# Machine Learning I Lecture V: Linear Regression Revisited

Jakob H Macke
Max Planck Institute for Biological Cybernetics
Bernstein Center for Computational Neuroscience

XY.XY.2012

# Plan for today

Linear regression revisited

Bayesian linear regression

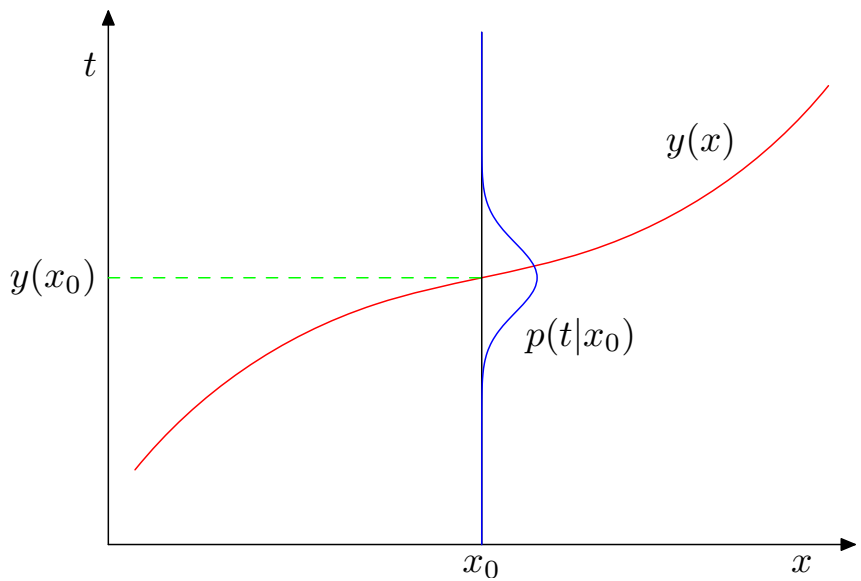
Fully Bayesian linear regression

# Linear regression can be considered as maximum likelihood estimation in a Gaussian model.

- Suppose that we have data $D = \{(\mathbf{x}_1, t_1), \ldots, (\mathbf{x}_N, t_N)\}$
- We assume that the data can be modelled by some function $t_n \approx y(x, \omega) + \epsilon$, where $\epsilon$ models additive noise.
- We assume that noise is independent, identically distributed and Gaussian:

$$\epsilon \sim \mathcal{N}(0, \beta^{-1}) \tag{1}$$

$$t|\mathbf{x}, \omega, \beta \sim \mathcal{N}(y(\mathbf{x}, \omega), \beta^{-1}) \tag{2}$$




Bishop PRML Figure 1.28

# Linear regression can be considered as maximum likelihood estimation in a Gaussian model.

- We consider a linear model $y(\mathbf{x}, \omega) = \omega^\top \mathbf{x}$
- [On board: Likelihood, Maximum Likelihood Solution, Predictive distribution from MLE]

- For a Bayesian treatment, we need priors.

We use a multivariate Gaussian as a prior on the parameters $\omega$.

