Maximum Likelihood and Bayesian regression

Jerónimo Arenas-García

Universidad Carlos III de Madrid

jeronimo.arenas@uc3m.es

October 18, 2020

Contents

Maximum Likelihood Estimation of pdf parameters

Assume we have observations i.i.d. $\{\mathbf{x_k}\}$ from a distribution with unknown parameters $\mathbf{w}$.

Maximum likelihood (ML) estimation of $\mathbf{w}$

- We can measure fitness of data for particular $\mathbf{w}$ using the likelihood function

$$p(\mathbf{x} \mid \mathbf{w}) = p(x_0, x_1, \cdots, x_{K-1} \mid \mathbf{w}) = \prod_{k=0}^{K-1} p(x_k \mid \mathbf{w})$$

- The ML estimator is the one with largest likelihood

$$\hat{\mathbf{w}}_{ML} = \arg\max_{\mathbf{w}} p(\mathbf{x} \mid \mathbf{w}) = \arg\max_{\mathbf{w}} \sum_{k=0}^{K-1} \log p(x_k \mid \mathbf{w})$$

ML Estimation: parameters of a Gaussian distribution

Exercise: Assume we have observations i.i.d. $\{\mathbf{x_k}\}$ from a Gaussian distribution with unknown mean and variance

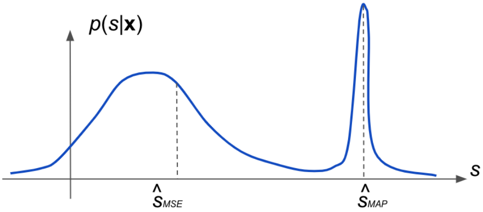
$$x \sim \mathcal{N}(m, v)$$

Obtain the ML estimator of the mean and the variance of the distribution.

Estimation Theory

- Knowing $p(s \mid \mathbf{x})$ we can design different analytical estimators of s
- $p(s \mid \mathbf{x})$ summarizes all that can be known about the possible values of s for every single $\mathbf{x}$

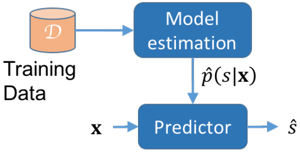
$$\hat{s}_{\text{MSE}} = \mathbb{E}\{s \mid \mathbf{x}\}$$
$$\hat{s}_{\text{MAP}} = \arg\max_{s} p(s \mid \mathbf{x})$$



Maximum Likelihood Regression

ML regression

- Since we know how to proceed from $p(s \mid \mathbf{x})$, we will estimate it
- We assume a parametric model $p(s \mid \mathbf{x}, \mathbf{w})$, and estimate $\mathbf{w}$ using ML
- If the model is not good, the regression model will be poor



Summary steps

1. Propose a model $p(s \mid \mathbf{x}, \mathbf{w})$
2. Calculate $p(\mathbf{s} \mid \mathbf{X}, \mathbf{w})$ (i.e., for the available training data)
3. Calculate $\mathbf{w}_{ML}$
4. Obtain $\hat{s}_{\mathrm{MSE}}$ or $\hat{s}_{\mathrm{MAP}}$ from $p(s \mid \mathbf{x}, \mathbf{w}_{ML})$

Model Assumptions

Notation

- $\mathbf{s} = (s_0, \ldots, s_{K-1})^\top$
- $\mathbf{X} = (\mathbf{x}_0, \ldots, \mathbf{x}_{K-1})^\top$
- $\mathcal{D} = (\mathbf{s}, \mathbf{X})$

Model Assumptions

Some assumptions are generally required for step 2

1. All samples in $\mathcal{D}$ have been generated by the same distribution, $p(s, \mathbf{x} \mid \mathbf{w})$
2. Input variables $\mathbf{x}$ do not depend on $\mathbf{w}$: $p(\mathbf{X} \mid \mathbf{w}) = p(\mathbf{X})$
3. Targets s_k are independent, given $\mathbf{w}$ and the inputs $\mathbf{x}_k$:

$$p(\mathbf{s} \mid \mathbf{X}, \mathbf{w}) = \prod_{k=0}^{K-1} p(s_k \mid \mathbf{x}_k, \mathbf{w})$$

