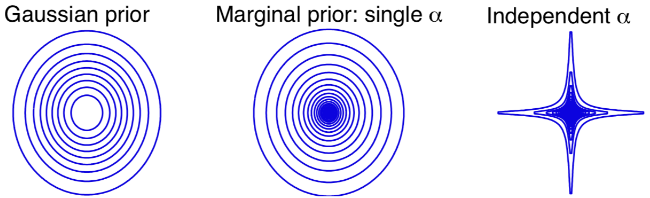
- Data log-likelihood has the form

$$\log p(\mathbf{Y}_m|\mathbf{X}_m, \mathbf{w}, \beta) = \sum_{i=1}^{m} \log \mathcal{N}(y_i|\mathbf{w} \cdot \boldsymbol{\phi}(\mathbf{x}_i), \beta^{-1})$$
$$= \frac{m}{2} \log \beta - \frac{m}{2} \log(2\pi) - \beta E_D(\mathbf{w})$$

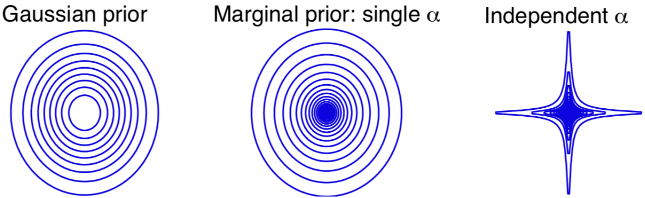
where $E_D(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{m} (y_i - \mathbf{w} \cdot \boldsymbol{\phi}(\mathbf{x}_i)^\top)^2$

Priors Compared

- Specifying independent hyperparameters α_i is the key to sparsity
- Example marginal priors $p(w_1, w_2)$ illustrated below



- Specifying independent hyperparameters α_i is the key to sparsity
- Example marginal priors $p(w_1, w_2)$ illustrated below



- The prior is

$$p(\mathbf{w}) = p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{i=1}^{M} \mathcal{N}(w_i|\mathbf{0}, \alpha_i^{-1})$$

- The posterior is defined by

$$p(\mathbf{w}|\mathcal{D}_m, \boldsymbol{\alpha}, \beta) = \mathcal{N}(\mathbf{w}|\boldsymbol{\omega}_m, \mathbf{S}_m),$$
$$\boldsymbol{\omega}_m = \beta \mathbf{S}_m \boldsymbol{\Phi}^\top \mathbf{Y}_m$$
$$\mathbf{S}_m = (\boldsymbol{\alpha} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1},$$

where $\boldsymbol{\Phi} = \{\phi_i(\mathbf{x}_j)\} \in \mathbb{R}^{m \times M}$, $\boldsymbol{\alpha} = \mathrm{diag}(\alpha_i)$

- We maximize the evidence approximation to estimate $\boldsymbol{\alpha}$ and β

$$p(\mathcal{D}_m|\boldsymbol{\alpha}, \beta) = \int p(\mathcal{D}_m|\mathbf{w}, \beta) p(\mathbf{w}|\boldsymbol{\alpha}) d\mathbf{w} =$$

- The prior is

$$p(\mathbf{w}) = p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{i=1}^{M} \mathcal{N}(w_i|\mathbf{0}, \alpha_i^{-1})$$

- The posterior is defined by

$$\begin{aligned} p(\mathbf{w}|\mathcal{D}_m, \boldsymbol{\alpha}, \beta) &= \mathcal{N}(\mathbf{w}|\boldsymbol{\omega}_m, \mathbf{S}_m), \\ \boldsymbol{\omega}_m &= \beta \mathbf{S}_m \boldsymbol{\Phi}^\top \mathbf{Y}_m \\ \mathbf{S}_m &= (\boldsymbol{\alpha} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1}, \end{aligned}$$

where $\boldsymbol{\Phi} = \{\phi_i(\mathbf{x}_j)\} \in \mathbb{R}^{m \times M}$, $\boldsymbol{\alpha} = \mathrm{diag}(\alpha_i)$

- We maximize the evidence approximation to estimate α and β

$$p(\mathcal{D}_m|\boldsymbol{\alpha}, \beta) = \int p(\mathcal{D}_m|\mathbf{w}, \beta) p(\mathbf{w}|\boldsymbol{\alpha}) d\mathbf{w} =$$

- The prior is

$$p(\mathbf{w}) = p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{i=1}^{M} \mathcal{N}(w_i|\mathbf{0}, \alpha_i^{-1})$$

- The posterior is defined by

$$\begin{aligned} p(\mathbf{w}|\mathcal{D}_m, \boldsymbol{\alpha}, \beta) &= \mathcal{N}(\mathbf{w}|\boldsymbol{\omega}_m, \mathbf{S}_m), \\ \boldsymbol{\omega}_m &= \beta \mathbf{S}_m \boldsymbol{\Phi}^\top \mathbf{Y}_m \\ \mathbf{S}_m &= (\boldsymbol{\alpha} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1}, \end{aligned}$$

where $\boldsymbol{\Phi} = \{\phi_i(\mathbf{x}_j)\} \in \mathbb{R}^{m \times M}$, $\boldsymbol{\alpha} = \mathrm{diag}(\alpha_i)$

- We maximize the evidence approximation to estimate $\boldsymbol{\alpha}$ and β

$$p(\mathcal{D}_m|\boldsymbol{\alpha}, \beta) = \int p(\mathcal{D}_m|\mathbf{w}, \beta) p(\mathbf{w}|\boldsymbol{\alpha}) d\mathbf{w} =$$

Parameter distribution

- The prior is

$$p(\mathbf{w}) = p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{i=1}^{M} \mathcal{N}(w_i|\mathbf{0}, \alpha_i^{-1})$$

- The posterior is defined by

$$\begin{aligned} p(\mathbf{w}|\mathcal{D}_m, \boldsymbol{\alpha}, \beta) &= \mathcal{N}(\mathbf{w}|\boldsymbol{\omega}_m, \mathbf{S}_m), \\ \boldsymbol{\omega}_m &= \beta \mathbf{S}_m \boldsymbol{\Phi}^\top \mathbf{Y}_m \\ \mathbf{S}_m &= (\boldsymbol{\alpha} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1}, \end{aligned}$$

  where $\boldsymbol{\Phi} = \{\phi_i(\mathbf{x}_j)\} \in \mathbb{R}^{m \times M}$, $\boldsymbol{\alpha} = \mathrm{diag}(\alpha_i)$
- We maximize the evidence approximation to estimate $\boldsymbol{\alpha}$ and β

$$\begin{aligned} p(\mathcal{D}_m|\boldsymbol{\alpha}, \beta) &= \int p(\mathcal{D}_m|\mathbf{w}, \beta) p(\mathbf{w}|\boldsymbol{\alpha}) d\mathbf{w} = \\ &= \int p(\mathbf{Y}_m|\mathbf{X}_m, \mathbf{w}, \beta) p(\mathbf{w}|\boldsymbol{\alpha}) d\mathbf{w} \end{aligned}$$

- In our case

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{i=1}^{M} \mathcal{N}(w_i|\mathbf{0}, \alpha_i^{-1})$$

$$p(\mathbf{Y}_m|\mathbf{X}_m, \mathbf{w}, \beta) = \mathcal{N}(\mathbf{Y}_m|\boldsymbol{\Phi} \cdot \mathbf{w}^{\top}, \beta^{-1}\mathbf{I})$$

- The log evidence has the form

- In our case

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{i=1}^{M} \mathcal{N}(w_i|\mathbf{0}, \alpha_i^{-1})$$

$$p(\mathbf{Y}_m|\mathbf{X}_m, \mathbf{w}, \beta) = \mathcal{N}(\mathbf{Y}_m|\boldsymbol{\Phi} \cdot \mathbf{w}^\top, \beta^{-1}\mathbf{I})$$

- The log evidence has the form

- In our case

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{i=1}^{M} \mathcal{N}(w_i|\mathbf{0}, \alpha_i^{-1})$$

$$p(\mathbf{Y}_m|\mathbf{X}_m, \mathbf{w}, \beta) = \mathcal{N}(\mathbf{Y}_m|\boldsymbol{\Phi} \cdot \mathbf{w}^\top, \beta^{-1}\mathbf{I})$$

- The log evidence has the form

- In our case

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{i=1}^{M} \mathcal{N}(w_i|\mathbf{0}, \alpha_i^{-1})$$

$$p(\mathbf{Y}_m|\mathbf{X}_m, \mathbf{w}, \beta) = \mathcal{N}(\mathbf{Y}_m|\boldsymbol{\Phi} \cdot \mathbf{w}^\top, \beta^{-1}\mathbf{I})$$

- The log evidence has the form

$$\log p(\mathbf{Y}_m|\mathbf{X}_m, \boldsymbol{\alpha}, \beta) = \log \int p(\mathbf{Y}_m|\mathbf{X}_m, \mathbf{w}, \beta) p(\mathbf{w}|\boldsymbol{\alpha}) d\mathbf{w}$$

- In our case

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{i=1}^{M} \mathcal{N}(w_i|\mathbf{0}, \alpha_i^{-1})$$

$$p(\mathbf{Y}_m|\mathbf{X}_m, \mathbf{w}, \beta) = \mathcal{N}(\mathbf{Y}_m|\boldsymbol{\Phi} \cdot \mathbf{w}^\top, \beta^{-1}\mathbf{I})$$

