UNIVERSITÄT BONN

Juergen Gall

Regression

MA-INF 2213 - Advanced Computer Vision

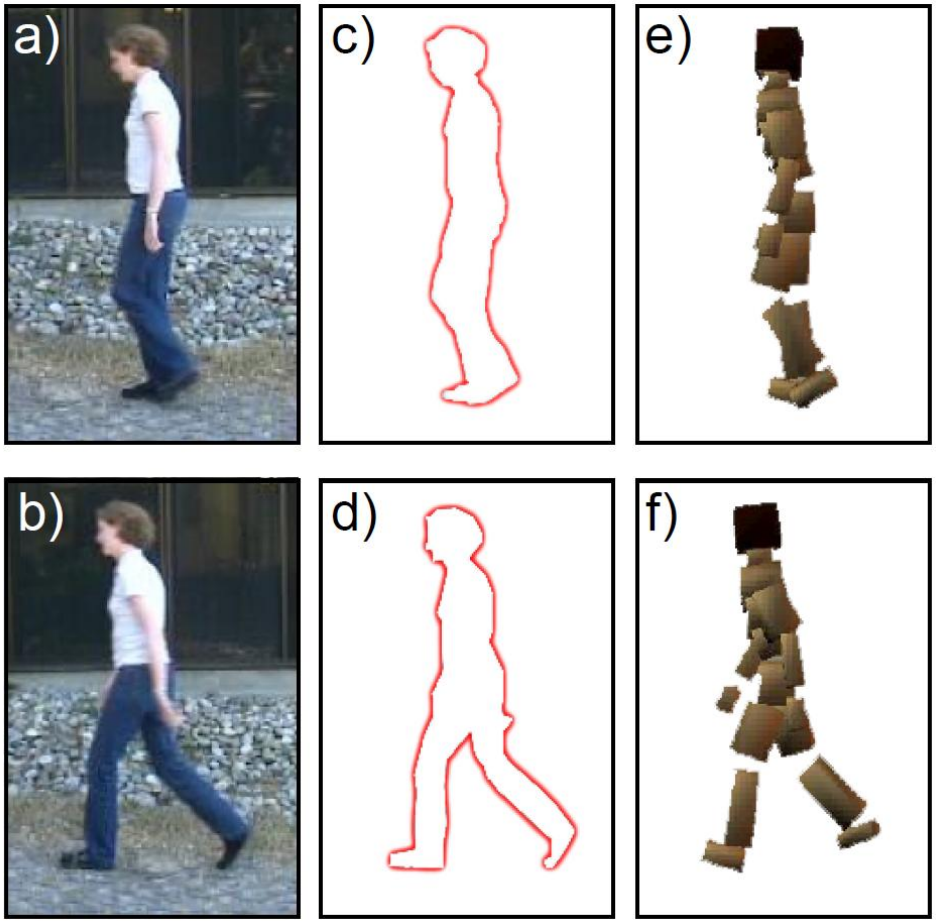
SS25

Literature

Chapter 8 Regression Models

S. Prince. **Computer Vision: Models, Learning, and Inference.** Cambridge University Press 2012

Body Pose Regression



Encode silhouette as 100x1 vector, encode body pose as 55 x1 vector. Learn relationship

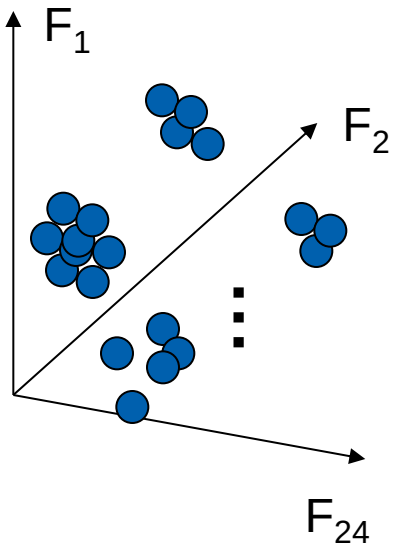
[A. Agarwal and B. Triggs. **3D Human Pose from Silhouettes by Relevance Vector Regression**. CVPR 2004]

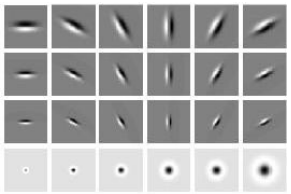
Simon J.D. Prince

Recall: Segmentation as clustering

Depending on what we choose as the *feature space*, we can group pixels in different ways.

Grouping pixels based on **texture** similarity





Filter bank of 24 filters

Feature space: filter bank responses (e.g., 24-d)

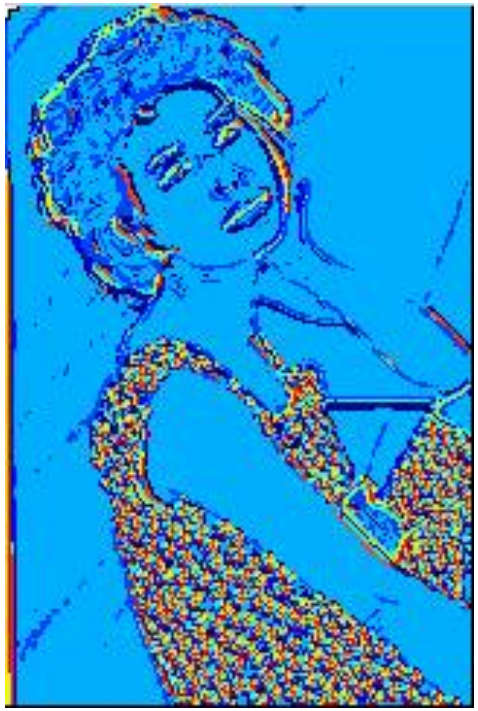
Source: K. Grauman

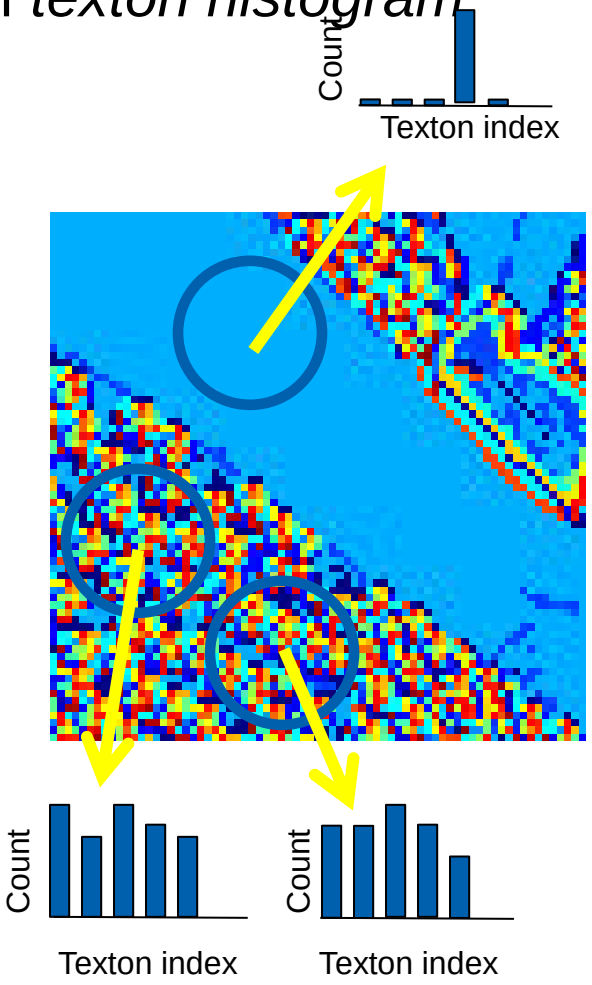
Recall: Segmentation with texture features