Steps 2 and 3 using the assumptions

Using assumptions 1 and 2:

$$p(\mathcal{D}|\mathbf{w}) = p(\mathbf{s}, \mathbf{X}|\mathbf{w}) = p(\mathbf{s}|\mathbf{X}, \mathbf{w})p(\mathbf{X}|\mathbf{w}) = p(\mathbf{s}|\mathbf{X}, \mathbf{w})p(\mathbf{X})$$

Using 3:

$$\begin{aligned}
\hat{\mathbf{w}}_{\mathrm{ML}} &= \arg\max_{\mathbf{w}} p(\mathbf{s}|\mathbf{X}, \mathbf{w}) \\
&= \arg\max_{\mathbf{w}} \prod_{k=0}^{K-1} p(s_k \mid \mathbf{x}_k, \mathbf{w}) \\
&= \arg\max_{\mathbf{w}} \sum_{k=0}^{K-1} \log p(s_k \mid \mathbf{x}_k, \mathbf{w})
\end{aligned}$$

Gaussian model (I)

Step 1

We assume that targets are generated as

$$s_k = \mathbf{w}^\top \mathbf{x}_k + \varepsilon_k; \qquad \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$$

This is equivalent to assumming the parametric form

$$p(s \mid \mathbf{x}, \mathbf{w}) \sim \mathcal{N}(\mathbf{w}^\top \mathbf{x}, \sigma_\varepsilon^2)$$

Step 4

Since the distribution is Gaussian:

$$\hat{s}_{\text{MSE}} = \hat{s}_{\text{MAP}} = \mathbf{w}^\top \mathbf{x}$$

We will turn our attention to steps 2 and 3 shortly. Before that, we explain how we can extend the model to obtain a non-linear regression in an easy manner

Non-linear models using feature transformations

- We have seen that the previous setup will produce a linear regression model
- However, we can apply any transformations to the available features

$$\mathbf{x} \longrightarrow \boxed{t(\mathbf{x})} \longrightarrow \mathbf{z} = [t_0(\mathbf{x}),\ t_1(\mathbf{x}),\ \ldots,\ t_m(\mathbf{x})]^\top$$

E.g.:

$$\mathbf{z} = [1,\ x_0,\ x_1,\ x_0^2 \cdot x_1,\ \log(x_0 \cdot x_1)]^\top$$

- We can transform all training patterns and build a model based on $\mathbf{z}$

$$\mathcal{D}_\mathbf{x} = \{\mathbf{s}, \mathbf{X}\} \longrightarrow \mathcal{D}_\mathbf{z} = \{\mathbf{s}, \mathbf{Z}\}$$

- For evaluating the model, test data is accordingly trasnformed
- A linear regression model w.r.t. $\mathbf{z}$ is non-linear w.r.t. $\mathbf{x}$

Gaussian model (II)

Step 1

We assume that targets are generated as

$$s_k = \mathbf{w}^\top \mathbf{z}_k + \varepsilon_k; \qquad \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$$

This is equivalent to assumming the parametric form

$$p(s \mid \mathbf{x}, \mathbf{w}) \sim \mathcal{N}(\mathbf{w}^\top \mathbf{z}, \sigma_\varepsilon^2)$$

Step 4

Since the distribution is Gaussian:

$$\hat{s}_{\text{MSE}} = \hat{s}_{\text{MAP}} = \mathbf{w}^\top \mathbf{z}$$

Gaussian model (III): Step 2

Since $p(s \mid \mathbf{x}, \mathbf{w}) \sim \mathcal{N}(\mathbf{w}^\top \mathbf{z}, \sigma_\varepsilon^2)$