$$\omega_i \sim \mathcal{N}(0, \alpha^{-1}) \tag{3}$$

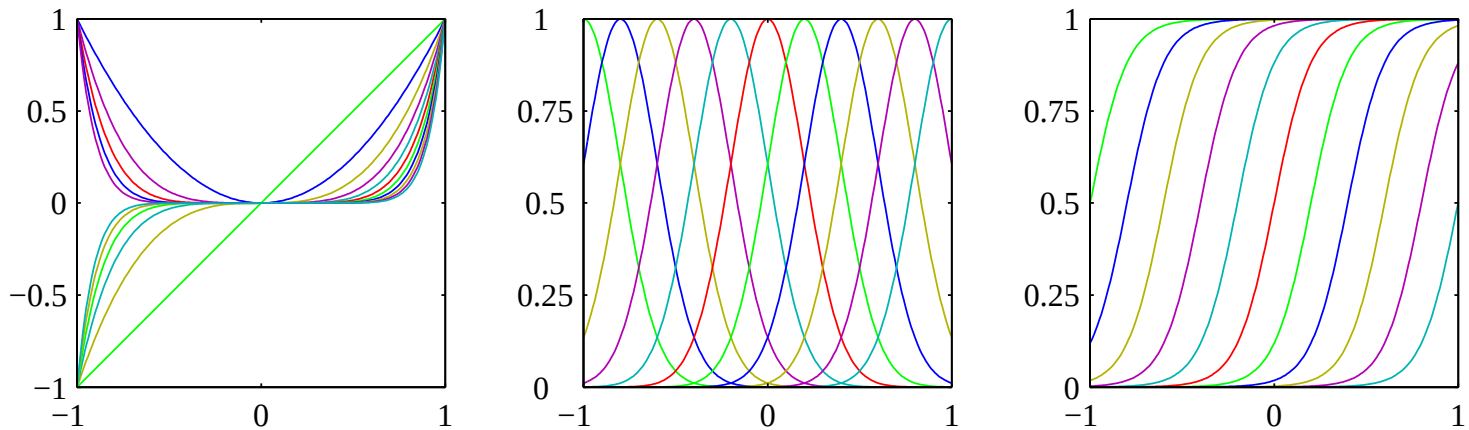
$$p(\omega|\alpha) = \prod_{i=1}^{M} \sqrt{\frac{\alpha}{2\pi}} \exp\left(-\frac{\alpha}{2}\omega_i^2\right) \tag{4}$$

$$= \left(\frac{\alpha}{2\pi}\right)^{M/2} \exp\left(-\frac{\alpha}{2}\omega^\top\omega\right) \tag{5}$$

- Finding the maximum-a-posteriori of $\omega$: [on board]

# If you are smart about choosing good basis functions ('features'), linear regression can get you pretty far.

- If we use nonlinear basis functions $\phi(x)$, can model nonlinear relationships with $y(\omega, \mathbf{x}) = \omega^\top \phi(x)$.
- Polynomial regression: $\phi(x) = (1, x, x^2, x^3)$
- 'Gaussian bumps': $\phi_i(x) = \exp\left((x - s_i)^2/\sigma_i^2\right)$
- Sigmoids $\phi_i(x) = 1/(1 + \exp(-x - s_i))$
- 'Kernel methods' are essentially linear algorithms which take one basis function per data-point.
- Predictive Mean [on board]