- The log evidence has the form

$$\begin{aligned}\log p(\mathbf{Y}_m|\mathbf{X}_m, \boldsymbol{\alpha}, \beta) &= \log \int p(\mathbf{Y}_m|\mathbf{X}_m, \mathbf{w}, \beta) p(\mathbf{w}|\boldsymbol{\alpha}) d\mathbf{w} \\ &= \log \mathcal{N}(\mathbf{Y}_m|\mathbf{0}, \mathbf{C})\end{aligned}$$

- In our case

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{i=1}^{M} \mathcal{N}(w_i|\mathbf{0}, \alpha_i^{-1})$$

$$p(\mathbf{Y}_m|\mathbf{X}_m, \mathbf{w}, \beta) = \mathcal{N}(\mathbf{Y}_m|\boldsymbol{\Phi} \cdot \mathbf{w}^\top, \beta^{-1}\mathbf{I})$$

- The log evidence has the form

$$\begin{aligned}\log p(\mathbf{Y}_m|\mathbf{X}_m, \boldsymbol{\alpha}, \beta) &= \log \int p(\mathbf{Y}_m|\mathbf{X}_m, \mathbf{w}, \beta) p(\mathbf{w}|\boldsymbol{\alpha}) d\mathbf{w} \\ &= \log \mathcal{N}(\mathbf{Y}_m|\mathbf{0}, \mathbf{C}) \\ &= -\frac{1}{2}\left\{m \log(2\pi) + \log|\mathbf{C}| + \mathbf{Y}_m^\top \mathbf{C}^{-1} \mathbf{Y}_m\right\},\end{aligned}$$

- In our case

$$p(\mathbf{w}|\boldsymbol{\alpha}) = \prod_{i=1}^{M} \mathcal{N}(w_i|\mathbf{0}, \alpha_i^{-1})$$

$$p(\mathbf{Y}_m|\mathbf{X}_m, \mathbf{w}, \beta) = \mathcal{N}(\mathbf{Y}_m|\boldsymbol{\Phi} \cdot \mathbf{w}^\top, \beta^{-1}\mathbf{I})$$

- The log evidence has the form

$$\begin{aligned}\log p(\mathbf{Y}_m|\mathbf{X}_m, \boldsymbol{\alpha}, \beta) &= \log \int p(\mathbf{Y}_m|\mathbf{X}_m, \mathbf{w}, \beta) p(\mathbf{w}|\boldsymbol{\alpha}) d\mathbf{w} \\ &= \log \mathcal{N}(\mathbf{Y}_m|\mathbf{0}, \mathbf{C}) \\ &= -\frac{1}{2}\left\{m \log(2\pi) + \log|\mathbf{C}| + \mathbf{Y}_m^\top \mathbf{C}^{-1} \mathbf{Y}_m\right\},\end{aligned}$$

where $\mathbf{Y}_m = (y_1, \dots, y_m)^\top$, and $\mathbf{C} \in \mathbb{R}^{m \times m}$ is

$$\mathbf{C} = \beta^{-1}\mathbf{I} + \boldsymbol{\Phi}\boldsymbol{\alpha}^{-1}\boldsymbol{\Phi}^\top$$

Re-estimation equations

- We set the derivatives of the log evidence w.r.t. $\boldsymbol{\alpha}$ and β to zero

$$\log p(\mathbf{Y}_m|\mathbf{X}_m, \boldsymbol{\alpha}, \beta) \sim -\frac{1}{2}\left\{\log|\mathbf{C}| + \mathbf{Y}_m^\top \mathbf{C}^{-1}\mathbf{Y}_m\right\} \to \max_{\boldsymbol{\alpha},\beta}$$

where $\mathbf{C} = \beta^{-1}\mathbf{I} + \boldsymbol{\Phi}\boldsymbol{\alpha}^{-1}\boldsymbol{\Phi}^\top$

- As is the case of isotropic prior on $\mathbf{w}$ we obtain similar re-estimation equations

$$\alpha_i^{new} = \frac{\gamma_i}{([\boldsymbol{\omega}_m]_i)^2},\ (\beta^{new})^{-1} = \frac{\|\mathbf{Y}_m - \boldsymbol{\Phi}\boldsymbol{\omega}_m^\top\|^2}{m - \sum_i \gamma_i}$$

- Here $[\boldsymbol{\omega}_m]_i$ is the i-th component of the posterior mean

$$\boldsymbol{\omega}_m = \beta \mathbf{S}_m \boldsymbol{\Phi}^\top \mathbf{Y}_m,$$
$$\mathbf{S}_m = (\boldsymbol{\alpha} + \beta\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}$$

- The parameter γ_i measures how well the corresponding parameter w_i is determined by the data and is defined by

$$\gamma_i = 1 - \alpha_i \cdot [\mathbf{S}_m]_{ii}$$

- We set the derivatives of the log evidence w.r.t. $\boldsymbol{\alpha}$ and β to zero

$$\log p(\mathbf{Y}_m|\mathbf{X}_m, \boldsymbol{\alpha}, \beta) \sim -\frac{1}{2}\left\{\log|\mathbf{C}| + \mathbf{Y}_m^\top \mathbf{C}^{-1}\mathbf{Y}_m\right\} \to \max_{\boldsymbol{\alpha},\beta}$$

  where $\mathbf{C} = \beta^{-1}\mathbf{I} + \boldsymbol{\Phi}\boldsymbol{\alpha}^{-1}\boldsymbol{\Phi}^\top$

- As is the case of isotropic prior on $\mathbf{w}$ we obtain similar re-estimation equations

$$\alpha_i^{new} = \frac{\gamma_i}{([\boldsymbol{\omega}_m]_i)^2},\ (\beta^{new})^{-1} = \frac{\|\mathbf{Y}_m - \boldsymbol{\Phi}\boldsymbol{\omega}_m^\top\|^2}{m - \sum_i \gamma_i}$$

- Here $[\boldsymbol{\omega}_m]_i$ is the i-th component of the posterior mean

$$\boldsymbol{\omega}_m = \beta \mathbf{S}_m \boldsymbol{\Phi}^\top \mathbf{Y}_m,$$
$$\mathbf{S}_m = (\boldsymbol{\alpha} + \beta\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}$$

- The parameter γ_i measures how well the corresponding parameter w_i is determined by the data and is defined by

$$\gamma_i = 1 - \alpha_i \cdot [\mathbf{S}_m]_{ii}$$

- We set the derivatives of the log evidence w.r.t. $\boldsymbol{\alpha}$ and β to zero

$$\log p(\mathbf{Y}_m|\mathbf{X}_m, \boldsymbol{\alpha}, \beta) \sim -\frac{1}{2}\left\{\log|\mathbf{C}| + \mathbf{Y}_m^\top \mathbf{C}^{-1}\mathbf{Y}_m\right\} \to \max_{\boldsymbol{\alpha},\beta}$$

where $\mathbf{C} = \beta^{-1}\mathbf{I} + \boldsymbol{\Phi}\boldsymbol{\alpha}^{-1}\boldsymbol{\Phi}^\top$

- As is the case of isotropic prior on $\mathbf{w}$ we obtain similar re-estimation equations

$$\alpha_i^{new} = \frac{\gamma_i}{([\boldsymbol{\omega}_m]_i)^2},\ (\beta^{new})^{-1} = \frac{\|\mathbf{Y}_m - \boldsymbol{\Phi}\boldsymbol{\omega}_m^\top\|^2}{m - \sum_i \gamma_i}$$

- Here $[\boldsymbol{\omega}_m]_i$ is the i-th component of the posterior mean

$$\boldsymbol{\omega}_m = \beta \mathbf{S}_m \boldsymbol{\Phi}^\top \mathbf{Y}_m,$$
$$\mathbf{S}_m = (\boldsymbol{\alpha} + \beta\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}$$

- The parameter γ_i measures how well the corresponding parameter w_i is determined by the data and is defined by

$$\gamma_i = 1 - \alpha_i \cdot [\mathbf{S}_m]_{ii}$$

- We set the derivatives of the log evidence w.r.t. $\boldsymbol{\alpha}$ and β to zero

$$\log p(\mathbf{Y}_m|\mathbf{X}_m, \boldsymbol{\alpha}, \beta) \sim -\frac{1}{2}\left\{\log|\mathbf{C}| + \mathbf{Y}_m^\top \mathbf{C}^{-1}\mathbf{Y}_m\right\} \to \max_{\boldsymbol{\alpha},\beta}$$

where $\mathbf{C} = \beta^{-1}\mathbf{I} + \boldsymbol{\Phi}\boldsymbol{\alpha}^{-1}\boldsymbol{\Phi}^\top$

- As is the case of isotropic prior on $\mathbf{w}$ we obtain similar re-estimation equations

$$\alpha_i^{new} = \frac{\gamma_i}{([\boldsymbol{\omega}_m]_i)^2},\ (\beta^{new})^{-1} = \frac{\|\mathbf{Y}_m - \boldsymbol{\Phi}\boldsymbol{\omega}_m^\top\|^2}{m - \sum_i \gamma_i}$$

- Here $[\boldsymbol{\omega}_m]_i$ is the i-th component of the posterior mean

$$\boldsymbol{\omega}_m = \beta \mathbf{S}_m \boldsymbol{\Phi}^\top \mathbf{Y}_m,$$
$$\mathbf{S}_m = (\boldsymbol{\alpha} + \beta\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}$$