$$
\begin{aligned}
p(\mathbf{s}|\mathbf{X}, \mathbf{w}) &= \prod_{k=0}^{K-1} p(s_k|\mathbf{x}_k, \mathbf{w}) \\
&= \prod_{k=0}^{K-1} \frac{1}{\sqrt{2\pi}\sigma_\varepsilon} \exp\left(-\frac{\left(s_k - \mathbf{w}^\top \mathbf{z}_k\right)^2}{2\sigma_\varepsilon^2}\right) \\
&= \left(\frac{1}{\sqrt{2\pi}\sigma_\varepsilon}\right)^K \exp\left(-\sum_{k=1}^{K} \frac{\left(s_k - \mathbf{w}^\top \mathbf{z}_k\right)^2}{2\sigma_\varepsilon^2}\right) \\
&= \left(\frac{1}{\sqrt{2\pi}\sigma_\varepsilon}\right)^K \exp\left(-\frac{1}{2\sigma_\varepsilon^2}\|\mathbf{s} - \mathbf{Z}\mathbf{w}\|^2\right)
\end{aligned}
$$

where $\mathbf{Z}$ is the *transformed* data input matrix, containing vectors $\mathbf{z}_k$ arranged rowwise.

Note that $p(\mathbf{s}|\mathbf{X}, \mathbf{w})$ is a function of $\mathbf{w}$ (when evaluated over a fixed training data set)

Gaussian model (IV): Step 3

- Maximizing the likelihood is equivalent to minimizing $\|\mathbf{s} - \mathbf{Z}\mathbf{w}\|^2$
- This is the least squares solution
- The problem is quadratic and thus can be solved by differentiation

$$\nabla_{\mathbf{w}} \|\mathbf{s} - \mathbf{Z}\mathbf{w}\|^2 \Big|_{\mathbf{w}=\mathbf{w}_{\mathrm{ML}}} = -2\mathbf{Z}^\top \mathbf{s} + 2\mathbf{Z}^\top \mathbf{Z}\mathbf{w}_{\mathrm{ML}} = \mathbf{0}$$

$$\mathbf{w}_{\mathrm{ML}} = (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{s}$$

- A closed-form expression based just on training data set
- Stable implementations avoid the computation of the inverse matrix (Notebook, Exercise 5)

Parametric exponential model (Notebook, Exercise 6)

Consider a one-dimension regression problem with $\mathcal{D} = \{s_k, x_k\}$

Step 1: Assumption of a parametric model

$$p(s \mid x, w) = wx \exp(-wxs), \qquad s \geq 0,\ x \geq 0,\ w \geq 0$$

Step 2: Compute the likelihood function

Step 3: Obtain the ML solution

Step 4: Obtain $\hat{s}_{\text{MSE}}$ using $\mathbf{w}_{\text{ML}}$

Parametric exponential model (Notebook, Exercise 6)

Consider a one-dimension regression problem with $\mathcal{D} = \{s_k, x_k\}$

Step 1: Assumption of a parametric model

$$p(s \mid x, w) = wx \exp(-wxs), \qquad s \geq 0,\ x \geq 0,\ w \geq 0$$

Step 2: Compute the likelihood function

$$p(\mathbf{s}|w, \mathbf{X}) = \prod_{k=0}^{K-1} wx_k \exp(-wx_k s_k) = w^K \left(\prod_{k=0}^{K-1} x_k\right) \exp\left(-w \sum_{k=0}^{K-1} x_k s_k\right)$$

Step 3: Obtain the ML solution

Step 4: Obtain $\hat{s}_{\text{MSE}}$ using $\mathbf{w}_{\text{ML}}$

Parametric exponential model (Notebook, Exercise 6)

Consider a one-dimension regression problem with $\mathcal{D} = \{s_k, x_k\}$

Step 1: Assumption of a parametric model

$$p(s \mid x, w) = wx \exp(-wxs), \qquad s \geq 0,\ x \geq 0,\ w \geq 0$$

Step 2: Compute the likelihood function

$$p(\mathbf{s}|w, \mathbf{X}) = \prod_{k=0}^{K-1} wx_k \exp(-wx_k s_k) = w^K \left(\prod_{k=0}^{K-1} x_k \right) \exp\left(-w \sum_{k=0}^{K-1} x_k s_k \right)$$