- Find “textons” by **clustering** vectors of filter bank outputs
- Describe texture in a window based on *texton histogram*

Image

Texton map

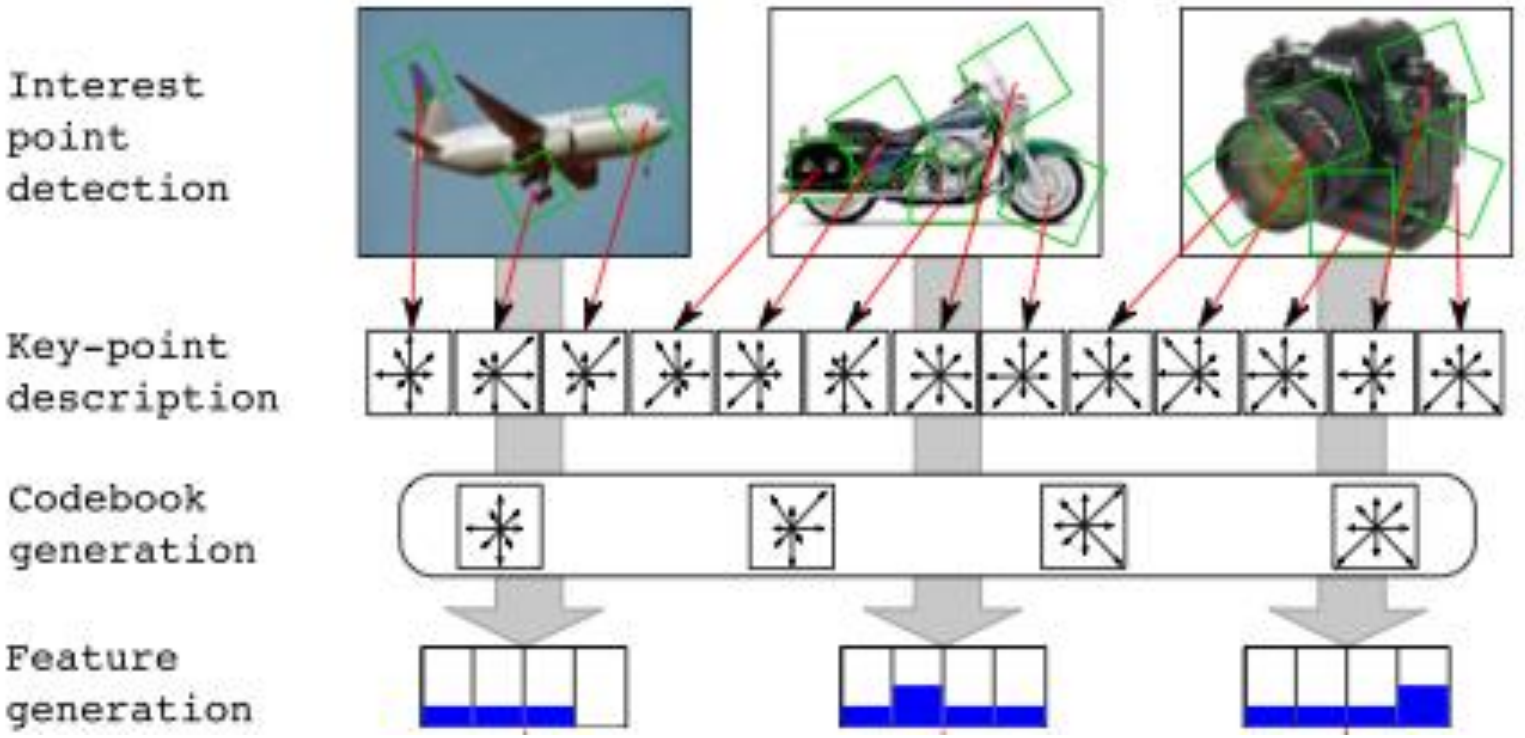




Source: S. Lazebnik

Codebook

- Codebook is used to reduce dimensionality and get a more compact representation
- Codebook can be obtained by any clustering method, e.g., K-means



Recall: Material classification example

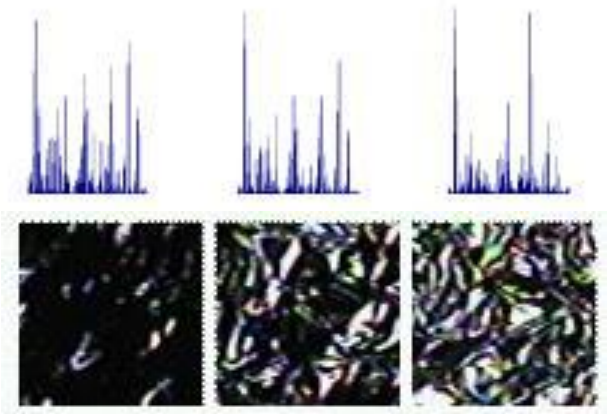
UNIVERSITÄT BONN

For an image of a single texture, we can classify it according to its global (image-wide) texton histogram.

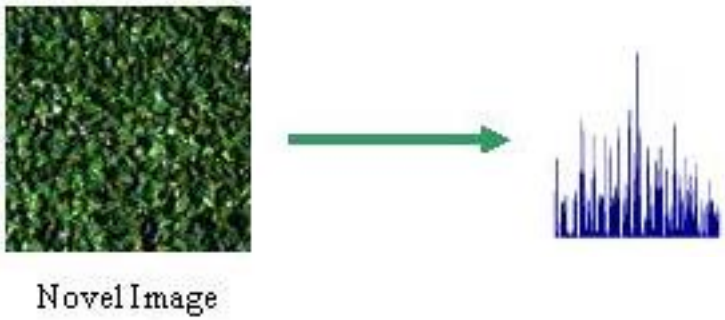
Source: M. Varma

Recall: Material classification example

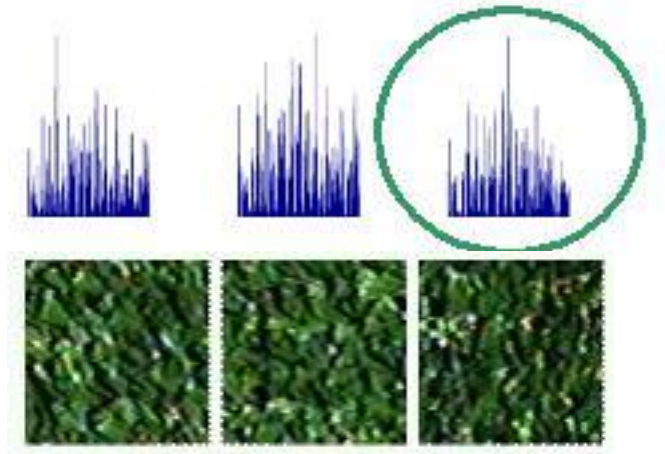
Nearest neighbor classification: label the input according to the nearest known example's label.