- The parameter γ_i measures how well the corresponding parameter w_i is determined by the data and is defined by

$$\gamma_i = 1 - \alpha_i \cdot [\mathbf{S}_m]_{ii}$$

- We get that

$$\mathbf{C} = \beta^{-1}\mathbf{I} + \boldsymbol{\Phi}\boldsymbol{\alpha}^{-1}\boldsymbol{\Phi}^\top = \beta^{-1}\mathbf{I} + \sum_{j=1}^{M} \alpha_j^{-1}\boldsymbol{\phi}_j\boldsymbol{\phi}_j^\top,$$

where the column vector $\boldsymbol{\phi}_i = (\phi_i(\mathbf{x}_1), \ldots, \phi_i(\mathbf{x}_m))$

- Let us pull out the contribution from α_i in $\mathbf{C}$ and get

$$\mathbf{C} = \beta^{-1}\mathbf{I} + \sum_{j\neq i} \alpha_j^{-1}\boldsymbol{\phi}_j\boldsymbol{\phi}_j^\top + \alpha_i^{-1}\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top$$
$$= \mathbf{C}_{-i} + \alpha_i^{-1}\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top$$

- Due to Woodbury identity

$$(\mathbf{A} + \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{D} + \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1}$$

- If $\mathbf{a}$ and $\mathbf{b}$ are N-dimensional column vectors, then

$$|\mathbf{I}_N + \mathbf{a}\mathbf{b}^\top| = 1 + \mathbf{a}^\top\mathbf{b}$$

- Since $\mathbf{C} = \mathbf{C}_{-i} + \alpha_i^{-1}\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top = \mathbf{C}_{-i}(\mathbf{I} + \alpha_i^{-1}\mathbf{C}_{-i}^{-1}\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top)$ we get that

Analysis of sparsity

- We get that

$$\mathbf{C} = \beta^{-1}\mathbf{I} + \boldsymbol{\Phi}\boldsymbol{\alpha}^{-1}\boldsymbol{\Phi}^\top = \beta^{-1}\mathbf{I} + \sum_{j=1}^{M} \alpha_j^{-1}\boldsymbol{\phi}_j\boldsymbol{\phi}_j^\top,$$

  where the column vector $\boldsymbol{\phi}_i = (\phi_i(\mathbf{x}_1), \dots, \phi_i(\mathbf{x}_m))$
- Let us pull out the contribution from α_i in $\mathbf{C}$ and get

$$\begin{aligned}\mathbf{C} &= \beta^{-1}\mathbf{I} + \sum_{j\neq i} \alpha_j^{-1}\boldsymbol{\phi}_j\boldsymbol{\phi}_j^\top + \alpha_i^{-1}\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top \\ &= \mathbf{C}_{-i} + \alpha_i^{-1}\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top\end{aligned}$$

- Due to Woodbury identity

$$(\mathbf{A} + \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{D} + \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1}$$

- If $\mathbf{a}$ and $\mathbf{b}$ are N-dimensional column vectors, then

$$|\mathbf{I}_N + \mathbf{a}\mathbf{b}^\top| = 1 + \mathbf{a}^\top\mathbf{b}$$

- Since $\mathbf{C} = \mathbf{C}_{-i} + \alpha_i^{-1}\phi_i\phi_i^\top = \mathbf{C}_{-i}(\mathbf{I} + \alpha_i^{-1}\mathbf{C}_{-i}^{-1}\phi_i\phi_i^\top)$ we get that

- We get that

$$\mathbf{C} = \beta^{-1}\mathbf{I} + \boldsymbol{\Phi}\boldsymbol{\alpha}^{-1}\boldsymbol{\Phi}^\top = \beta^{-1}\mathbf{I} + \sum_{j=1}^{M} \alpha_j^{-1}\boldsymbol{\phi}_j\boldsymbol{\phi}_j^\top,$$

where the column vector $\boldsymbol{\phi}_i = (\phi_i(\mathbf{x}_1), \ldots, \phi_i(\mathbf{x}_m))$

- Let us pull out the contribution from α_i in $\mathbf{C}$ and get

$$\begin{aligned}\mathbf{C} &= \beta^{-1}\mathbf{I} + \sum_{j\neq i} \alpha_j^{-1}\boldsymbol{\phi}_j\boldsymbol{\phi}_j^\top + \alpha_i^{-1}\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top \\ &= \mathbf{C}_{-i} + \alpha_i^{-1}\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top\end{aligned}$$

- Due to Woodbury identity

$$(\mathbf{A} + \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{D} + \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1}$$

- If $\mathbf{a}$ and $\mathbf{b}$ are N-dimensional column vectors, then

$$|\mathbf{I}_N + \mathbf{a}\mathbf{b}^\top| = 1 + \mathbf{a}^\top\mathbf{b}$$

- Since $\mathbf{C} = \mathbf{C}_{-i} + \alpha_i^{-1}\phi_i\phi_i^\top = \mathbf{C}_{-i}(\mathbf{I} + \alpha_i^{-1}\mathbf{C}_{-i}^{-1}\phi_i\phi_i^\top)$ we get that

- We get that

$$\mathbf{C} = \beta^{-1}\mathbf{I} + \boldsymbol{\Phi}\boldsymbol{\alpha}^{-1}\boldsymbol{\Phi}^\top = \beta^{-1}\mathbf{I} + \sum_{j=1}^{M} \alpha_j^{-1}\boldsymbol{\phi}_j\boldsymbol{\phi}_j^\top,$$

  where the column vector $\boldsymbol{\phi}_i = (\phi_i(\mathbf{x}_1), \ldots, \phi_i(\mathbf{x}_m))$
- Let us pull out the contribution from α_i in $\mathbf{C}$ and get

$$\begin{aligned}\mathbf{C} &= \beta^{-1}\mathbf{I} + \sum_{j \neq i} \alpha_j^{-1}\boldsymbol{\phi}_j\boldsymbol{\phi}_j^\top + \alpha_i^{-1}\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top \\ &= \mathbf{C}_{-i} + \alpha_i^{-1}\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top\end{aligned}$$

- Due to Woodbury identity

$$(\mathbf{A} + \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{D} + \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1}$$

- If $\mathbf{a}$ and $\mathbf{b}$ are N-dimensional column vectors, then

$$|\mathbf{I}_N + \mathbf{a}\mathbf{b}^\top| = 1 + \mathbf{a}^\top\mathbf{b}$$

- Since $\mathbf{C} = \mathbf{C}_{-i} + \alpha_i^{-1}\phi_i\phi_i^\top = \mathbf{C}_{-i}(\mathbf{I} + \alpha_i^{-1}\mathbf{C}_{-i}^{-1}\phi_i\phi_i^\top)$ we get that

- We get that

$$\mathbf{C} = \beta^{-1}\mathbf{I} + \boldsymbol{\Phi}\boldsymbol{\alpha}^{-1}\boldsymbol{\Phi}^\top = \beta^{-1}\mathbf{I} + \sum_{j=1}^{M} \alpha_j^{-1}\boldsymbol{\phi}_j\boldsymbol{\phi}_j^\top,$$

  where the column vector $\boldsymbol{\phi}_i = (\phi_i(\mathbf{x}_1), \ldots, \phi_i(\mathbf{x}_m))$
- Let us pull out the contribution from α_i in $\mathbf{C}$ and get

$$\begin{aligned}\mathbf{C} &= \beta^{-1}\mathbf{I} + \sum_{j \neq i} \alpha_j^{-1}\boldsymbol{\phi}_j\boldsymbol{\phi}_j^\top + \alpha_i^{-1}\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top \\ &= \mathbf{C}_{-i} + \alpha_i^{-1}\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top\end{aligned}$$

- Due to Woodbury identity

$$(\mathbf{A} + \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{D} + \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1}$$

- If $\mathbf{a}$ and $\mathbf{b}$ are N-dimensional column vectors, then

$$|\mathbf{I}_N + \mathbf{a}\mathbf{b}^\top| = 1 + \mathbf{a}^\top\mathbf{b}$$

- Since $\mathbf{C} = \mathbf{C}_{-i} + \alpha_i^{-1}\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top = \mathbf{C}_{-i}(\mathbf{I} + \alpha_i^{-1}\mathbf{C}_{-i}^{-1}\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top)$ we get that

- We get that
$$\mathbf{C} = \beta^{-1}\mathbf{I} + \boldsymbol{\Phi}\boldsymbol{\alpha}^{-1}\boldsymbol{\Phi}^\top = \beta^{-1}\mathbf{I} + \sum_{j=1}^{M} \alpha_j^{-1}\boldsymbol{\phi}_j\boldsymbol{\phi}_j^\top,$$
where the column vector $\boldsymbol{\phi}_i = (\phi_i(\mathbf{x}_1), \ldots, \phi_i(\mathbf{x}_m))$
- Let us pull out the contribution from α_i in $\mathbf{C}$ and get
$$\begin{aligned}\mathbf{C} &= \beta^{-1}\mathbf{I} + \sum_{j\neq i} \alpha_j^{-1}\boldsymbol{\phi}_j\boldsymbol{\phi}_j^\top + \alpha_i^{-1}\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top \\ &= \mathbf{C}_{-i} + \alpha_i^{-1}\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top\end{aligned}$$
- Due to Woodbury identity
$$(\mathbf{A} + \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{D} + \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1}$$
- If $\mathbf{a}$ and $\mathbf{b}$ are N-dimensional column vectors, then
$$|\mathbf{I}_N + \mathbf{a}\mathbf{b}^\top| = 1 + \mathbf{a}^\top\mathbf{b}$$
- Since $\mathbf{C} = \mathbf{C}_{-i} + \alpha_i^{-1}\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top = \mathbf{C}_{-i}(\mathbf{I} + \alpha_i^{-1}\mathbf{C}_{-i}^{-1}\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top)$ we get that
$$|\mathbf{C}| = |\mathbf{C}_{-i}|(1 + \alpha_i^{-1}\boldsymbol{\phi}_i^\top\mathbf{C}_{-i}^{-1}\boldsymbol{\phi}_i),$$

