Multivariate Bayesian Linear Regression

MLAI Lecture 12

Neil D. Lawrence

Department of Computer Science
Sheffield University

21st October 2012

Outline

Bayesian Polynomials

Revisit Olympics Data

- Use Bayesian approach on olympics data with polynomials.
- Choose a prior $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \alpha \mathbf{I})$ with $\alpha = 1$.
- Choose noise variance $\sigma^2 = 0.01$

Sampling the Prior

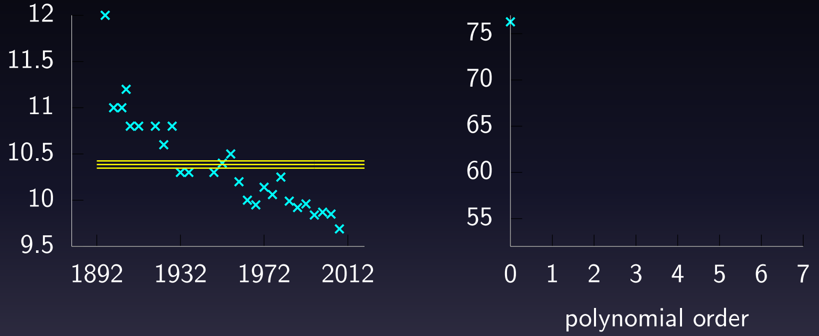
- Always useful to perform a 'sanity check' and sample from the prior before observing the data.
- Since $\mathbf{t} = \mathbf{\Phi}\mathbf{w} + \boldsymbol{\epsilon}$ just need to sample

$$w \sim \mathcal{N}(0, \alpha)$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\right)$$

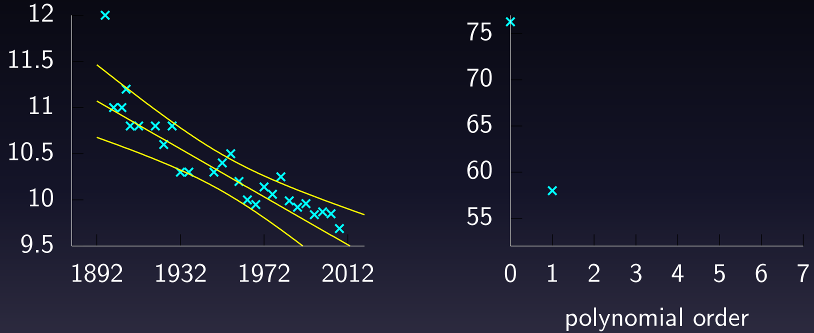
with $\alpha = 1$ and $\boldsymbol{\epsilon} = 0.01$.

Polynomial Fits to Olympics Data



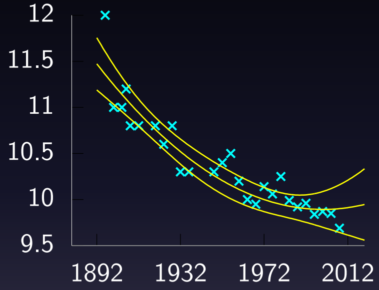
Left: fit to data, *Right*: marginal log likelihood. Polynomial order 0, model error 76.292, $\sigma^2 = 0.268$, $\sigma = 0.518$.

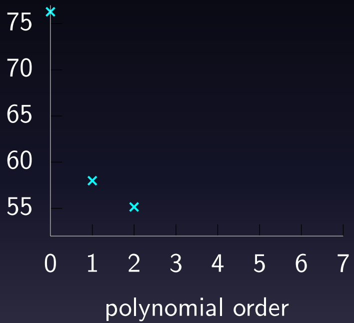
Polynomial Fits to Olympics Data



Left: fit to data, *Right*: marginal log likelihood. Polynomial order 1, model error 57.991, $\sigma^2 = 0.0609$, $\sigma = 0.247$.