Foil



$$\chi^2(h_i, h_j) = \frac{1}{2}\sum_{k=1}^{K}\frac{[h_i(k) - h_j(k)]^2}{h_i(k) + h_j(k)}$$



Grass

Source: M. Varma

Tricks of the Trade: Normalization

UNIVERSITÄT BONN

We formed histograms as raw counts of feature points:

- Their number of entries can vary strongly:
  - Normalize each histogram by its L^1 or L^2 norm:

$$\hat{h}_j := \frac{1}{n_1} h_j \qquad \text{with } n_1 = \sum_j h_j$$

$$\hat{h}_j := \frac{1}{n_2} h_j \qquad \text{with } n_2 = \sqrt{\sum_j h_j^2}$$

  - Suppress strong peaks in the histogram by non-linear *preprocessing*, e.g.

$$\hat{h}_j := \sqrt{h_j}, \qquad \hat{h}_j := \sqrt[3]{h_j}, \qquad \hat{h}_j := \begin{cases} \theta_j & \text{for } h_j > \theta_j \\ h_j & \text{else.} \end{cases}$$

- No golden rule how to normalize/preprocess
  - rough hint: adjust normalization+kernel for scaling invariance.

Christoph Lampert

Kernels for Comparing Histograms

UNIVERSITÄT BONN

Many *kernels* have been suggest for histogram data:

- We can treat K-bin histograms h, h' as vectors in $\mathbb{R}^K$:
  - $k(h,h') = \sum_j h_j h'_j$ *linear kernel*
  - $k(h,h') = (c + \sum_j h_j h'_j)^d$ *polynomial kernel*
  - $k(h,h') = \exp\left(-\frac{1}{\gamma}\sum_j \|h_j - h'_j\|^2\right)$ *Gaussian kernel*

- If we normalize them (to sum 1), we can treat histograms as discrete probability distributions:
  - $k_{HI}(h,h') = \sum_j \min(h_j, h'_j)$.
  - $k_{bhattacharya}(h,h') = \sum_j \sqrt{h_j h'_j}$
  - $k_{symKL}(h,h') = \exp\left(-\frac{1}{2}(KL(h|h') + KL(h'|h))\right)$
    where $KL(h|h') = \sum_j h_j \log \frac{h_j}{h'_j}$
  - $k_{\chi^2}(h,h') = \exp\left(-\frac{1}{\gamma}\chi^2(h,h')\right)$ with $\chi^2(h,h') = \sum_j \frac{(h_j - h'_j)^2}{h_j + h'_j}$.

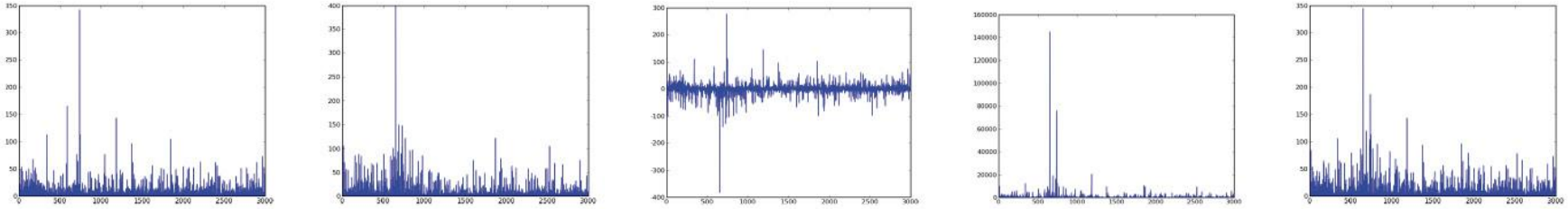
[M. Hein, O. Bousquet, "Hilbertian metrics and positive definite kernels on probability measures", AISTATS 2005]

Christoph Lampert

Robust Kernels

Especially k_{HI} and k_{χ^2} seem to work very well for computer vision:

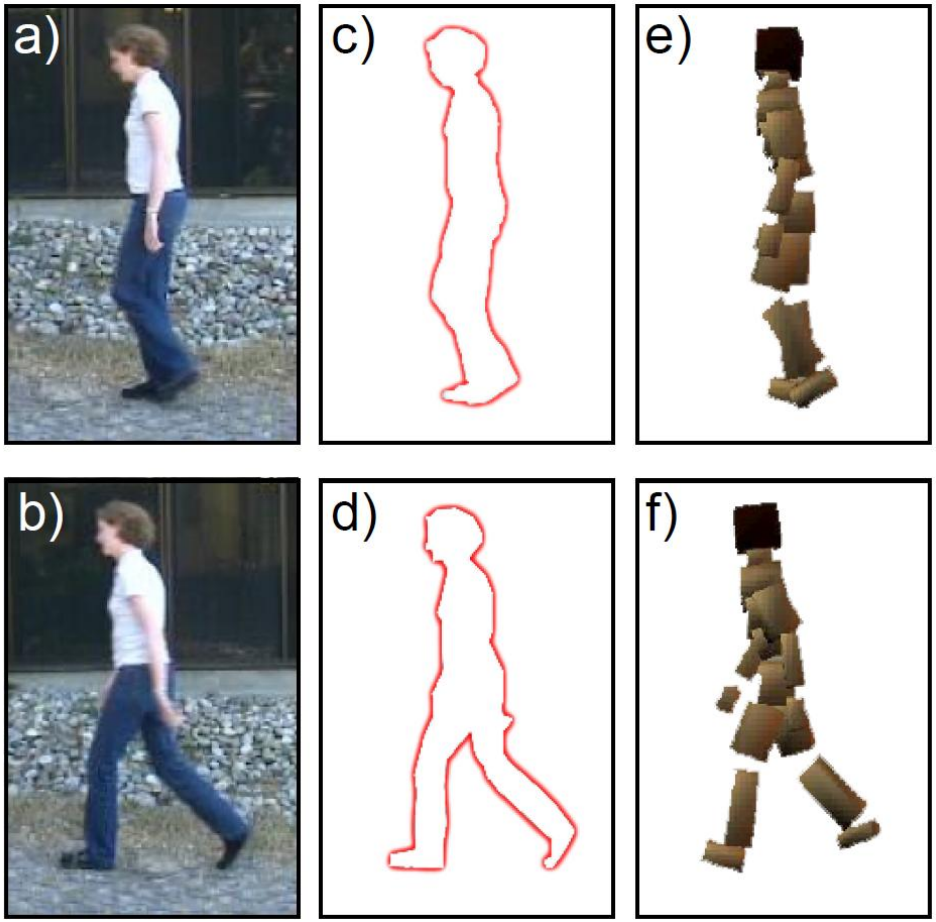
$$k_{HI}(h, h') = \sum_j \min(h_j, h'_j) \qquad k_{\chi^2}(h, h') = \exp\left(-\frac{1}{\gamma}\sum_j \frac{(h_j - h'_j)^2}{h_j + h'_j}\right).$$



h_j h'_j $h_j - h'_j$ $(h_j - h'_j)^2$ $\frac{(h_j - h'_j)^2}{h_j + h'_j}$

- Feature-histograms have few large and many small entries.
- Quadratic measures (L^2 or Gaussian kernel) concentrate on the largest differences: 3 bins (out of 3000) contribute 25%
- $1st$-order (HI or χ^2-kernel) consider bins more balanced: 3 largest terms contribute 3.5%