- We get that

$$\mathbf{C} = \beta^{-1}\mathbf{I} + \boldsymbol{\Phi}\boldsymbol{\alpha}^{-1}\boldsymbol{\Phi}^\top = \beta^{-1}\mathbf{I} + \sum_{j=1}^{M} \alpha_j^{-1}\boldsymbol{\phi}_j\boldsymbol{\phi}_j^\top,$$

  where the column vector $\boldsymbol{\phi}_i = (\phi_i(\mathbf{x}_1), \ldots, \phi_i(\mathbf{x}_m))$
- Let us pull out the contribution from α_i in $\mathbf{C}$ and get

$$\begin{aligned}\mathbf{C} &= \beta^{-1}\mathbf{I} + \sum_{j \neq i} \alpha_j^{-1}\boldsymbol{\phi}_j\boldsymbol{\phi}_j^\top + \alpha_i^{-1}\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top \\ &= \mathbf{C}_{-i} + \alpha_i^{-1}\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top\end{aligned}$$

- Due to Woodbury identity

$$(\mathbf{A} + \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{D} + \mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1}$$

- If $\mathbf{a}$ and $\mathbf{b}$ are N-dimensional column vectors, then

$$|\mathbf{I}_N + \mathbf{a}\mathbf{b}^\top| = 1 + \mathbf{a}^\top\mathbf{b}$$

- Since $\mathbf{C} = \mathbf{C}_{-i} + \alpha_i^{-1}\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top = \mathbf{C}_{-i}(\mathbf{I} + \alpha_i^{-1}\mathbf{C}_{-i}^{-1}\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top)$ we get that

$$|\mathbf{C}| = |\mathbf{C}_{-i}|(1 + \alpha_i^{-1}\boldsymbol{\phi}_i^\top\mathbf{C}_{-i}^{-1}\boldsymbol{\phi}_i), \ \mathbf{C}^{-1} = \mathbf{C}_{-i}^{-1} - \frac{\mathbf{C}_{-i}^{-1}\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top\mathbf{C}_{-i}^{-1}}{\alpha_i + \boldsymbol{\phi}_i^\top\mathbf{C}_{-i}^{-1}\boldsymbol{\phi}_i}$$

Analysis of sparsity

- The log evidence function

$$L(\boldsymbol{\alpha}) = \log p(\mathbf{Y}_m|\mathbf{X}_m, \boldsymbol{\alpha}, \beta) \sim -\frac{1}{2}\left\{\log|\mathbf{C}| + \mathbf{Y}_m^\top \mathbf{C}^{-1}\mathbf{Y}_m\right\} \to \max_{\boldsymbol{\alpha},\beta}$$

- Thanks to pulling out the contribution from α_i in $\mathbf{C}$ we get that

Analysis of sparsity

- The log evidence function

$$L(\boldsymbol{\alpha}) = \log p(\mathbf{Y}_m|\mathbf{X}_m, \boldsymbol{\alpha}, \beta) \sim -\frac{1}{2}\left\{\log|\mathbf{C}| + \mathbf{Y}_m^\top \mathbf{C}^{-1}\mathbf{Y}_m\right\} \to \max_{\boldsymbol{\alpha},\beta}$$

- Thanks to pulling out the contribution from α_i in $\mathbf{C}$ we get that

- The log evidence function

$$L(\boldsymbol{\alpha}) = \log p(\mathbf{Y}_m | \mathbf{X}_m, \boldsymbol{\alpha}, \beta) \sim -\frac{1}{2} \left\{ \log |\mathbf{C}| + \mathbf{Y}_m^\top \mathbf{C}^{-1} \mathbf{Y}_m \right\} \to \max_{\boldsymbol{\alpha}, \beta}$$

- Thanks to pulling out the contribution from α_i in $\mathbf{C}$ we get that

$$L(\boldsymbol{\alpha}) = L(\boldsymbol{\alpha}_{-i}) + \lambda(\alpha_i),$$

- The log evidence function

$$L(\boldsymbol{\alpha}) = \log p(\mathbf{Y}_m | \mathbf{X}_m, \boldsymbol{\alpha}, \beta) \sim -\frac{1}{2} \left\{ \log |\mathbf{C}| + \mathbf{Y}_m^\top \mathbf{C}^{-1} \mathbf{Y}_m \right\} \to \max_{\boldsymbol{\alpha}, \beta}$$

- Thanks to pulling out the contribution from α_i in $\mathbf{C}$ we get that

$$L(\boldsymbol{\alpha}) = L(\boldsymbol{\alpha}_{-i}) + \lambda(\alpha_i),$$

where is $L(\boldsymbol{\alpha}_{-i})$ is the same function but for the linear model without basis function $\boldsymbol{\phi}_i$, and

$$\lambda(\alpha_i) = \frac{1}{2} \left[\log \alpha_i - \log(\alpha_i + s_i) + \frac{q_i^2}{\alpha_i + s_i} \right],$$

- The log evidence function

$$L(\boldsymbol{\alpha}) = \log p(\mathbf{Y}_m|\mathbf{X}_m, \boldsymbol{\alpha}, \beta) \sim -\frac{1}{2}\left\{\log|\mathbf{C}| + \mathbf{Y}_m^\top \mathbf{C}^{-1}\mathbf{Y}_m\right\} \to \max_{\boldsymbol{\alpha},\beta}$$

- Thanks to pulling out the contribution from α_i in $\mathbf{C}$ we get that

$$L(\boldsymbol{\alpha}) = L(\boldsymbol{\alpha}_{-i}) + \lambda(\alpha_i),$$

where is $L(\boldsymbol{\alpha}_{-i})$ is the same function but for the linear model without basis function $\boldsymbol{\phi}_i$, and

$$\lambda(\alpha_i) = \frac{1}{2}\left[\log\alpha_i - \log(\alpha_i + s_i) + \frac{q_i^2}{\alpha_i + s_i}\right],$$

$$s_i = \boldsymbol{\phi}_i^\top \mathbf{C}_{-i}^{-1}\boldsymbol{\phi}_i$$

$$q_i = \boldsymbol{\phi}_i^\top \mathbf{C}_{-i}^{-1}\mathbf{Y}_m$$

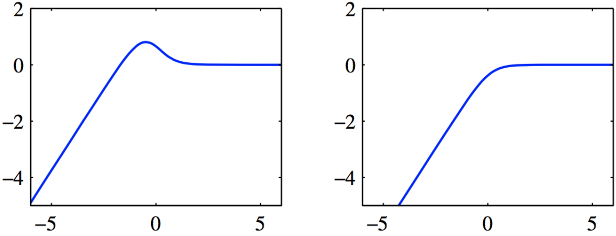


Figure – Plots of the log marginal likelihood $\lambda(\alpha_i)$ versus $\log \alpha_i$ showing on the left, the single maximum at a finite α_i for $q_i^2 > s_i$ and on the right, the maximum at $\alpha_i = \infty$ for $q_i^2 < s_i$

- The stationary points of the marginal likelihood w.r.t. α_i

$$\frac{d\lambda(\alpha_i)}{d\alpha_i} = \frac{\alpha_i^{-1} s_i^2 - (q_i^2 - s_i)}{2(\alpha_i + s_i)^2} = 0$$

- Recall that $\alpha_i \geq 0$, then if you $q_i^2 < s_i$ we get that $\alpha_i^{opt} \to \infty$
- If $q_i^2 > s_i$, then

$$\alpha_i^{opt} = \frac{s_i^2}{q_i^2 - s_i}$$

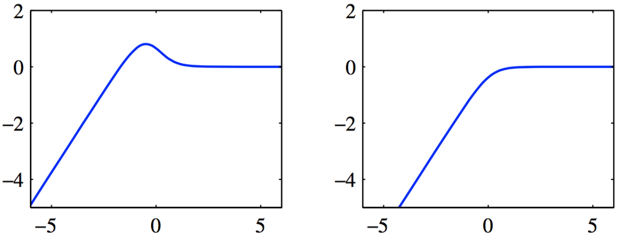


Figure – Plots of the log marginal likelihood $\lambda(\alpha_i)$ versus $\log \alpha_i$ showing on the left, the single maximum at a finite α_i for $q_i^2 > s_i$ and on the right, the maximum at $\alpha_i = \infty$ for $q_i^2 < s_i$

- The stationary points of the marginal likelihood w.r.t. α_i

$$\frac{d\lambda(\alpha_i)}{d\alpha_i} = \frac{\alpha_i^{-1} s_i^2 - (q_i^2 - s_i)}{2(\alpha_i + s_i)^2} = 0$$

- Recall that $\alpha_i \geq 0$, then if you $q_i^2 < s_i$ we get that $\alpha_i^{opt} \to \infty$
- If $q_i^2 > s_i$, then

$$\alpha_i^{opt} = \frac{s_i^2}{q_i^2 - s_i}$$

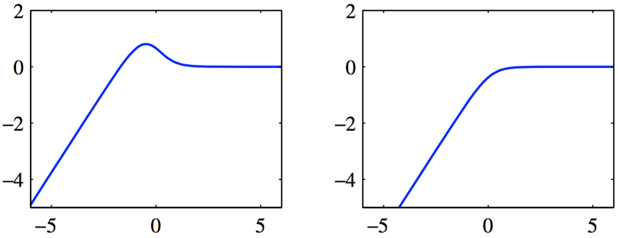


Figure – Plots of the log marginal likelihood $\lambda(\alpha_i)$ versus $\log \alpha_i$ showing on the left, the single maximum at a finite α_i for $q_i^2 > s_i$ and on the right, the maximum at $\alpha_i = \infty$ for $q_i^2 < s_i$