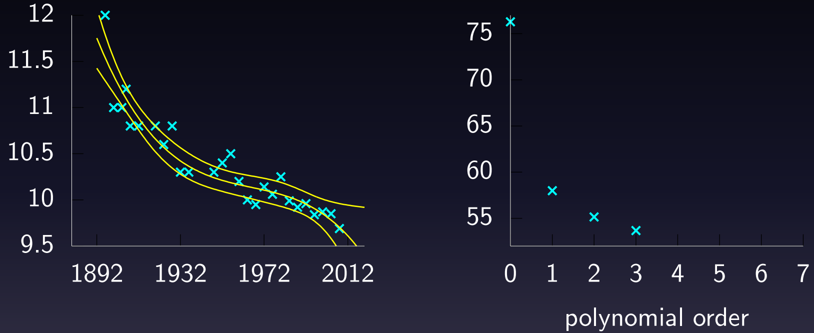
Polynomial Fits to Olympics Data





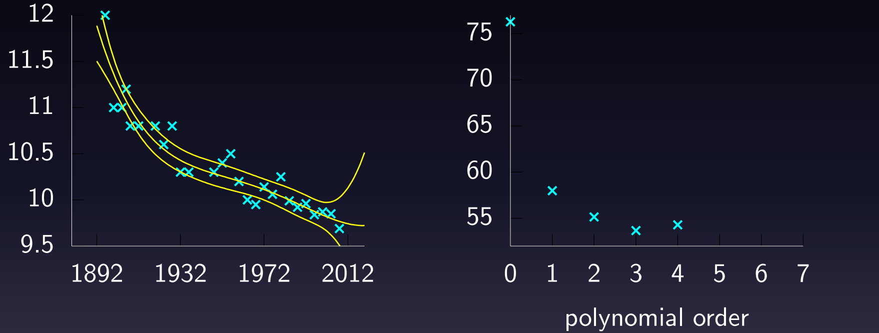
Left: fit to data, *Right*: marginal log likelihood. Polynomial order 2, model error 55.155, $\sigma^2 = 0.0391$, $\sigma = 0.198$.

Polynomial Fits to Olympics Data



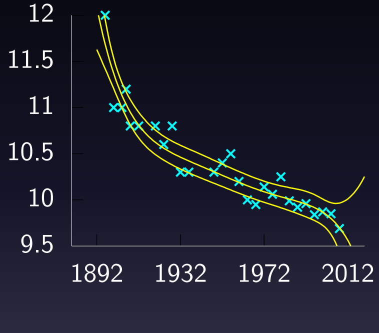
Left: fit to data, *Right*: marginal log likelihood. Polynomial order 3, model error 53.683, $\sigma^2 = 0.0301$, $\sigma = 0.173$.

Polynomial Fits to Olympics Data



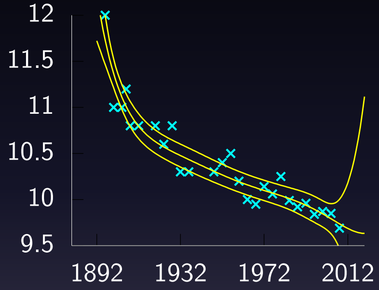
Left: fit to data, *Right*: marginal log likelihood. Polynomial order 4, model error 54.301, $\sigma^2 = 0.0277$, $\sigma = 0.166$.

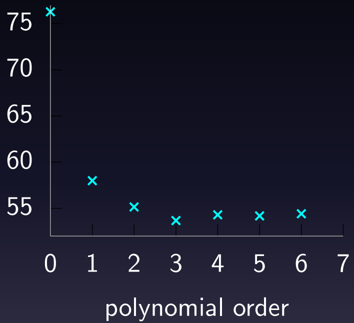
Polynomial Fits to Olympics Data



Left: fit to data, *Right*: marginal log likelihood. Polynomial order 5, model error 54.177, $\sigma^2 = 0.0249$, $\sigma = 0.158$.

Polynomial Fits to Olympics Data



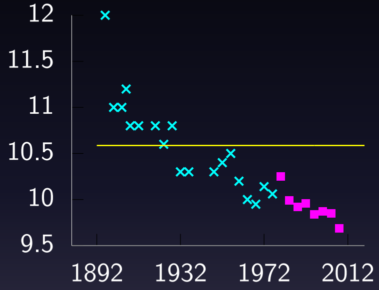


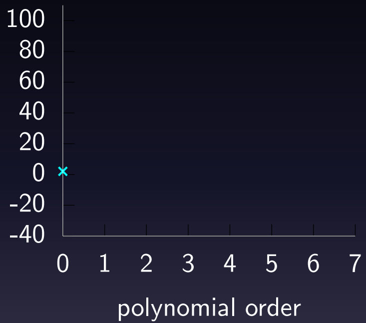
Left: fit to data, *Right*: marginal log likelihood. Polynomial order 6, model error 54.415, $\sigma^2 = 0.0236$, $\sigma = 0.154$.