Body Pose Regression

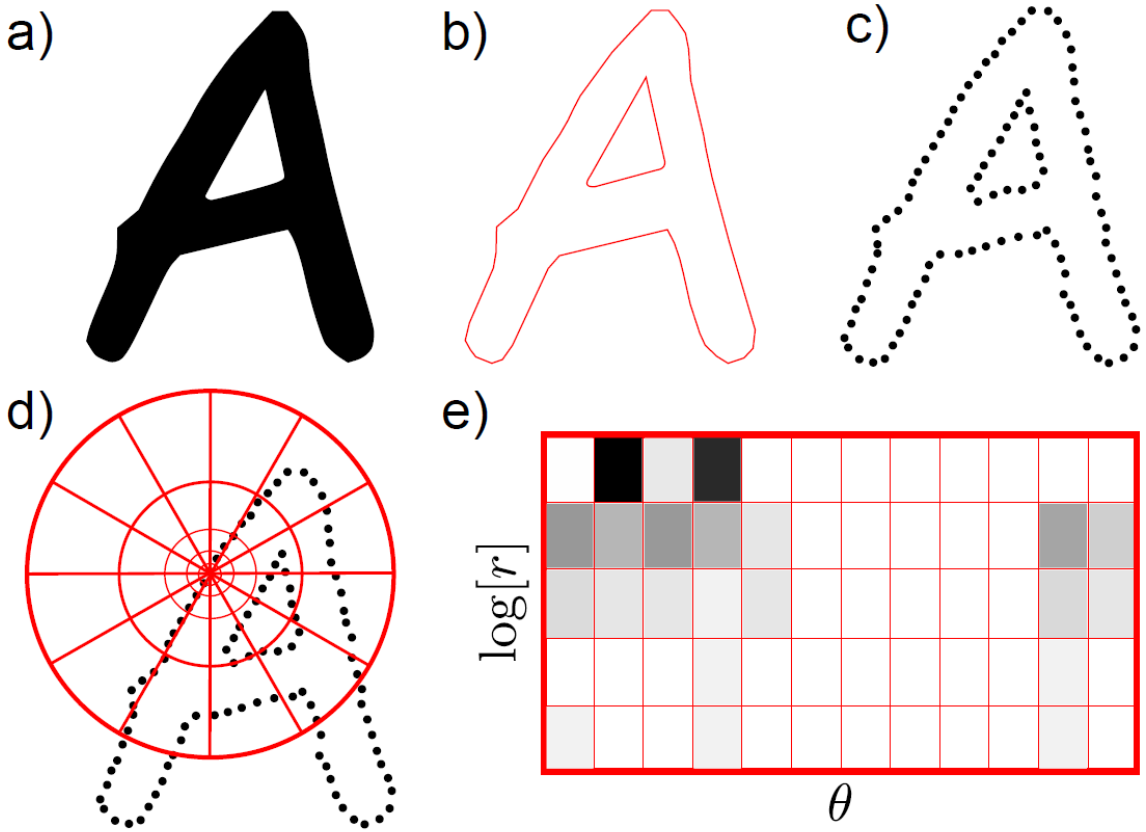


Encode silhouette as 100x1 vector, encode body pose as 55 x1 vector. Learn relationship

[A. Agarwal and B. Triggs. **3D Human Pose from Silhouettes by Relevance Vector Regression**. CVPR 2004]

Simon J.D. Prince

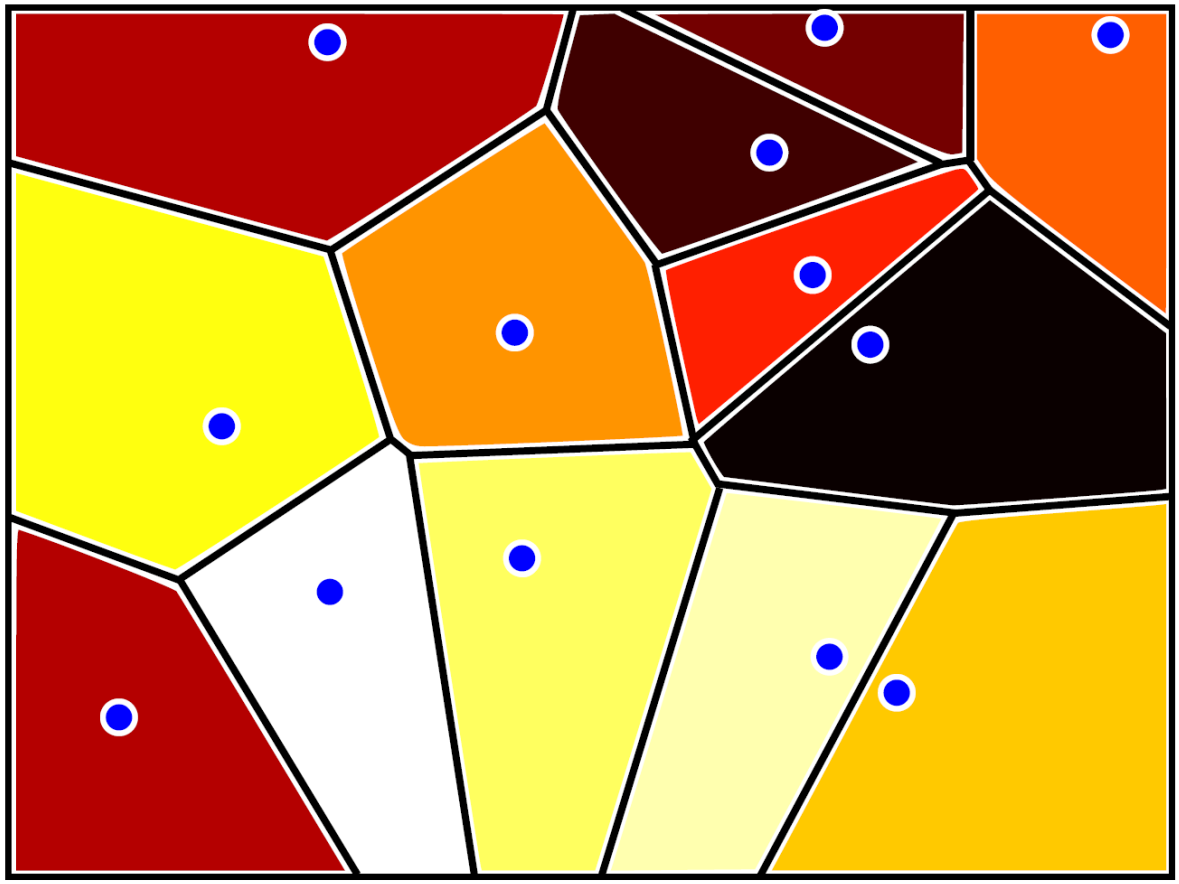
Shape Context



Returns 60 x 1 vector for each of 400 points around the silhouette

Simon J.D. Prince

Codebook



Cluster 60D space (based on all training data) into 100 vectors
Assign each 60x1 vector to closest cluster (Voronoi partition)
Final data vector is 100x1 histogram over distribution of assignments

Simon J.D. Prince

Learn relationship

Probabilistic: Given feature **x,** model pose **w** as probability distribution Pr(**w**|**x**)

How to model Pr(**w**|**x**)?