- The stationary points of the marginal likelihood w.r.t. α_i

$$\frac{d\lambda(\alpha_i)}{d\alpha_i} = \frac{\alpha_i^{-1} s_i^2 - (q_i^2 - s_i)}{2(\alpha_i + s_i)^2} = 0$$

- Recall that $\alpha_i \geq 0$, then if you $q_i^2 < s_i$ we get that $\alpha_i^{opt} \to \infty$
- If $q_i^2 > s_i$, then

$$\alpha_i^{opt} = \frac{s_i^2}{q_i^2 - s_i}$$

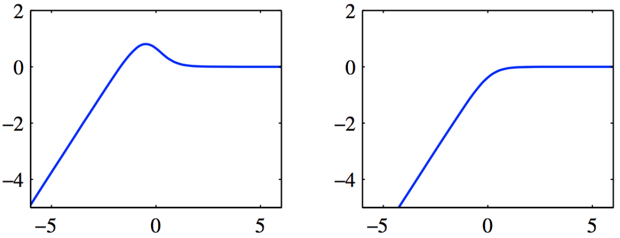


Figure – Plots of the log marginal likelihood $\lambda(\alpha_i)$ versus $\log \alpha_i$ showing on the left, the single maximum at a finite α_i for $q_i^2 > s_i$ and on the right, the maximum at $\alpha_i = \infty$ for $q_i^2 < s_i$

- The stationary points of the marginal likelihood w.r.t. α_i

$$\frac{d\lambda(\alpha_i)}{d\alpha_i} = \frac{\alpha_i^{-1} s_i^2 - (q_i^2 - s_i)}{2(\alpha_i + s_i)^2} = 0$$

- Recall that $\alpha_i \geq 0$, then if you $q_i^2 < s_i$ we get that $\alpha_i^{opt} \to \infty$
- If $q_i^2 > s_i$, then

$$\alpha_i^{opt} = \frac{s_i^2}{q_i^2 - s_i}$$

- For any given basis function $\phi_i(\mathbf{x})$ and associated hyperparameter α_i we can compute the quantities s_i and q_i^2 (true even if $\alpha_i = \infty$)
- Depending on the criterion $q_i^2 > s_i$ and the value of α_i we can then perform the following updates, all of which will increase $p(\mathcal{D}_m|\boldsymbol{\alpha}, \beta)$:

- For any given basis function $\phi_i(\mathbf{x})$ and associated hyperparameter α_i we can compute the quantities s_i and q_i^2 (true even if $\alpha_i = \infty$)
- Depending on the criterion $q_i^2 > s_i$ and the value of α_i we can then perform the following updates, all of which will increase $p(\mathcal{D}_m|\boldsymbol{\alpha}, \beta)$:

- For any given basis function $\phi_i(\mathbf{x})$ and associated hyperparameter α_i we can compute the quantities s_i and q_i^2 (true even if $\alpha_i = \infty$)
- Depending on the criterion $q_i^2 > s_i$ and the value of α_i we can then perform the following updates, all of which will increase $p(\mathcal{D}_m|\boldsymbol{\alpha}, \beta)$:

	"In model": $\alpha_i < \infty$	**"Out of model"**: $\alpha_i = \infty$
$q_i^2 > s_i$	*re-estimation* of α_i	*addition* of $\phi_i(\mathbf{x})$
$q_i^2 \leq s_i$	*deletion* of $\phi_i(\mathbf{x})$	—

1. Initialize β and all $\alpha_j = \infty$, i.e. the empty model
2. Select a function $\phi_i(\mathbf{x})$ from the set of all M functions
3. Compute relevance $R_i = q_i^2 - s_i$
   - if $R_i > 0$ and $\alpha_i < \infty$: re-estimate α_i
   - if $R_i > 0$ and $\alpha_i = \infty$: add ϕ_i to the model with updated α_i
   - if $R_i \leq 0$ and $\alpha_i < \infty$: delete ϕ_i from the model and set $\alpha_i = \infty$
4. If solving a regression problem update β
5. Recalculate all q_i and s_i
6. If convergence terminate, otherwise repeat

1. Initialize β and all $\alpha_j = \infty$, i.e. the empty model
2. Select a function $\phi_i(\mathbf{x})$ from the set of all M functions
3. Compute relevance $R_i = q_i^2 - s_i$
   - if $R_i > 0$ and $\alpha_i < \infty$: re-estimate α_i
   - if $R_i > 0$ and $\alpha_i = \infty$: add ϕ_i to the model with updated α_i
   - if $R_i \leq 0$ and $\alpha_i < \infty$: delete ϕ_i from the model and set $\alpha_i = \infty$
4. If solving a regression problem update β
5. Recalculate all q_i and s_i
6. If convergence terminate, otherwise repeat

1. Initialize β and all $\alpha_j = \infty$, i.e. the empty model
2. Select a function $\phi_i(\mathbf{x})$ from the set of all M functions
3. Compute relevance $R_i = q_i^2 - s_i$
   - if $R_i > 0$ and $\alpha_i < \infty$: re-estimate α_i
   - if $R_i > 0$ and $\alpha_i = \infty$: add ϕ_i to the model with updated α_i
   - if $R_i \leq 0$ and $\alpha_i < \infty$: delete ϕ_i from the model and set $\alpha_i = \infty$
4. If solving a regression problem update β
5. Recalculate all q_i and s_i
6. If convergence terminate, otherwise repeat

1. Initialize β and all $\alpha_j = \infty$, i.e. the empty model
2. Select a function $\phi_i(\mathbf{x})$ from the set of all M functions
3. Compute relevance $R_i = q_i^2 - s_i$
   - if $R_i > 0$ and $\alpha_i < \infty$: re-estimate α_i
   - if $R_i > 0$ and $\alpha_i = \infty$: add ϕ_i to the model with updated α_i
   - if $R_i \leq 0$ and $\alpha_i < \infty$: delete ϕ_i from the model and set $\alpha_i = \infty$
4. If solving a regression problem update β
5. Recalculate all q_i and s_i
6. If convergence terminate, otherwise repeat

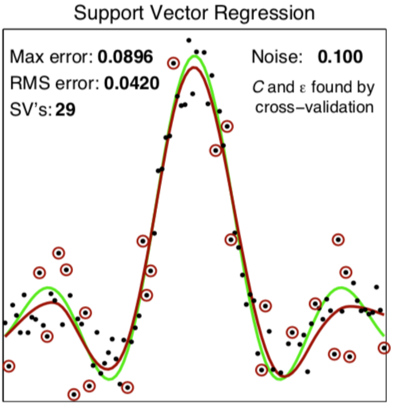
1. Initialize β and all $\alpha_j = \infty$, i.e. the empty model
2. Select a function $\phi_i(\mathbf{x})$ from the set of all M functions
3. Compute relevance $R_i = q_i^2 - s_i$
   - if $R_i > 0$ and $\alpha_i < \infty$: re-estimate α_i
   - if $R_i > 0$ and $\alpha_i = \infty$: add ϕ_i to the model with updated α_i
   - if $R_i \leq 0$ and $\alpha_i < \infty$: delete ϕ_i from the model and set $\alpha_i = \infty$
4. If solving a regression problem update β
5. Recalculate all q_i and s_i
6. If convergence terminate, otherwise repeat

1. Initialize β and all $\alpha_j = \infty$, i.e. the empty model
2. Select a function $\phi_i(\mathbf{x})$ from the set of all M functions
3. Compute relevance $R_i = q_i^2 - s_i$
   - if $R_i > 0$ and $\alpha_i < \infty$: re-estimate α_i
   - if $R_i > 0$ and $\alpha_i = \infty$: add ϕ_i to the model with updated α_i
   - if $R_i \leq 0$ and $\alpha_i < \infty$: delete ϕ_i from the model and set $\alpha_i = \infty$
4. If solving a regression problem update β
5. Recalculate all q_i and s_i
6. If convergence terminate, otherwise repeat

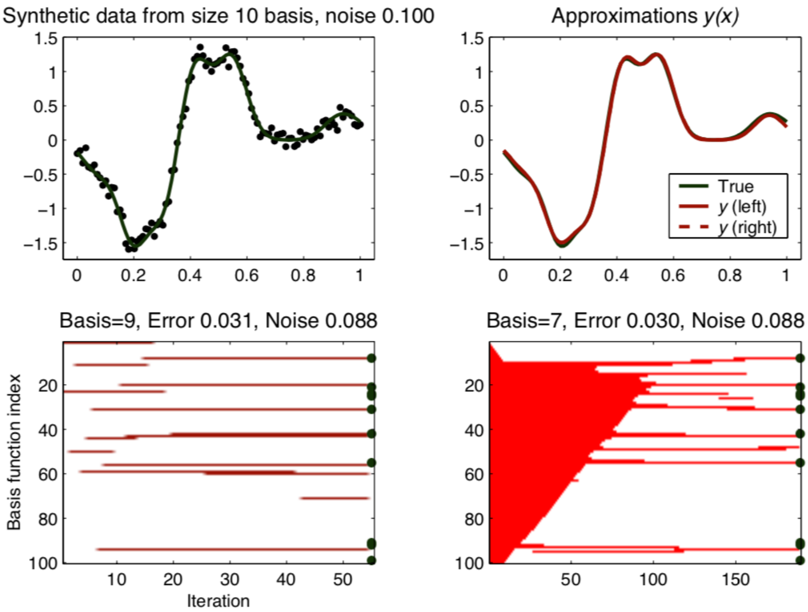
1. Initialize β and all $\alpha_j = \infty$, i.e. the empty model
2. Select a function $\phi_i(\mathbf{x})$ from the set of all M functions
3. Compute relevance $R_i = q_i^2 - s_i$
   - if $R_i > 0$ and $\alpha_i < \infty$: re-estimate α_i
   - if $R_i > 0$ and $\alpha_i = \infty$: add ϕ_i to the model with updated α_i
   - if $R_i \leq 0$ and $\alpha_i < \infty$: delete ϕ_i from the model and set $\alpha_i = \infty$
4. If solving a regression problem update β
5. Recalculate all q_i and s_i
6. If convergence terminate, otherwise repeat