Model Fit

- Marginal likelihood doesn't always increase as model order increases.
- Bayesian model always has 2 parameters, regardless of how many basis functions (and here we didn't even fit them).
- Maximum likelihood model over fits through increasing number of parameters.
- Revisit maximum likelihood solution with validation set.

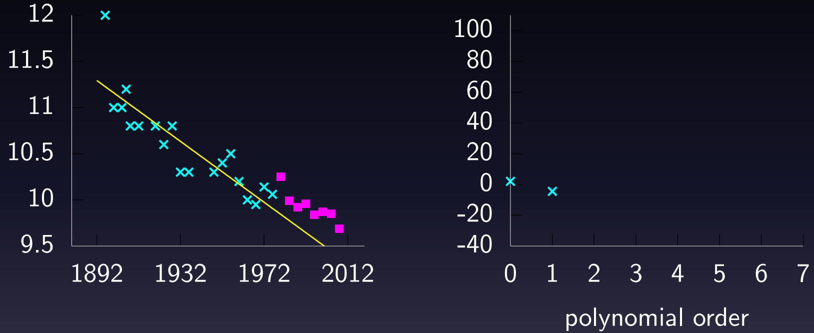
Recall: Validation Set for Maximum Likelihood





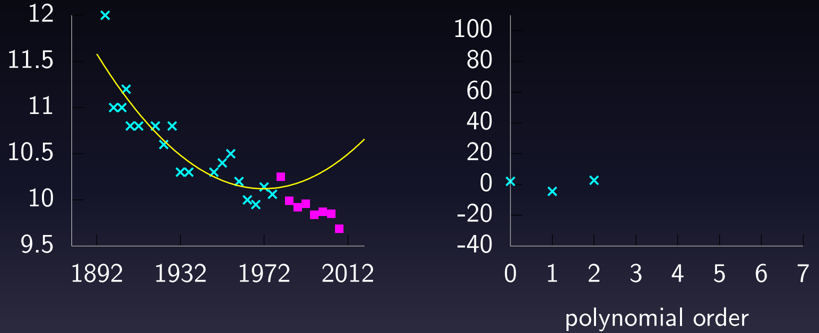
Left: fit to data, *Right*: model error. Polynomial order 0, training error -4.0526, validation error 2.0524, $\sigma^2 = 0.240$, $\sigma = 0.490$.

Recall: Validation Set for Maximum Likelihood



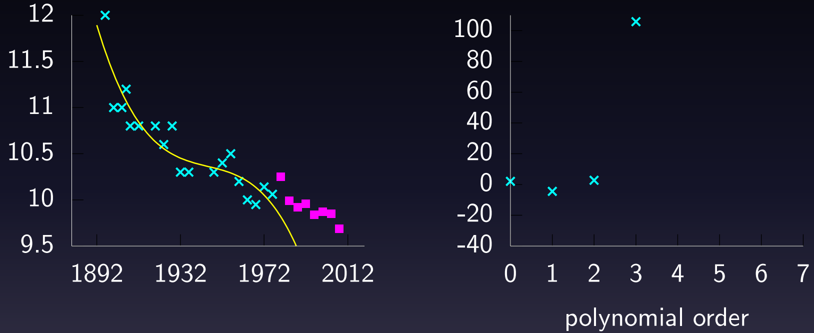
Left: fit to data, *Right*: model error. Polynomial order 1, training error -17.519, validation error -4.4127, $\sigma^2 = 0.0582$, $\sigma = 0.241$.

Recall: Validation Set for Maximum Likelihood



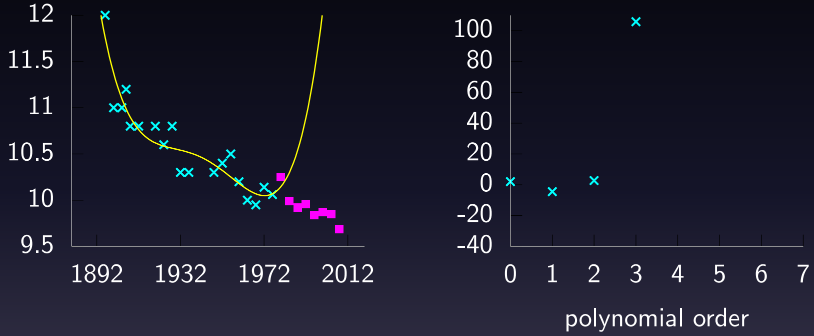
Left: fit to data, *Right*: model error. Polynomial order 2, training error -20.159, validation error 2.7275, $\sigma^2 = 0.0441$, $\sigma = 0.210$.