- Choose an appropriate form for prior Pr(**w**)
- Parameterize a function of type **w** = f(**x; θ)**

Learning algorithm: learn parameters θ from training data **x**,**w**

Inference algorithm: just evaluate **Pr(w|x)**

Linear Regression

- For simplicity we will assume that each dimension of world is predicted separately.
- Concentrate on predicting a univariate world state w.

Choose normal distribution over world w

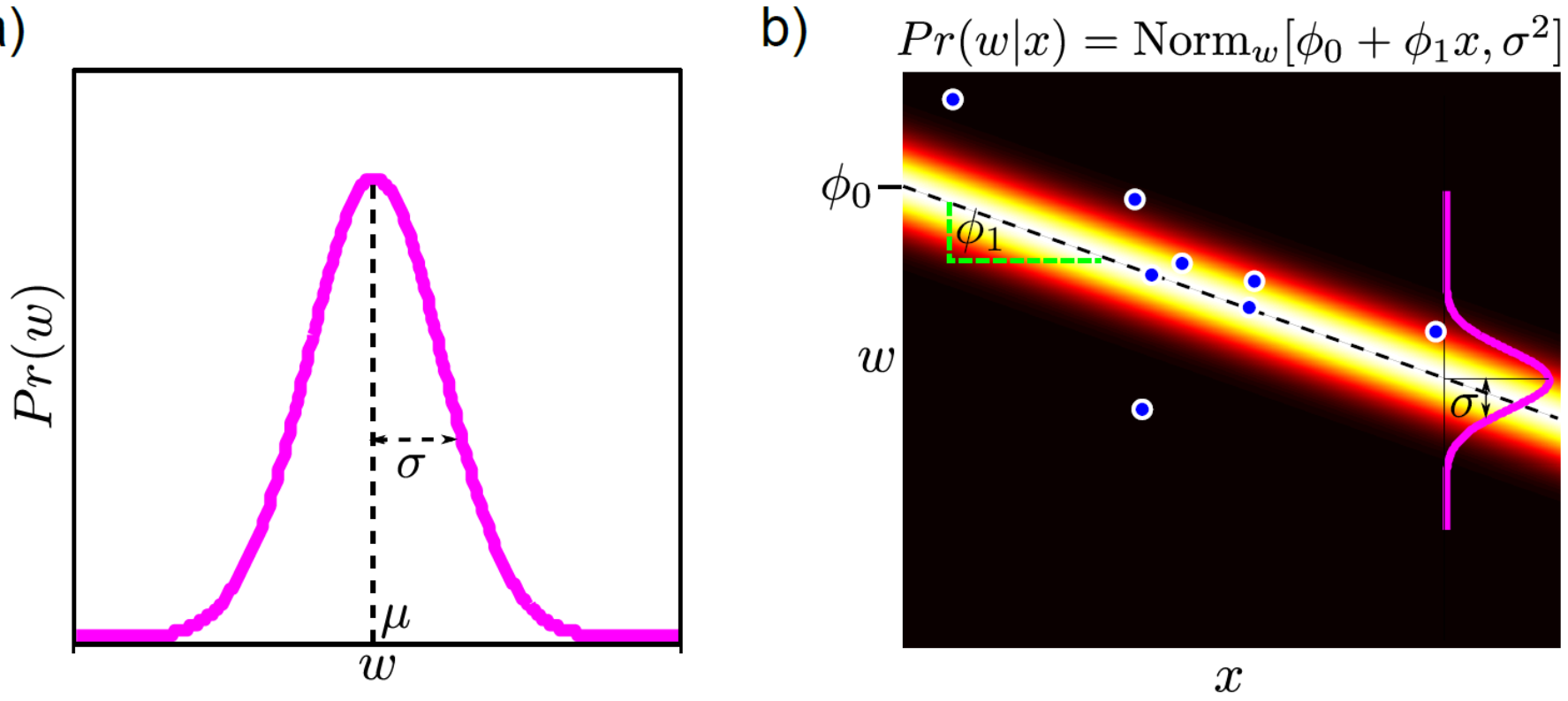
$$Pr(w_i|\mathbf{x}_i, \boldsymbol{\theta}) = \text{Norm}_{w_i}\left[\phi_0 + \boldsymbol{\phi}^T\mathbf{x}_i, \sigma^2\right]$$

Make

- Mean a linear function of data x
- Variance constant

Simon J.D. Prince

Linear Regression



$$Pr(w_i|\mathbf{x}_i, \boldsymbol{\theta}) = \mathrm{Norm}_{w_i}\left[\phi_0 + \boldsymbol{\phi}^T\mathbf{x}_i, \sigma^2\right]$$

Simon J.D. Prince

Neater Notation

$$Pr(w_i|\mathbf{x}_i, \boldsymbol{\theta}) = \text{Norm}_{w_i}\left[\phi_0 + \boldsymbol{\phi}^T\mathbf{x}_i, \sigma^2\right]$$

To make notation easier to handle, we

- Attach a 1 to the start of every data vector

$$\mathbf{x}_i \leftarrow \begin{bmatrix}1 & \mathbf{x}_i^T\end{bmatrix}^T$$

- Attach the offset to the start of the gradient vector $\boldsymbol{\phi}$

$$\boldsymbol{\phi} \leftarrow \begin{bmatrix}\phi_0 & \boldsymbol{\phi}^T\end{bmatrix}^T$$

New model:

$$Pr(w_i|\mathbf{x}_i, \boldsymbol{\theta}) = \text{Norm}_{w_i}\left[\boldsymbol{\phi}^T\mathbf{x}_i, \sigma^2\right]$$

Simon J.D. Prince

Combining Equations

We have one equation for each (x_i,w_i) training pair:

$$Pr(w_i|\mathbf{x}_i,\boldsymbol{\theta}) = \text{Norm}_{w_i}\left[\boldsymbol{\phi}^T\mathbf{x}_i, \sigma^2\right]$$

The likelihood of the whole dataset is the product of these individual distributions and can be written as

$$Pr(\mathbf{w}|\mathbf{X},\boldsymbol{\theta}) = \text{Norm}_{\mathbf{w}}[\mathbf{X}^T\boldsymbol{\phi}, \sigma^2\mathbf{I}]$$

Where for *I* training pairs

$$\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2 \ldots \mathbf{x}_I] \quad \mathbf{w} = [w_1, w_2, \ldots, w_I]^T$$