1. Initialize β and all $\alpha_j = \infty$, i.e. the empty model
2. Select a function $\phi_i(\mathbf{x})$ from the set of all M functions
3. Compute relevance $R_i = q_i^2 - s_i$
   - if $R_i > 0$ and $\alpha_i < \infty$: re-estimate α_i
   - if $R_i > 0$ and $\alpha_i = \infty$: add ϕ_i to the model with updated α_i
   - if $R_i \leq 0$ and $\alpha_i < \infty$: delete ϕ_i from the model and set $\alpha_i = \infty$
4. If solving a regression problem update β
5. Recalculate all q_i and s_i
6. If convergence terminate, otherwise repeat

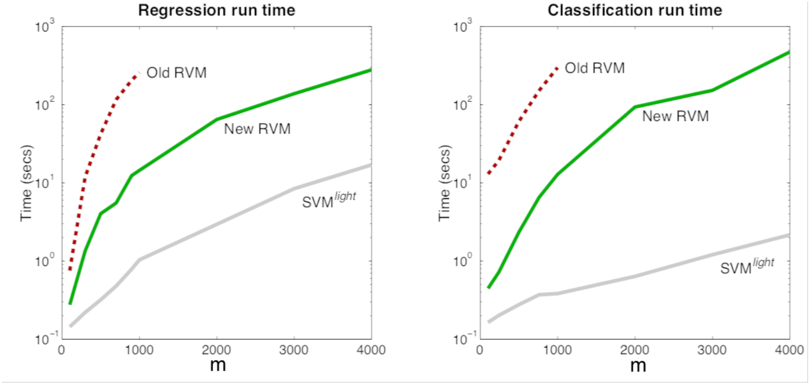
1. Initialize β and all $\alpha_j = \infty$, i.e. the empty model
2. Select a function $\phi_i(\mathbf{x})$ from the set of all M functions
3. Compute relevance $R_i = q_i^2 - s_i$
   - if $R_i > 0$ and $\alpha_i < \infty$: re-estimate α_i
   - if $R_i > 0$ and $\alpha_i = \infty$: add ϕ_i to the model with updated α_i
   - if $R_i \leq 0$ and $\alpha_i < \infty$: delete ϕ_i from the model and set $\alpha_i = \infty$
4. If solving a regression problem update β
5. Recalculate all q_i and s_i
6. If convergence terminate, otherwise repeat


Support Vector Regression
Max error: 0.0896
RMS error: 0.0420
SV's: 29
Noise: 0.100
C and ε found by
cross–validation
Relevance Vector Regression
Max error: 0.0664
RMS error: 0.0322
RV's: 7
Noise: 0.100
Estimate: 0.107

Kernel regression



Data set	*N*	*d*	*errors* SVM	RVM	*vectors* SVM	RVM
Sinc (Gaussian noise)	100	1	0.378	0.326	45.2	6.7
Sinc (Uniform noise)	100	1	0.215	0.187	44.3	7.0
Friedman #1	240	10	2.92	2.80	116.6	59.4
Friedman #2	240	4	4140	3505	110.3	6.9
Friedman #3	240	4	0.0202	0.0164	106.5	11.5
Boston Housing	481	13	8.04	7.46	142.8	39.0
Normalised Mean			1.00	**0.86**	1.00	**0.15**



Here we either

- Optimize greedily and delete basis functions (New RVM)
- Optimize w.r.t. all existing basis functions and delete some of them only at the end of the learning process (Old RVM)

- Comparing at $m = 1000$ we have

	Regression	Classification
Old RVM	4 mins 17 secs	4 mins 58 secs
New RVM	14.42 secs	12.84 secs
SVM^{light}	1.03 secs	0.38 secs

- In practice usually it takes $20 - 50$ iterations to learn RVM
- On each iteration we calculate $\boldsymbol{\omega}_m$ (inversion of a matrix of size $M \times M$ is required), and re-calculate $\boldsymbol{\alpha}$ and β (usually $O(1)$). As a result RVM is $20 - 50$ times slower than ordinary linear regression
- If we use kernel functions $K(\mathbf{x}, \mathbf{x}_i)$ as basis functions $\phi_i(\mathbf{x})$, then we have to perform cross-validation to select the kernel width. As a result we get a sparse kernel regression, as only a small subset of the initial sample will be used to define kernel basis functions and the final decision rule

- In practice usually it takes $20 - 50$ iterations to learn RVM
- On each iteration we calculate $\boldsymbol{\omega}_m$ (inversion of a matrix of size $M \times M$ is required), and re-calculate $\boldsymbol{\alpha}$ and β (usually $O(1)$). As a result RVM is $20 - 50$ times slower than ordinary linear regression
- If we use kernel functions $K(\mathbf{x}, \mathbf{x}_i)$ as basis functions $\phi_i(\mathbf{x})$, then we have to perform cross-validation to select the kernel width. As a result we get a sparse kernel regression, as only a small subset of the initial sample will be used to define kernel basis functions and the final decision rule

- In practice usually it takes $20-50$ iterations to learn RVM
- On each iteration we calculate $\boldsymbol{\omega}_m$ (inversion of a matrix of size $M \times M$ is required), and re-calculate $\boldsymbol{\alpha}$ and β (usually $O(1)$). As a result RVM is $20-50$ times slower than ordinary linear regression
- If we use kernel functions $K(\mathbf{x}, \mathbf{x}_i)$ as basis functions $\phi_i(\mathbf{x})$, then we have to perform cross-validation to select the kernel width. As a result we get a sparse kernel regression, as only a small subset of the initial sample will be used to define kernel basis functions and the final decision rule

- With RVM we can obtain not only point prediction, but also its uncertainty. Let $\boldsymbol{\alpha}^*$ and β^* be the hyperparameters that maximize the marginal likelihood, then the predictive distribution

- RVM concept can be used not only for regression, but also for classification, kernel density estimation, sparse kernel PCA, robust regression, etc.

- With RVM we can obtain not only point prediction, but also its uncertainty. Let $\boldsymbol{\alpha}^*$ and β^* be the hyperparameters that maximize the marginal likelihood, then the predictive distribution

$$p(y|\mathbf{x}, \mathbf{X}_m, \mathbf{Y}_m, \boldsymbol{\alpha}^*, \beta^*) = \int p(y|\mathbf{x}, \mathbf{w}, \beta^*) p(\mathbf{w}|\mathbf{X}_m, \mathbf{Y}_m, \boldsymbol{\alpha}^*, \beta^*) d\mathbf{w}$$

- RVM concept can be used not only for regression, but also for classification, kernel density estimation, sparse kernel PCA, robust regression, etc.

- With RVM we can obtain not only point prediction, but also its uncertainty. Let $\boldsymbol{\alpha}^*$ and β^* be the hyperparameters that maximize the marginal likelihood, then the predictive distribution

$$\begin{aligned} p(y|\mathbf{x}, \mathbf{X}_m, \mathbf{Y}_m, \boldsymbol{\alpha}^*, \beta^*) &= \int p(y|\mathbf{x}, \mathbf{w}, \beta^*) p(\mathbf{w}|\mathbf{X}_m, \mathbf{Y}_m, \boldsymbol{\alpha}^*, \beta^*) d\mathbf{w} \\ &= \mathcal{N}(y|\boldsymbol{\omega}_m \cdot \boldsymbol{\phi}_i(\mathbf{x})^\top, \sigma^2(\mathbf{x})), \end{aligned}$$

- RVM concept can be used not only for regression, but also for classification, kernel density estimation, sparse kernel PCA, robust regression, etc.

- With RVM we can obtain not only point prediction, but also its uncertainty. Let $\boldsymbol{\alpha}^*$ and β^* be the hyperparameters that maximize the marginal likelihood, then the predictive distribution

$$p(y|\mathbf{x}, \mathbf{X}_m, \mathbf{Y}_m, \boldsymbol{\alpha}^*, \beta^*) = \int p(y|\mathbf{x}, \mathbf{w}, \beta^*) p(\mathbf{w}|\mathbf{X}_m, \mathbf{Y}_m, \boldsymbol{\alpha}^*, \beta^*) d\mathbf{w}$$
$$= \mathcal{N}(y|\boldsymbol{\omega}_m \cdot \boldsymbol{\phi}_i(\mathbf{x})^\top, \sigma^2(\mathbf{x})),$$

where

- RVM concept can be used not only for regression, but also for classification, kernel density estimation, sparse kernel PCA, robust regression, etc.