Recall: Validation Set for Maximum Likelihood



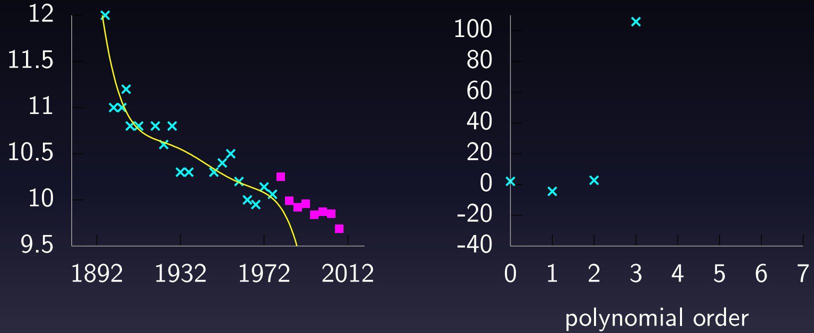
Left: fit to data, *Right*: model error. Polynomial order 3, training error -22.172, validation error 105.8, $\sigma^2 = 0.0357$, $\sigma = 0.189$.

Recall: Validation Set for Maximum Likelihood



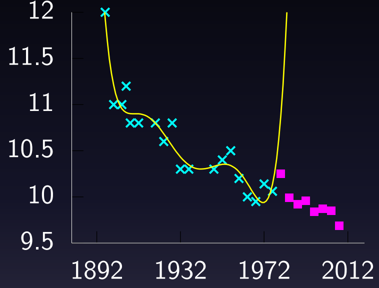
Left: fit to data, *Right*: model error. Polynomial order 4, training error -23.781, validation error 578.29, $\sigma^2 = 0.0301$, $\sigma = 0.173$.

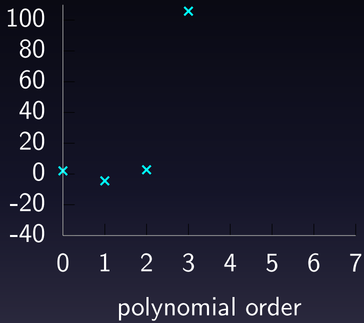
Recall: Validation Set for Maximum Likelihood



Left: fit to data, *Right*: model error. Polynomial order 5, training error -24.136, validation error 746.57, $\sigma^2 = 0.0290$, $\sigma = 0.170$.

Recall: Validation Set for Maximum Likelihood





Left: fit to data, *Right*: model error. Polynomial order 6, training error -28.528, validation error 3.3585e+05, $\sigma^2 = 0.0183$, $\sigma = 0.135$.

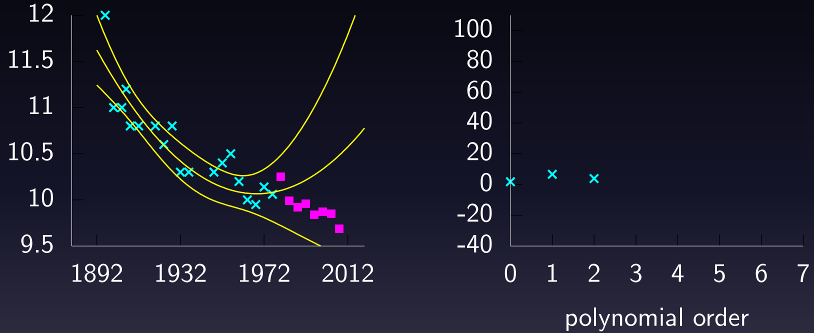
Validation Set

Left: fit to data, *Right*: model error. Polynomial order 0, training error 76.292, validation error 1.761, $\sigma^2 = 0.240$, $\sigma = 0.490$.