Simon J.D. Prince

Derivatives and Notation

UNIVERSITÄT BONN

$$\mathrm{Norm}_{\mathbf{x}}\left[\boldsymbol{\mu}, \boldsymbol{\Sigma}\right] = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-0.5(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

$$\frac{\partial \mathbf{x}^T\mathbf{a}}{\partial \mathbf{x}} = \mathbf{a}$$

$$\frac{\partial \mathbf{a}^T\mathbf{x}}{\partial \mathbf{x}} = \mathbf{a}$$

$$\frac{\partial \mathbf{a}^T\mathbf{X}\mathbf{b}}{\partial \mathbf{X}} = \mathbf{a}\mathbf{b}^T$$

$$\frac{\partial \mathbf{a}^T\mathbf{X}^T\mathbf{b}}{\partial \mathbf{X}} = \mathbf{b}\mathbf{a}^T$$

$$\frac{\partial \mathbf{b}^T\mathbf{X}^T\mathbf{X}\mathbf{c}}{\partial \mathbf{X}} = \mathbf{X}(\mathbf{b}\mathbf{c}^T + \mathbf{c}\mathbf{b}^T)$$

$$\frac{\partial (\mathbf{B}\mathbf{x}+\mathbf{b})^T\mathbf{C}(\mathbf{D}\mathbf{x}+\mathbf{d})}{\partial \mathbf{x}} = \mathbf{B}^T\mathbf{C}(\mathbf{D}\mathbf{x}+\mathbf{d}) + \mathbf{D}^T\mathbf{C}^T(\mathbf{B}\mathbf{x}+\mathbf{b})$$

$$\frac{\partial \mathbf{x}^T\mathbf{B}\mathbf{x}}{\partial \mathbf{x}} = (\mathbf{B}+\mathbf{B}^T)\mathbf{x}$$

$$\frac{\partial \mathbf{b}^T\mathbf{X}^T\mathbf{D}\mathbf{X}\mathbf{c}}{\partial \mathbf{X}} = \mathbf{D}^T\mathbf{X}\mathbf{b}\mathbf{c}^T + \mathbf{D}\mathbf{X}\mathbf{c}\mathbf{b}^T$$

$$\frac{\partial (\mathbf{X}\mathbf{b}+\mathbf{c})^T\mathbf{D}(\mathbf{X}\mathbf{b}+\mathbf{c})}{\partial \mathbf{X}} = (\mathbf{D}+\mathbf{D}^T)(\mathbf{X}\mathbf{b}+\mathbf{c})\mathbf{b}^T.$$

Combining Equations

The likelihood of the whole dataset is the product of these individual distributions and can be written as

$$Pr(\mathbf{w}|\mathbf{X},\boldsymbol{\theta}) = \text{Norm}_{\mathbf{w}}[\mathbf{X}^T\boldsymbol{\phi}, \sigma^2\mathbf{I}]$$

$$\text{Norm}_{\mathbf{w}}[\mathbf{X}^T\phi, \sigma^2\mathbf{I}] = \frac{1}{(2\pi)^{I/2}(\sigma^2)^{I/2}} \exp\left[-\frac{(\mathbf{w}-\mathbf{X}^T\phi)^T(\mathbf{w}-\mathbf{X}^T\phi)}{2\sigma^2}\right]$$

$$\log \text{Norm}_{\mathbf{w}}[\mathbf{X}^T\phi, \sigma^2\mathbf{I}] = \log[(2\pi)^{-I/2}] + \log[(\sigma^2)^{-I/2}] + \log\exp\left[-\frac{(\mathbf{w}-\mathbf{X}^T\phi)^T(\mathbf{w}-\mathbf{X}^T\phi)}{2\sigma^2}\right]$$

$$= -\frac{I\log[(2\pi)]}{2} - \frac{I\log[\sigma^2]}{2} - \frac{(\mathbf{w}-\mathbf{X}^T\phi)^T(\mathbf{w}-\mathbf{X}^T\phi)}{2\sigma^2}$$

Simon J.D. Prince

Learning

UNIVERSITÄT BONN

Maximum likelihood

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[Pr(\mathbf{w}|\mathbf{X}, \boldsymbol{\theta}) \right] = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[\log Pr(\mathbf{w}|\mathbf{X}, \boldsymbol{\theta}) \right]$$

Substituting in

$$\hat{\boldsymbol{\phi}}, \hat{\sigma}^2 = \underset{\boldsymbol{\phi}, \sigma^2}{\operatorname{argmax}} \left[-\frac{I \log[2\pi]}{2} - \frac{I \log[\sigma^2]}{2} - \frac{(\mathbf{w} - \mathbf{X}^T\boldsymbol{\phi})^T(\mathbf{w} - \mathbf{X}^T\boldsymbol{\phi})}{2\sigma^2} \right]$$

Take derivative, set result to zero and re-arrange:

$$\begin{aligned} \hat{\boldsymbol{\phi}} &= (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{w} \\ \hat{\sigma}^2 &= \frac{(\mathbf{w} - \mathbf{X}^T\boldsymbol{\phi})^T(\mathbf{w} - \mathbf{X}^T\boldsymbol{\phi})}{I} \end{aligned}$$

Simon J.D. Prince

Learning

$$\hat{\boldsymbol{\phi}}, \hat{\sigma}^2 = \underset{\boldsymbol{\phi},\sigma^2}{\operatorname{argmax}} \left[-\frac{I\log[2\pi]}{2} - \frac{I\log[\sigma^2]}{2} - \frac{(\mathbf{w}-\mathbf{X}^T\boldsymbol{\phi})^T(\mathbf{w}-\mathbf{X}^T\boldsymbol{\phi})}{2\sigma^2} \right]$$

$$\frac{\partial(\mathbf{Bx}+\mathbf{b})^T\mathbf{C}(\mathbf{Dx}+\mathbf{d})}{\partial \mathbf{x}} = \mathbf{B}^T\mathbf{C}(\mathbf{Dx}+\mathbf{d}) + \mathbf{D}^T\mathbf{C}^T(\mathbf{Bx}+\mathbf{b})$$

Take derivative, set result to zero:

$$-\frac{1}{2\sigma^2}\frac{\partial}{\partial\phi}(w-X^T\phi)^T(w-X^T\phi) = \frac{1}{\sigma^2}X(w-X^T\phi) = 0$$

Re-arrange:

$$Xw = XX^T\phi$$

$$\phi = (XX^T)^{-1}Xw$$

Learning

$$\hat{\phi}, \hat{\sigma}^2 = \underset{\phi,\sigma^2}{\operatorname{argmax}} \left[-\frac{I \log[2\pi]}{2} - \frac{I \log[\sigma^2]}{2} - \frac{(\mathbf{w} - \mathbf{X}^T\phi)^T(\mathbf{w} - \mathbf{X}^T\phi)}{2\sigma^2} \right]$$

Take derivative, set result to zero and re-arrange:

$$-\frac{I}{2}\frac{\partial}{\partial\sigma^2}\log(\sigma^2) - \frac{(w - X^T\phi)^T(w - X^T\phi)}{2}\frac{\partial}{\partial\sigma^2}\frac{1}{\sigma^2}$$

$$= -\frac{I}{2\sigma^2} + \frac{(w - X^T\phi)^T(w - X^T\phi)}{2\sigma^4} = 0$$

Re-arrange:

$$\frac{(w - X^T\phi)^T(w - X^T\phi)}{\sigma^2} = I$$

$$\sigma^2 = \frac{(w - X^T\phi)^T(w - X^T\phi)}{I}$$

Learning

UNIVERSITÄT BONN

Maximum likelihood

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[Pr(\mathbf{w}|\mathbf{X}, \boldsymbol{\theta}) \right] = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \left[\log Pr(\mathbf{w}|\mathbf{X}, \boldsymbol{\theta}) \right]$$