- With RVM we can obtain not only point prediction, but also its uncertainty. Let $\boldsymbol{\alpha}^*$ and β^* be the hyperparameters that maximize the marginal likelihood, then the predictive distribution

$$\begin{aligned} p(y|\mathbf{x}, \mathbf{X}_m, \mathbf{Y}_m, \boldsymbol{\alpha}^*, \beta^*) &= \int p(y|\mathbf{x}, \mathbf{w}, \beta^*) p(\mathbf{w}|\mathbf{X}_m, \mathbf{Y}_m, \boldsymbol{\alpha}^*, \beta^*) d\mathbf{w} \\ &= \mathcal{N}(y|\boldsymbol{\omega}_m \cdot \boldsymbol{\phi}_i(\mathbf{x})^\top, \sigma^2(\mathbf{x})), \end{aligned}$$

where

$$\sigma^2(\mathbf{x}) = (\beta^*)^{-1} + \boldsymbol{\phi}(\mathbf{x})^\top \mathbf{S}_m \boldsymbol{\phi}(\mathbf{x})$$

- RVM concept can be used not only for regression, but also for classification, kernel density estimation, sparse kernel PCA, robust regression, etc.

- With RVM we can obtain not only point prediction, but also its uncertainty. Let $\boldsymbol{\alpha}^*$ and β^* be the hyperparameters that maximize the marginal likelihood, then the predictive distribution

$$
\begin{aligned}
p(y|\mathbf{x}, \mathbf{X}_m, \mathbf{Y}_m, \boldsymbol{\alpha}^*, \beta^*) &= \int p(y|\mathbf{x}, \mathbf{w}, \beta^*) p(\mathbf{w}|\mathbf{X}_m, \mathbf{Y}_m, \boldsymbol{\alpha}^*, \beta^*) d\mathbf{w} \\
&= \mathcal{N}(y|\boldsymbol{\omega}_m \cdot \boldsymbol{\phi}_i(\mathbf{x})^\top, \sigma^2(\mathbf{x})),
\end{aligned}
$$

where

$$
\begin{aligned}
\sigma^2(\mathbf{x}) &= (\beta^*)^{-1} + \boldsymbol{\phi}(\mathbf{x})^\top \mathsf{S}_m \phi(\mathbf{x}) \\
\boldsymbol{\omega}_m &= \beta \mathsf{S}_m \boldsymbol{\Phi}^\top \mathbf{Y}_m \\
\mathsf{S}_m &= (\boldsymbol{\alpha} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1},
\end{aligned}
$$

- RVM concept can be used not only for regression, but also for classification, kernel density estimation, sparse kernel PCA, robust regression, etc.

- With RVM we can obtain not only point prediction, but also its uncertainty. Let $\boldsymbol{\alpha}^*$ and β^* be the hyperparameters that maximize the marginal likelihood, then the predictive distribution

$$\begin{aligned} p(y|\mathbf{x}, \mathbf{X}_m, \mathbf{Y}_m, \boldsymbol{\alpha}^*, \beta^*) &= \int p(y|\mathbf{x}, \mathbf{w}, \beta^*) p(\mathbf{w}|\mathbf{X}_m, \mathbf{Y}_m, \boldsymbol{\alpha}^*, \beta^*) d\mathbf{w} \\ &= \mathcal{N}(y|\boldsymbol{\omega}_m \cdot \boldsymbol{\phi}_i(\mathbf{x})^\top, \sigma^2(\mathbf{x})), \end{aligned}$$

  where

$$\begin{aligned} \sigma^2(\mathbf{x}) &= (\beta^*)^{-1} + \boldsymbol{\phi}(\mathbf{x})^\top \mathbf{S}_m \boldsymbol{\phi}(\mathbf{x}) \\ \boldsymbol{\omega}_m &= \beta \mathbf{S}_m \boldsymbol{\Phi}^\top \mathbf{Y}_m \\ \mathbf{S}_m &= (\boldsymbol{\alpha} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1}, \end{aligned}$$

  with $\boldsymbol{\Phi} = \{\phi_i(\mathbf{x}_j)\} \in \mathbb{R}^{m \times M}$, $\boldsymbol{\alpha} = \mathrm{diag}(\alpha_i)$

- RVM concept can be used not only for regression, but also for classification, kernel density estimation, sparse kernel PCA, robust regression, etc.

- With RVM we can obtain not only point prediction, but also its uncertainty. Let $\boldsymbol{\alpha}^*$ and β^* be the hyperparameters that maximize the marginal likelihood, then the predictive distribution

$$
\begin{aligned}
p(y|\mathbf{x}, \mathbf{X}_m, \mathbf{Y}_m, \boldsymbol{\alpha}^*, \beta^*) &= \int p(y|\mathbf{x}, \mathbf{w}, \beta^*) p(\mathbf{w}|\mathbf{X}_m, \mathbf{Y}_m, \boldsymbol{\alpha}^*, \beta^*) d\mathbf{w} \\
&= \mathcal{N}(y|\boldsymbol{\omega}_m \cdot \boldsymbol{\phi}_i(\mathbf{x})^\top, \sigma^2(\mathbf{x})),
\end{aligned}
$$

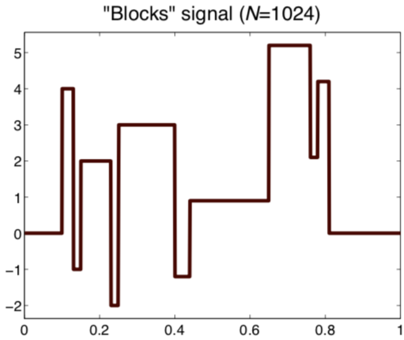
where

$$
\begin{aligned}
\sigma^2(\mathbf{x}) &= (\beta^*)^{-1} + \boldsymbol{\phi}(\mathbf{x})^\top \mathsf{S}_m \boldsymbol{\phi}(\mathbf{x}) \\
\boldsymbol{\omega}_m &= \beta \mathsf{S}_m \boldsymbol{\Phi}^\top \mathbf{Y}_m \\
\mathsf{S}_m &= (\boldsymbol{\alpha} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1},
\end{aligned}
$$

with $\boldsymbol{\Phi} = \{\phi_i(\mathbf{x}_j)\} \in \mathbb{R}^{m \times M}$, $\boldsymbol{\alpha} = \mathrm{diag}(\alpha_i)$

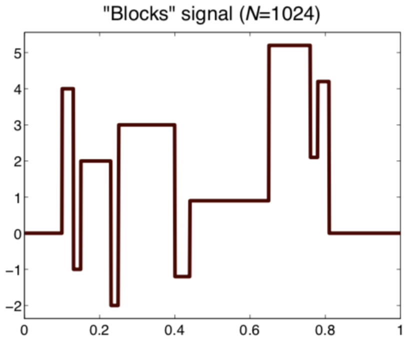
- RVM concept can be used not only for regression, but also for classification, kernel density estimation, sparse kernel PCA, robust regression, etc.

- Agglomerative algorithms (e.g. "matching pursuit") are often greedy — i.e. "early" additions can be significantly sub-optimal
- Demonstration: a popular signal processing test data set



- Approximate with a basis comprising:
  - "heavyside" step functions (easy)
  - "heayvyside" and Gaussians (hard?)

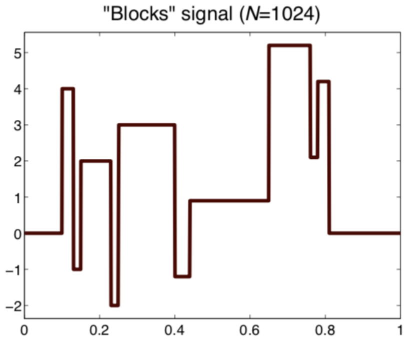
- Agglomerative algorithms (e.g. "matching pursuit") are often greedy — i.e. "early" additions can be significantly sub-optimal
- Demonstration: a popular signal processing test data set



- Approximate with a basis comprising:
  - "heavyside" step functions (easy)
  - "heayvyside" and Gaussians (hard?)

Greediness?

- Agglomerative algorithms (e.g. "matching pursuit") are often greedy — i.e. "early" additions can be significantly sub-optimal
- Demonstration: a popular signal processing test data set



- Approximate with a basis comprising:
  - "heavyside" step functions (easy)
  - "heayvyside" and Gaussians (hard?)