Step 3: Obtain the ML solution

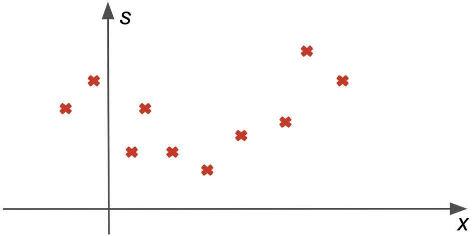
$$w_{\text{ML}} = K / \mathbf{x}^\top \mathbf{s}$$

Step 4: Obtain $\hat{s}_{\text{MSE}}$ using $\mathbf{w}_{\text{ML}}$

$$\hat{s}_{\text{MSE}} = \mathbb{E}\{s|x, w_{\text{ML}}\} = \int swx \exp(w_{\text{ML}} xs) ds = \frac{1}{w_{\text{ML}} x}$$

The problem of ML estimation (I)

Consider the regression problem with training data set

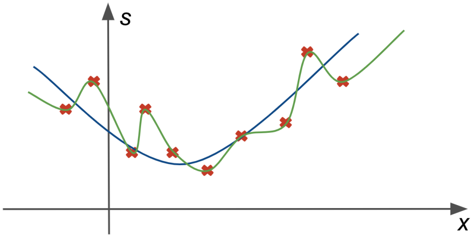


Considering the parametric model $p(s \mid x, \mathbf{w}) = \mathcal{N}(\mathbf{w}^\top \mathbf{z}, \sigma_\varepsilon^2)$, which model could achieve a larger likelihood?

1. $\mathbf{z} = [1, x, x^2]^\top \longrightarrow \hat{s} = w_0 + w_1 x + w_2 x^2$
2. $\mathbf{z} = [1, x, x^2, \cdots, x^9]^\top \longrightarrow \hat{s} = \sum_{m=0}^{9} w_m x^m$

The problem of ML estimation (II)

Consider the regression problem with training data set



Remember that larger likelihood is achieved for smaller LS error

1. $\mathbf{z} = [1, x, x^2]^\top \longrightarrow \hat{s} = w_0 + w_1 x + w_2 x^2$
2. $\mathbf{z} = [1, x, x^2, \cdots, x^9]^\top \longrightarrow \hat{s} = \sum_{m=0}^{9} w_m x^m$

Avoiding overfitting of the ML solution

ML overfitting

- As we have seen, ML is prone to overfitting
- Complex solutions with better fit are preferred
- We expect that smooth solutions offer better generalization

Controlling overfitting

- Smoother solutions are associated to smaller $\|\mathbf{w}\|$
- We encode our belief into a *prior* for the weight vector: $p(\mathbf{w})$
- The posterior $p(\mathbf{w}|\mathcal{D})$ can be obtained using Bayes' rule

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w}) \cdot p(\mathbf{w})}{p(\mathcal{D})}$$

- Instead of using $\mathbf{w}_{\mathrm{ML}}$ we can use the maximum of $p(\mathbf{w}|\mathcal{D})$

Maximizing the posterior distribution

$$\mathbf{w}_{\text{MAP}} = \arg\max_{\mathbf{w}} \frac{p(\mathcal{D}|\mathbf{w}) \cdot p(\mathbf{w})}{p(\mathcal{D})} = \arg\max_{\mathbf{w}} p(\mathcal{D}|\mathbf{w}) \cdot p(\mathbf{w})$$

- Likelihood $p(\mathcal{D}|\mathbf{w})$ controls model fit to training data
- Prior distribution $p(\mathbf{w})$ controls generalization
- $\mathbf{w}_{\text{MAP}}$ comes from a balance between both terms
- Other estimators of $\mathbf{w}$ can also be used, e.g., $\mathbf{w}_{\text{MSE}} = \mathbb{E}\{\mathbf{w} \mid \mathcal{D}\}$

Bayesian regression

Summary steps

1. Assume a parametric data model $p(s|\mathbf{x}, \mathbf{w})$
   Assume a prior distribution $p(\mathbf{w})$
2. Calculate the likelihood function $p(\mathbf{s}|\mathbf{w})$
3. Applying the Bayes' rule, compute the posterior distribution $p(\mathbf{w}|\mathbf{s})$
4. Compute the MAP or the MSE estimate of $\mathbf{w}$ given $\mathbf{x}$

$$\mathbf{w}_{\text{MAP}}/\mathbf{w}_{\text{MSE}} \longrightarrow \mathbf{w}^*$$

5. Obtain $\hat{s}_{\text{MSE}}$ or $\hat{s}_{\text{MAP}}$ from $p(s \mid \mathbf{x}, \mathbf{w}^*)$

Gaussian model (I)