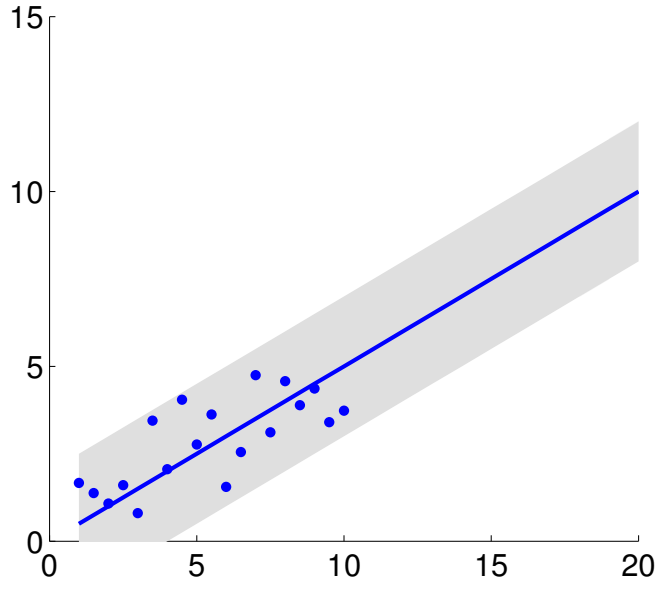


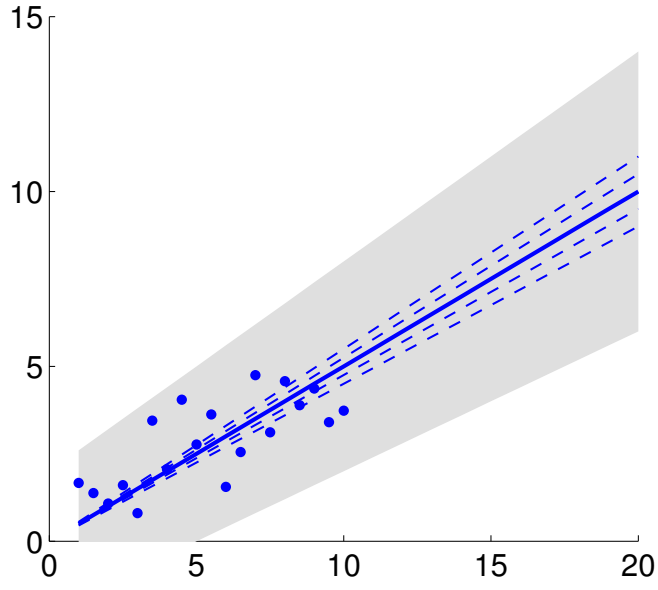

Bishop

PRML Figures 3.1a-c

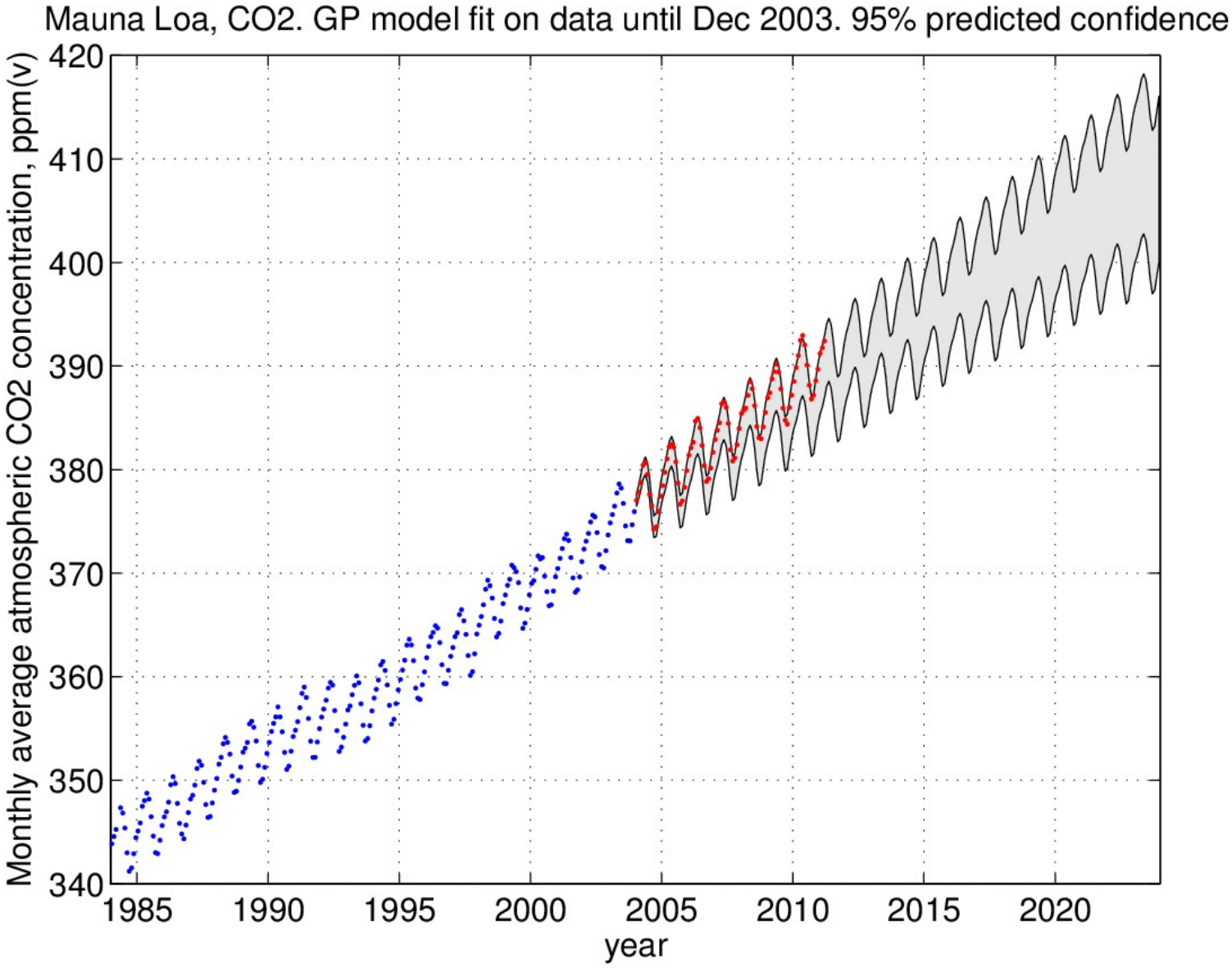
# Bayesian linear regression takes into account our uncertainty about parameters.