Validation Set

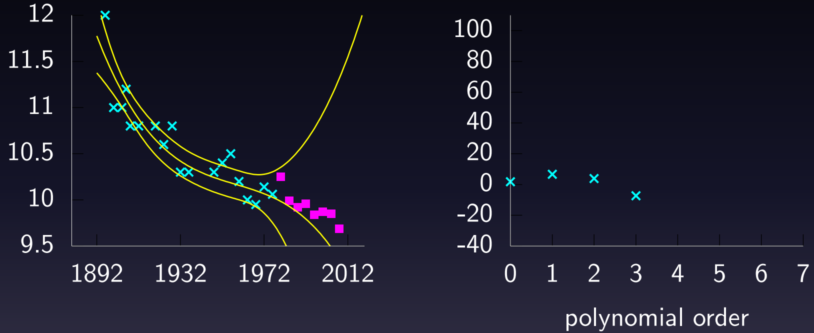
Left: fit to data, *Right*: model error. Polynomial order 1, training error 57.991, validation error 6.6482, $\sigma^2 = 0.0778$, $\sigma = 0.279$.

Validation Set



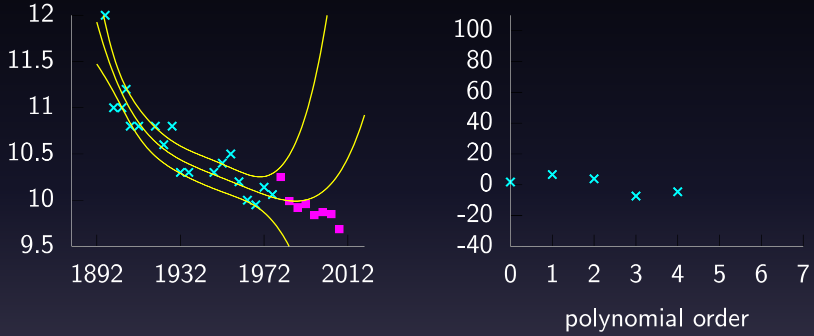
Left: fit to data, *Right*: model error. Polynomial order 2, training error 55.155, validation error 3.8897, $\sigma^2 = 0.0467$, $\sigma = 0.216$.

Validation Set



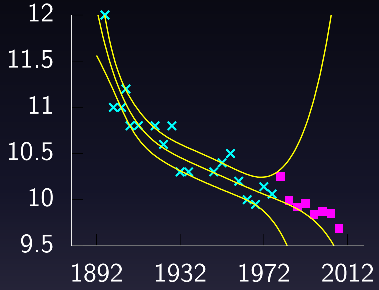
Left: fit to data, *Right*: model error. Polynomial order 3, training error 53.683, validation error -7.3484, $\sigma^2 = 0.0392$, $\sigma = 0.198$.

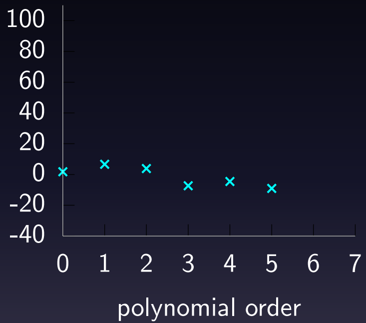
Validation Set



Left: fit to data, *Right*: model error. Polynomial order 4, training error 54.301, validation error -4.5232, $\sigma^2 = 0.0353$, $\sigma = 0.188$.

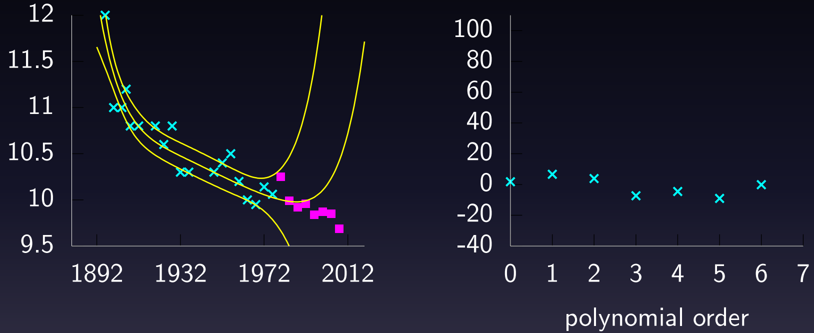
Validation Set





Left: fit to data, *Right*: model error. Polynomial order 5, training error 54.177, validation error -9.0875, $\sigma^2 = 0.0326$, $\sigma = 0.181$.