Step 1

We assume that targets are generated as

$$s_k = \mathbf{w}^\top \mathbf{z}_k + \varepsilon_k; \quad \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2) \longrightarrow p(s \mid \mathbf{x}, \mathbf{w}) \sim \mathcal{N}(\mathbf{w}^\top \mathbf{z}, \sigma_\varepsilon^2)$$

Prior distribution $\mathbf{w} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{V}_p\right)$

Step 2

$$p(\mathbf{s}|\mathbf{X}, \mathbf{w}) = \left(\frac{1}{\sqrt{2\pi}\sigma_\varepsilon}\right)^K \exp\left(-\frac{1}{2\sigma_\varepsilon^2}\|\mathbf{s} - \mathbf{Z}\mathbf{w}\|^2\right)$$

Step 5

Since the distribution is Gaussian:

$$\hat{s}_{\text{MSE}} = \hat{s}_{\text{MAP}} = \mathbf{w}^\top \mathbf{z}$$

Gaussian model (II): Steps 3 and 4

The posterior distribution of the weights can be computed using Bayes' rule

$$p(\mathbf{w}|\mathbf{s}) = \frac{p(\mathbf{s}|\mathbf{w})\ p(\mathbf{w})}{p(\mathbf{s})}$$

Since both $p(\mathbf{s}|\mathbf{w})$ and $p(\mathbf{w})$ follow a Gaussian distribution, we know also that the joint distribution and the posterior distribution of $\mathbf{w}$ given $\mathbf{s}$ are also Gaussian. Therefore,

$$\mathbf{w} \mid \mathbf{s} \sim \mathcal{N}\left(\mathbf{w}_{\mathrm{MSE}}, \mathbf{V}_{\mathbf{w}}\right)$$

After some algebra, it can be shown that mean and the covariance matrix of the distribution are:

$$\mathbf{V}_{\mathbf{w}} = \left[\frac{1}{\sigma_\varepsilon^2}\mathbf{Z}^\top\mathbf{Z} + \mathbf{V}_p^{-1}\right]^{-1}$$

$$\mathbf{w}_{\mathrm{MSE}} = \sigma_\varepsilon^{-2}\mathbf{V}_{\mathbf{w}}\mathbf{Z}^\top\mathbf{s}$$

Gaussian model (III): Steps 3 and 4 (Demo)

$$p(\mathbf{w}|\mathbf{s}) = \frac{p(\mathbf{s}|\mathbf{w})\ p(\mathbf{w})}{p(\mathbf{s})}$$

$$p(\mathbf{w}|\mathbf{s}) = \frac{\left(\frac{1}{\sqrt{2\pi}\sigma_\varepsilon}\right)^K \exp\left(-\frac{1}{2\sigma_\varepsilon^2}\|\mathbf{s}-\mathbf{Z}\mathbf{w}\|^2\right)\ \frac{1}{(2\pi)^{D/2}|\mathbf{V}_p|^{1/2}}\exp\left(-\frac{1}{2}\mathbf{w}^\intercal\mathbf{V}_p^{-1}\mathbf{w}\right)}{p(\mathbf{s})}$$

$$p(\mathbf{w}|\mathbf{s}) = \frac{1}{(2\pi)^{D/2}|\mathbf{V}_\mathbf{w}|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{w}-\mathbf{w}_{\mathrm{MSE}})^\intercal\mathbf{V}_\mathbf{w}^{-1}(\mathbf{w}-\mathbf{w}_{\mathrm{MSE}})\right)$$

1. Identifying quadratic terms $\mathbf{w}^\top\mathbf{A}\mathbf{w}$:

$$\mathbf{V}_\mathbf{w} = \left[\frac{1}{\sigma_\varepsilon^2}\mathbf{Z}^\top\mathbf{Z} + \mathbf{V}_p^{-1}\right]^{-1}$$