- Posterior distribution is Gaussian → Posterior mean and MAP coincide!
- However, neither the MLE nor the MAP solution take into account that we have (posterior) uncertainty about the parameters

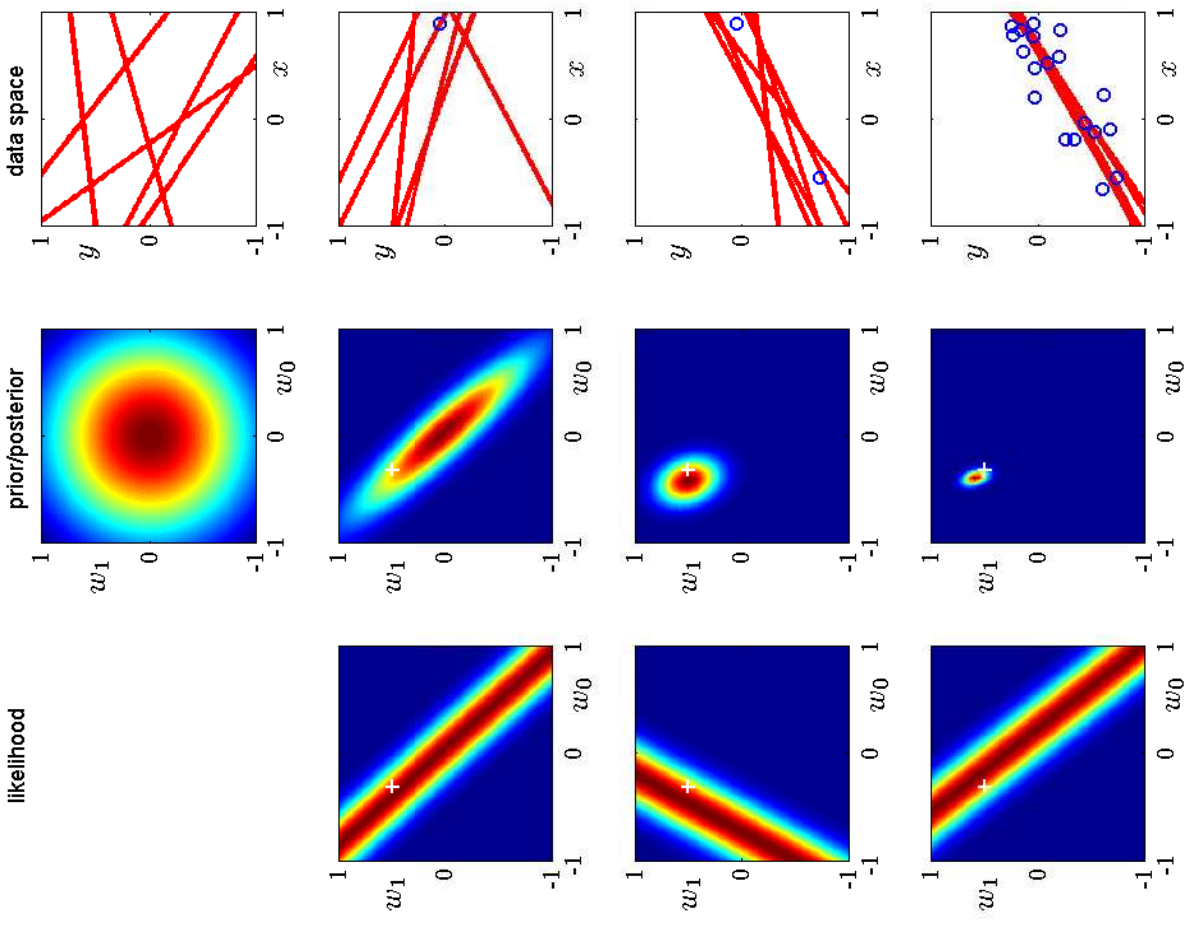


# Illustration: Climate prediction [by Carl Rasmussen, University of Cambridge, using Gaussian Processes]



# Bayesian regression illustrated: The more data we observe, the more constrained the parameters are.




Bishop PRML Figure 3.7

# Calculating the predictive mean and variance for Bayesian regression

Assume $\alpha$ and $\beta$ as given.
Posterior distribution: [derivation on board]

$$\Sigma_{post}^{-1} = \alpha \mathbf{I} + \beta \sum_i x_i x_i^\top \tag{6}$$

$$\mu_{post} = \Sigma_{post} \beta \sum_i x_i t_i \tag{7}$$

Predictive distribution: [derivation on board]

$$E(t^*|D, x^*) = \mu_{post}^\top x^* \tag{8}$$

$$\mathrm{Var}(t^*|D, x^*) = 1/\beta + x^{*\top} \Sigma_{post} x^* \tag{9}$$

What if there are basis functions? [on board]

# But, where do we get $\alpha$ and $\beta$ from?

- Bad news: Getting these makes things more complicated.
- Good news: This will not be on the exam (unless I take that back explicitly...).
- 'Full' Bayesian inference: Integrate out $\alpha$, $\beta$. No closed form solution. Use (e.g.) variational inference.
- Practical solution: optimize $\alpha$ and $\beta$ by maximizing the evidence, also known as marginal likelihood or likelihood type 2

$$E = \log P(\alpha, \beta | D) \tag{10}$$