Substituting in

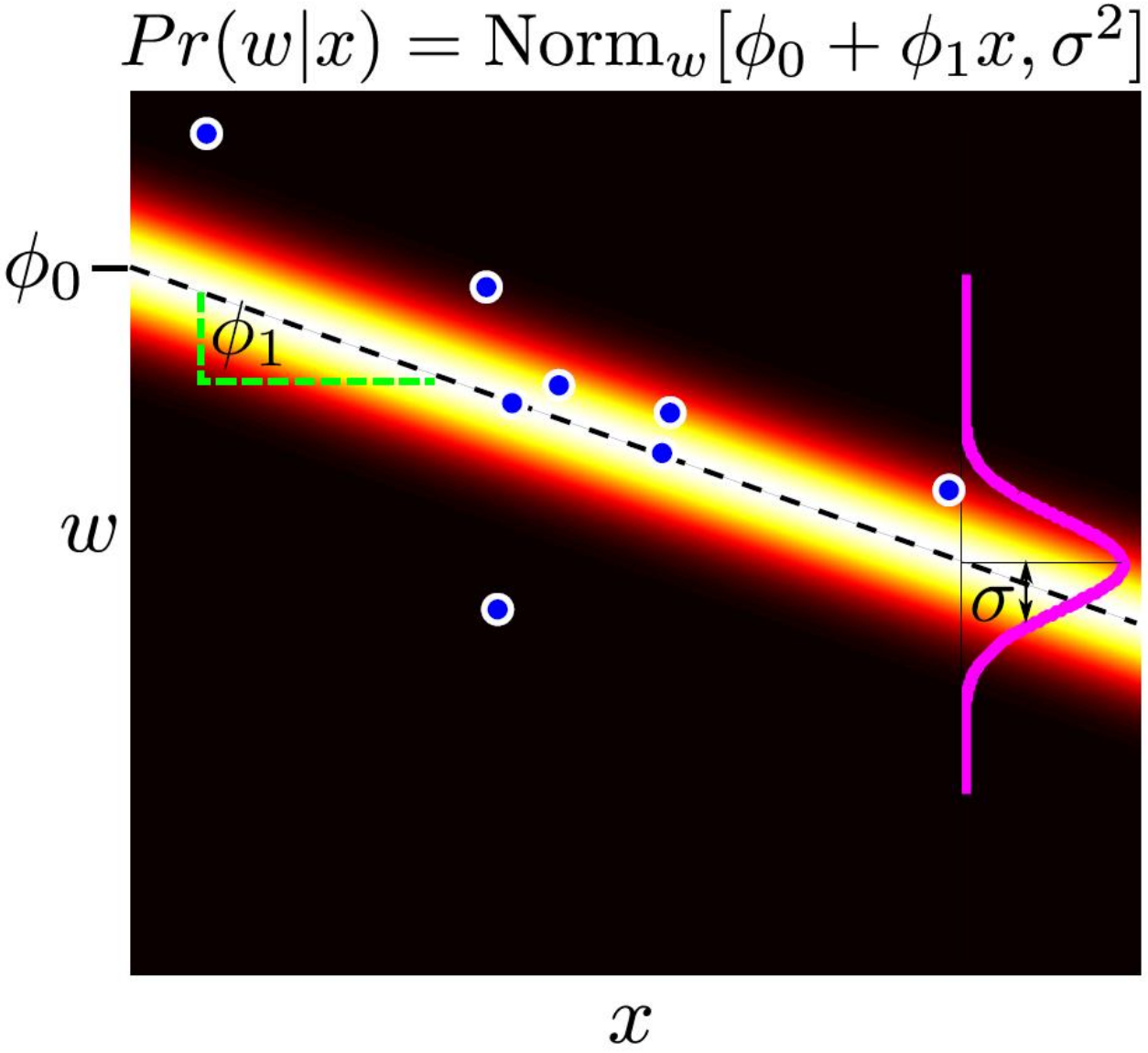
$$\hat{\boldsymbol{\phi}}, \hat{\sigma}^2 = \underset{\boldsymbol{\phi}, \sigma^2}{\operatorname{argmax}} \left[-\frac{I \log[2\pi]}{2} - \frac{I \log[\sigma^2]}{2} - \frac{(\mathbf{w} - \mathbf{X}^T\boldsymbol{\phi})^T(\mathbf{w} - \mathbf{X}^T\boldsymbol{\phi})}{2\sigma^2} \right]$$

Take derivative, set result to zero and re-arrange:

$$\begin{aligned} \hat{\boldsymbol{\phi}} &= (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{w} \\ \hat{\sigma}^2 &= \frac{(\mathbf{w} - \mathbf{X}^T\boldsymbol{\phi})^T(\mathbf{w} - \mathbf{X}^T\boldsymbol{\phi})}{I} \end{aligned}$$

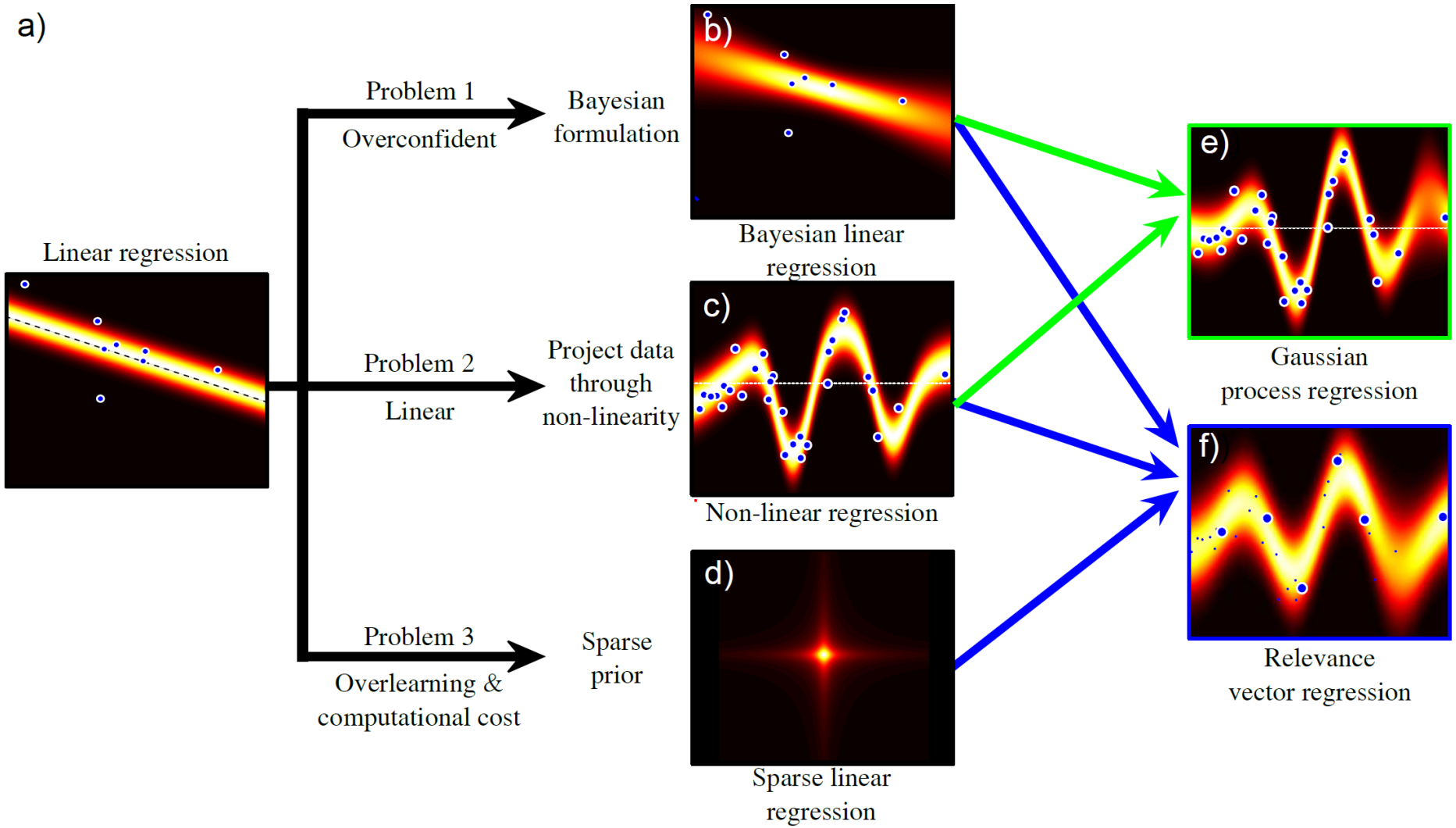
Simon J.D. Prince

Linear Regression



Simon J.D. Prince

Regression Models



Simon J.D. Prince

Bayesian Regression

(We concentrate on $\boldsymbol{\phi}$ – come back to σ^2 later!)