2. Identifying linear terms $\mathbf{w}^\top\mathbf{b}$:

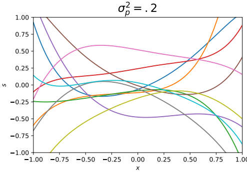
$$\mathbf{w}_{\mathrm{MSE}} = \sigma_\varepsilon^{-2}\mathbf{V}_\mathbf{w}\mathbf{Z}^\top\mathbf{s}$$

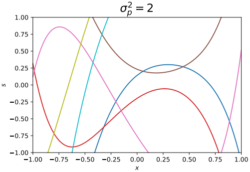
Example: The role of the prior distribution

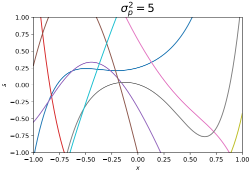
- Assume a unidimensional regression problem where the likelihood follows the previous Gaussian model with

$$s = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + w_5 x^5 + \varepsilon$$

- The prior for $\mathbf{w}$ is set as $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \sigma_p^2 \mathbf{I})$
- We depict 10 curves using weights drawn directly from the prior for three values of σ_p^2





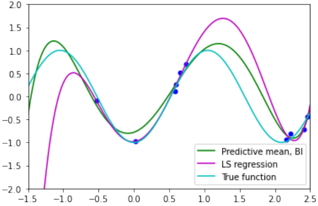


Example: Bayesian Inference vs Maximum Likelihood

- When maximizing the posterior, the prior distribution term as a sort of regularizer to control overfitting

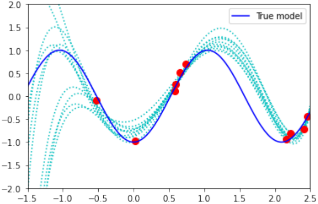
$$p(\mathbf{w}|\mathbf{s}) = \frac{p(\mathbf{s}|\mathbf{w})\ p(\mathbf{w})}{p(\mathbf{s})}$$

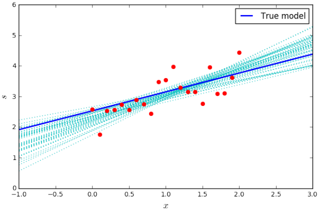
- We expect smoother solutions that generalize better



Model uncertainty (I)

- Instead of using just $\mathbf{w}_{\text{MAP}}$, we could draw solutions from $p(\mathbf{w}|\mathbf{s})$





- Drawn models tend to agree in some regions of the observation space and present more variance in other regions
- Obtain confidence intervals for the model prediction
- This we will do by computing the variance of the prediction "averaging" over the full posteriori distribution
- I.e., rather than using just one model ($\mathbf{w}_{\text{ML}}$ / $\mathbf{w}_{\text{MAP}}$) we will use all possible models averaged by $p(\mathbf{w}|\mathbf{s})$

Model uncertainty (II)

- When using $\mathbf{w}_{\text{ML}}$ / $\mathbf{w}_{\text{MAP}}$, model uncertainty is simply given by the assumed parametric model, for instance

$$p(s^*|\mathbf{w}_{\text{ML}}, \mathbf{x}^*) = \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left(-\frac{\left(s^* - \mathbf{w}_{\text{ML}}^\top \mathbf{z}^*\right)^2}{2\sigma_\varepsilon^2}\right)$$

  The variance of the prediction is constant for all $\mathbf{x}^*$

- Using Bayes Inference:

$$p(s^*|\mathbf{x}^*, \mathbf{s}) = \int p(s^*|\mathbf{w}, \mathbf{x}^*) p(\mathbf{w}|\mathbf{s}) d\mathbf{w}$$

- In general, the integral expression of the posterior distribution $p(s^*|\mathbf{x}^*, \mathbf{s})$ cannot be computed analytically. Fortunately, for the Gaussian model, the computation is feasible as the posterior of

$$s^* = \mathbf{w}^\top \mathbf{z}^* + \varepsilon$$

  is also Gaussian.