$$= \log \int_{\omega} p(D|\omega, \beta) p(\omega|\alpha) p(\alpha, \beta) d\omega \tag{11}$$

$$= \frac{M}{2} \log(\alpha) + \frac{N}{2} \log \beta - \mathrm{M}(\mu_{post}) + \frac{1}{2} |\Sigma_{post}| - \frac{N}{2} \log(2\pi) \tag{12}$$

- $\mu_{post}$ and $\Sigma_{post}$ are the posterior mean and covariance, and $\mathrm{M}(\mu_{post}) = \frac{\beta}{2} \sum_n (t_n - y(\mu_{post}, \mathbf{x}_n))^2 + \frac{\alpha}{2} \mu_{post} \mu_{post}^{\top}$ is the quadratic cost function evalulated at the posterior mean (see Bishop 3.5 for details).

Optimization of the marginal likelihood is a Bayesian alternative to parameter-setting by cross-validation.

[on board]

# What we have not had time to cover:

- Non-Gaussian priors: The most important non-Gaussian prior is the 'Laplace prior' ('L1 regularization'), which leads to sparse MAP-solutions.
- Non-Gaussian noise models: If you know that your noise is not Gaussian but, say, Poisson, use 'generalized linear regression'. Some choices of noise models are more robust to outliers than the Gaussian. We will do one specific example of generalized linear regression in the last lecture.
- Nonlinear regression models: The most important Bayesian nonlinear regression technique is *Gaussian process regression*. In a nutshell, GP regression is like linear regressions but the algorithm puts ne basis function at each data-point. To understand GP regression, you need to understand Gaussians.