Validation Set



Left: fit to data, *Right*: model error. Polynomial order 6, training error 54.415, validation error -0.077841, $\sigma^2 = 0.0305$, $\sigma = 0.175$.

Regularized Mean

- Validation fit here based on mean solution for **w** only.
- For Bayesian solution

$$\boldsymbol{\mu}_w = \left[\sigma^{-2}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \alpha^{-1}\mathbf{I}\right]^{-1}\sigma^{-2}\boldsymbol{\Phi}^\top\mathbf{t}$$

  instead of

$$\mathbf{w}^* = \left[\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right]^{-1}\boldsymbol{\Phi}^\top\mathbf{t}$$

- Two are equivalent when $\alpha \to \infty$.
- Equivalent to a prior for **w** with infinite variance.
- In other cases $\alpha\mathbf{I}$ *regularizes* the system (keeps parameters smaller).

Sampling the Posterior

- Now check samples by extracting $\mathbf{w}$ from the *posterior*.
- Now for $\mathbf{t} = \boldsymbol{\Phi}\mathbf{w} + \epsilon$ need

$$w \sim \mathcal{N}\left(\boldsymbol{\mu}_w, \mathbf{C}_w\right)$$

with $\mathbf{C}_w = \left[\sigma^{-2}\boldsymbol{\Phi}^\top\boldsymbol{\Phi} + \alpha^{-1}\mathbf{I}\right]^{-1}$ and $\boldsymbol{\mu}_w = \mathbf{C}_w\sigma^{-2}\boldsymbol{\Phi}^\top\mathbf{t}$

$$\boldsymbol{\epsilon} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\right)$$

with $\alpha = 1$ and $\boldsymbol{\epsilon} = 0.01$.

Marginal Likelihood

- The marginal likelihood can also be computed, it has the form:

$$p(\mathbf{t}|\mathbf{X}, \sigma^2, \alpha) = \frac{1}{(2\pi)^{\frac{N}{2}} |\mathbf{K}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\mathbf{t}^\top \mathbf{K}^{-1}\mathbf{t}\right)$$

  where $\mathbf{K} = \alpha\mathbf{\Phi}\mathbf{\Phi}^\top + \sigma^2\mathbf{I}$.
- So it is a zero mean N-dimensional Gaussian with covariance matrix $\mathbf{K}$.

Computing the Expected Output

- Given the posterior for the parameters, how can we compute the expected output at a given location?
- Output of model at location $\mathbf{x}_i$ is given by

$$y(\mathbf{x}_i; \mathbf{w}) = \phi_i^\top \mathbf{w}$$

- We want the expected output under the posterior density, $p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2, \alpha)$.
- Mean of mapping function will be given by

$$\begin{aligned}\left\langle y(\mathbf{x}_i; \mathbf{w})\right\rangle_{p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma^2,\alpha)} &= \phi_i^\top \left\langle \mathbf{w}\right\rangle_{p(\mathbf{w}|\mathbf{t},\mathbf{X},\sigma^2,\alpha)} \\ &= \phi_i^\top \boldsymbol{\mu}_w\end{aligned}$$

Variance of Expected Output

- Variance of model at location $\mathbf{x}_i$ is given by

$$\begin{aligned}\text{var}(y(\mathbf{x}_i;\mathbf{w})) &= \left\langle (y(\mathbf{x}_i;\mathbf{w}))^2 \right\rangle - \left\langle y(\mathbf{x}_i;\mathbf{w}) \right\rangle^2 \\ &= \phi_i^\top \left\langle \mathbf{w}\mathbf{w}^\top \right\rangle \phi_i - \phi_i^\top \left\langle \mathbf{w} \right\rangle \left\langle \mathbf{w} \right\rangle^\top \phi_i \\ &= \phi_i^\top \mathbf{C}_i \phi_i\end{aligned}$$

where all these expectations are taken under the posterior density, $p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2, \alpha)$.

Reading

- Section 3.7–3.8 of Rogers and Girolami (pg 122–133).
- Section 3.4 of Bishop (pg 161–165).

References I

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, 2006. [Google Books] .

S. Rogers and M. Girolami. *A First Course in Machine Learning*. CRC Press, 2011. [Google Books] .