Model uncertainty for the Gaussian model

Posterior distribution of $\hat{s} = \mathbf{w}^\top \mathbf{z}^*$

- $\mathbb{E}\{\mathbf{w}^\top \mathbf{z}^* | \mathbf{x}^*, \mathbf{s}\} = \mathbb{E}\{\mathbf{w}^\top | \mathbf{s}\}\mathbf{z}^* = \mathbf{w}_{\text{MSE}}^\top \mathbf{z}^*$
- $\text{Var}\{\mathbf{w}^\top \mathbf{z}^* | \mathbf{x}^*, \mathbf{s}\} = \mathbf{z}^{*\top} \mathbb{E}\{(\mathbf{w} - \mathbf{w}_{\text{MSE}})(\mathbf{w} - \mathbf{w}_{\text{MSE}})^\top | \mathbf{s}\}\mathbf{z}^* = \mathbf{z}^{*\top} \mathbf{V}_\mathbf{w} \mathbf{z}^*$

$$\hat{s} | \mathbf{x}^*, \mathbf{s} \;\sim\; \mathcal{N}(\mathbf{w}_{\text{MSE}}^\top \mathbf{z}^*, \mathbf{z}^{*\top} \mathbf{V}_\mathbf{w} \mathbf{z}^*)$$

Posterior distribution of $s^* = \mathbf{w}^\top \mathbf{z}^* + \varepsilon$

$$s^* | \mathbf{x}^*, \mathbf{s} \;\sim\; \mathcal{N}(\mathbf{w}_{\text{MSE}}^\top \mathbf{z}^*, \mathbf{z}^{*\top} \mathbf{V}_\mathbf{w} \mathbf{z}^* + \sigma_\varepsilon^2)$$

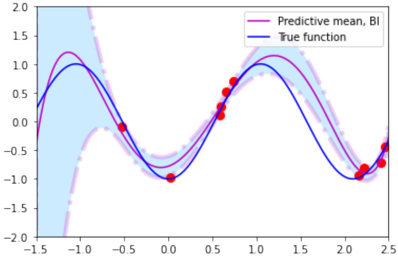
Model uncertainty for the Gaussian model

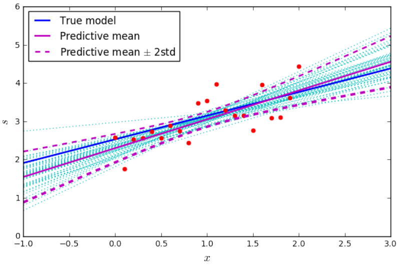
Posterior distribution of $\hat{s} = \mathbf{w}^\top \mathbf{z}^*$

$$\hat{s}|\mathbf{x}^*, \mathbf{s} \quad \sim \quad \mathcal{N}(\mathbf{w}_{\text{MSE}}^\top \mathbf{z}^*, \mathbf{z}^{*\top} \mathbf{V}_{\mathbf{w}} \mathbf{z}^*)$$

Posterior distribution of $s^* = \mathbf{w}^\top \mathbf{z}^* + \varepsilon$

$$s^*|\mathbf{x}^*, \mathbf{s} \quad \sim \quad \mathcal{N}(\mathbf{w}_{\text{MSE}}^\top \mathbf{z}^*, \mathbf{z}^{*\top} \mathbf{V}_{\mathbf{w}} \mathbf{z}^* + \sigma_\varepsilon^2)$$





Model hyperparameters selection

We have already addressed with Bayesian Inference the following two issues:

- For a given degree, how do we choose the weights?
- Should we focus on just one model, or can we use several models at once?

However, we still needed some assumptions: a parametric model (i.e., polynomial function and degree selection) and several parameters needed to be adjusted (e.g., noise variances, ...).

As we have already explained, it is possible to apply cross-validation techniques for choosing the model and its parameters.

Model hyperparameters selection

Bayesian inference opens the door to other strategies.

- We could select the most likely set of parameters according to an ML criterion
- We could argue that rather than keeping single selections of these parameters, we could use simultaneously several sets of parameters (and/or several parametric forms), and average them in a probabilistic way ... (like we did with the models)