"Blocks" Data Results Summary

	Heaviside		*Heaviside + Gauss*	
	Bayes	**ORMP**	**Bayes**	**ORMP**
M	1024	1024	5120	5120
$\widehat{M}$	**12**	**12**	**12**	**82**
Iterations	21	11	224	82
Additions	11	11	107	82
Deletions	0	–	96	–
Re-estimates	10	–	21	–
Time	1.34s	1.19s	43.3s	24.6s

- Assume the target is noise-free and is to be approximated more "cheaply", e.g. an image which is to be compressed
- Choose some appropriate basis set (e.g. Gabor wavelets)
- Fix σ^2 as desired
- Run the sparse Bayes regression algorithm
- Interpretation of σ^2 has changed — it now models the approximation error, the noise process

- Assume the target is noise-free and is to be approximated more "cheaply", e.g. an image which is to be compressed
- Choose some appropriate basis set (e.g. Gabor wavelets)
- Fix σ^2 as desired
- Run the sparse Bayes regression algorithm
- Interpretation of σ^2 has changed — it now models the approximation error, the noise process

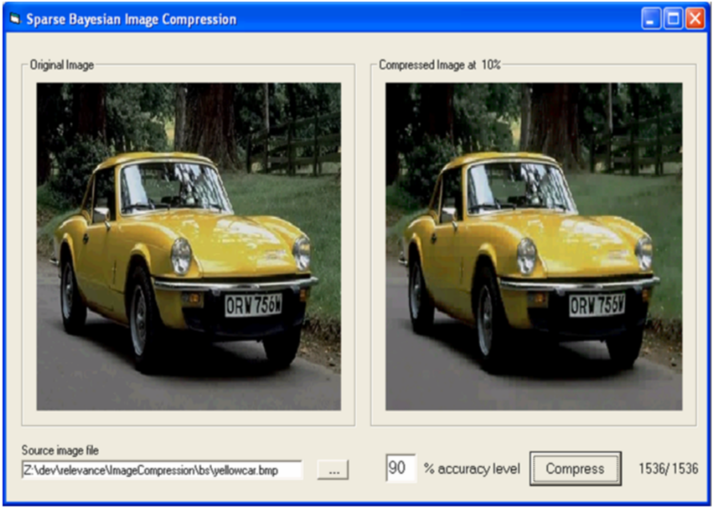
Applications: approximation (1)

- Assume the target is noise-free and is to be approximated more "cheaply", e.g. an image which is to be compressed
- Choose some appropriate basis set (e.g. Gabor wavelets)
- Fix σ^2 as desired
- Run the sparse Bayes regression algorithm
- Interpretation of σ^2 has changed — it now models the approximation error, the noise process

- Assume the target is noise-free and is to be approximated more "cheaply", e.g. an image which is to be compressed
- Choose some appropriate basis set (e.g. Gabor wavelets)
- Fix σ^2 as desired
- Run the sparse Bayes regression algorithm
- Interpretation of σ^2 has changed — it now models the approximation error, the noise process

- Assume the target is noise-free and is to be approximated more "cheaply", e.g. an image which is to be compressed
- Choose some appropriate basis set (e.g. Gabor wavelets)
- Fix σ^2 as desired
- Run the sparse Bayes regression algorithm
- Interpretation of σ^2 has changed — it now models the approximation error, the noise process

Image compression



Applications: approximation (2)

- Can approximate functions $f(\mathbf{x})$

$$\text{Likelihood} \sim \exp\left\{-\int \frac{1}{2\sigma^2}\|f(\mathbf{x};\mathbf{w}) - f(\mathbf{x})\|^2 d\mathbf{x}\right\}$$

- Condition: we need to compute all $\int \phi_i(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}$ and $\int \phi_i(\mathbf{x})\phi_j(\mathbf{x}) d\mathbf{x}$
- Practical example: $f(\mathbf{x}) = \sum_j \nu_j \psi_j(\mathbf{x})$, where ψ_j — Gaussian
- Potential target functions: Gaussian process, SVM, kernel density estimator

- Can approximate functions $f(\mathbf{x})$

$$\text{Likelihood} \sim \exp\left\{-\int \frac{1}{2\sigma^2}\|f(\mathbf{x};\mathbf{w}) - f(\mathbf{x})\|^2 d\mathbf{x}\right\}$$

- Condition: we need to compute all $\int \phi_i(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}$ and $\int \phi_i(\mathbf{x})\phi_j(\mathbf{x}) d\mathbf{x}$
- Practical example: $f(\mathbf{x}) = \sum_j \nu_j \psi_j(\mathbf{x})$, where ψ_j — Gaussian
- Potential target functions: Gaussian process, SVM, kernel density estimator

- Can approximate functions $f(\mathbf{x})$

$$\text{Likelihood} \sim \exp\left\{-\int \frac{1}{2\sigma^2}\|f(\mathbf{x};\mathbf{w}) - f(\mathbf{x})\|^2 d\mathbf{x}\right\}$$

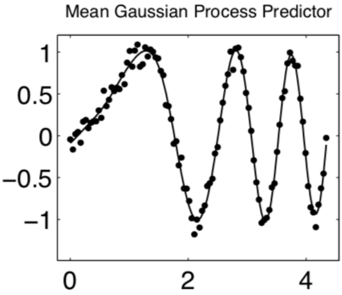
- Condition: we need to compute all $\int \phi_i(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}$ and $\int \phi_i(\mathbf{x})\phi_j(\mathbf{x}) d\mathbf{x}$
- Practical example: $f(\mathbf{x}) = \sum_j \nu_j \psi_j(\mathbf{x})$, where ψ_j — Gaussian
- Potential target functions: Gaussian process, SVM, kernel density estimator

- Can approximate functions $f(\mathbf{x})$

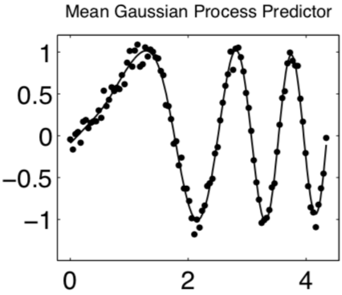
$$\text{Likelihood} \sim \exp\left\{-\int \frac{1}{2\sigma^2}\|f(\mathbf{x};\mathbf{w}) - f(\mathbf{x})\|^2 d\mathbf{x}\right\}$$

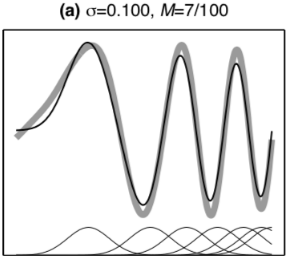
- Condition: we need to compute all $\int \phi_i(\mathbf{x}) f(\mathbf{x}) d\mathbf{x}$ and $\int \phi_i(\mathbf{x})\phi_j(\mathbf{x}) d\mathbf{x}$
- Practical example: $f(\mathbf{x}) = \sum_j \nu_j \psi_j(\mathbf{x})$, where ψ_j — Gaussian
- Potential target functions: Gaussian process, SVM, kernel density estimator

GP approximation

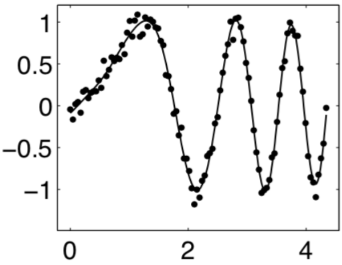


GP approximation

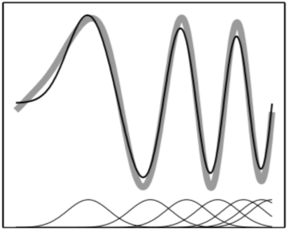




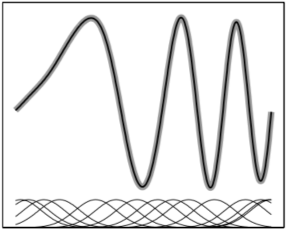
Mean Gaussian Process Predictor



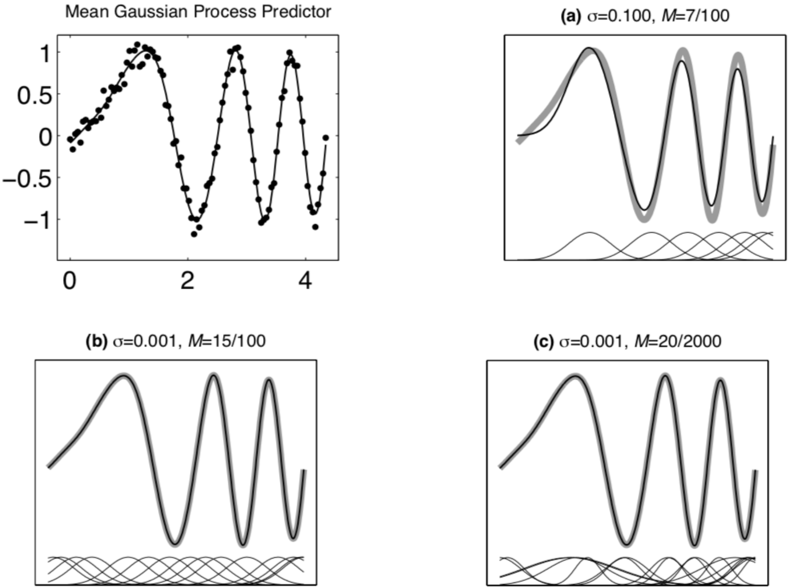
(a) σ=0.100, M=7/100



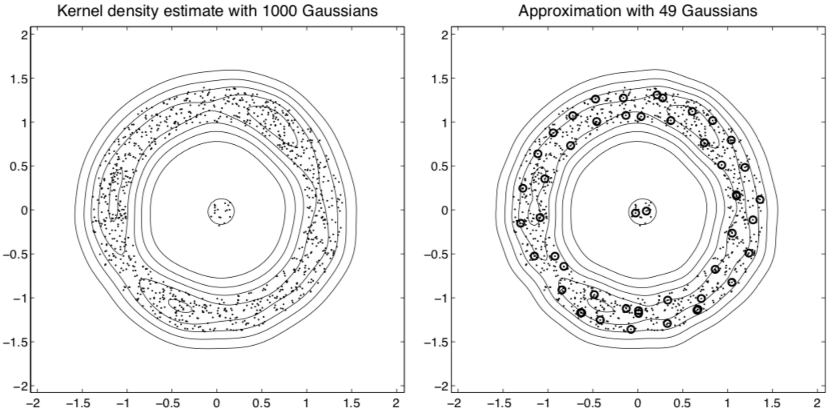
(b) σ=0.001, M=15/100



GP approximation



Kernel Density Estimator approximation



Applications: sparse kernel PCA

- Work directly with $\mathbf{C} = \sum_{i=1}^{m} \alpha_i^{-1} \phi_i \phi_i^\top + \sigma^2 \mathbf{I}$