Likelihood

$$Pr(\mathbf{w}|\mathbf{X}, \boldsymbol{\phi}) = \text{Norm}_{\mathbf{w}}[\mathbf{X}^T\boldsymbol{\phi}, \sigma^2\mathbf{I}]$$

Prior

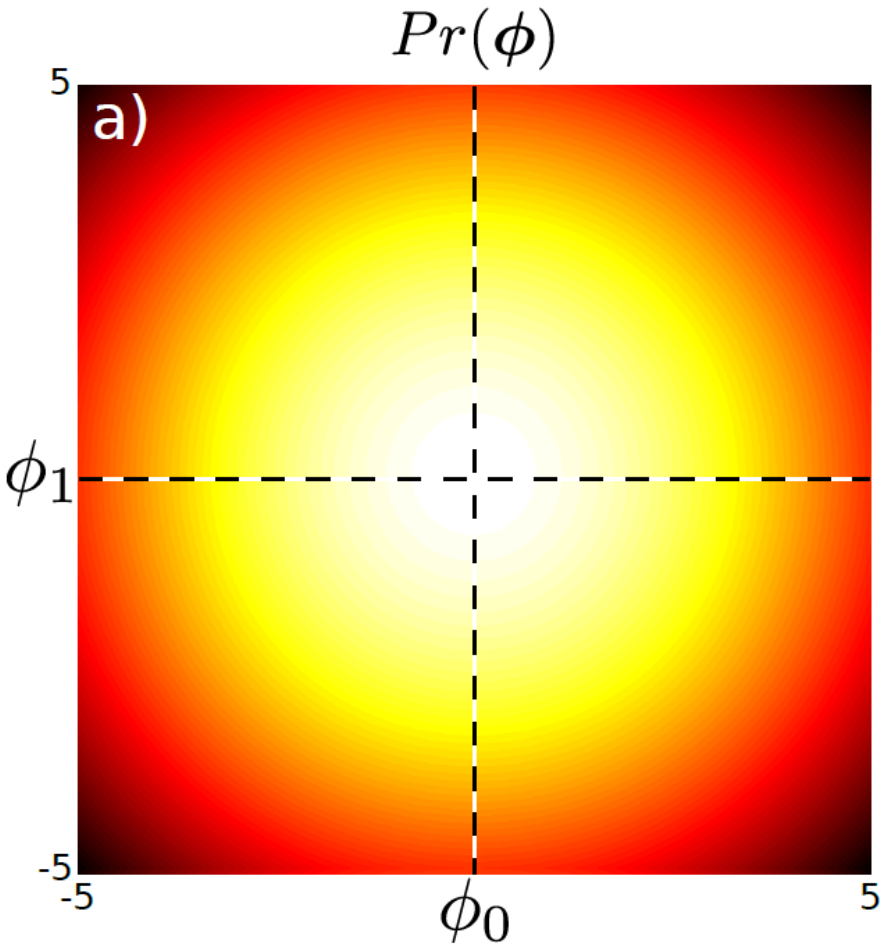
$$Pr(\boldsymbol{\phi}) = \text{Norm}_{\boldsymbol{\phi}}[\mathbf{0}, \sigma_p^2\mathbf{I}]$$

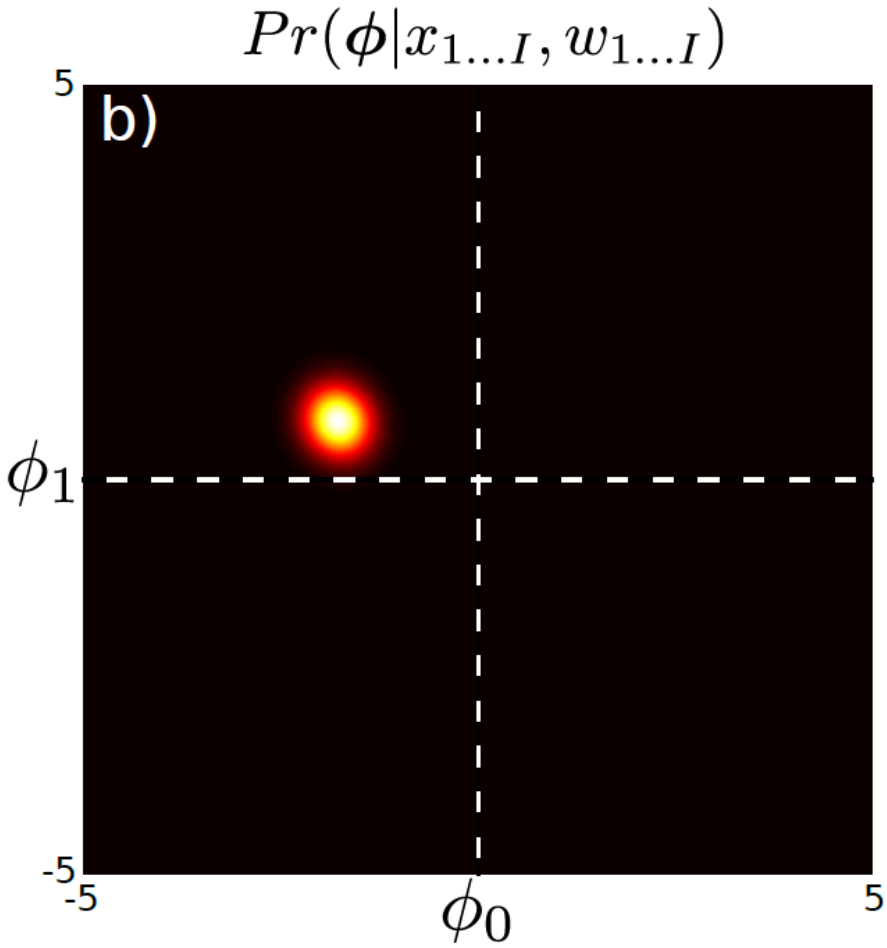
Bayes' rule

$$Pr(\boldsymbol{\phi}|\mathbf{X}, \mathbf{w}) = \frac{Pr(\mathbf{w}|\mathbf{X}, \boldsymbol{\phi})Pr(\boldsymbol{\phi})}{Pr(\mathbf{w}|\mathbf{X})}$$

Simon J.D. Prince

Bayesian Regression (Training)





$$Pr(\phi) = \text{Norm}_{\phi}[\mathbf{0}, \sigma_p^2\mathbf{I}]$$

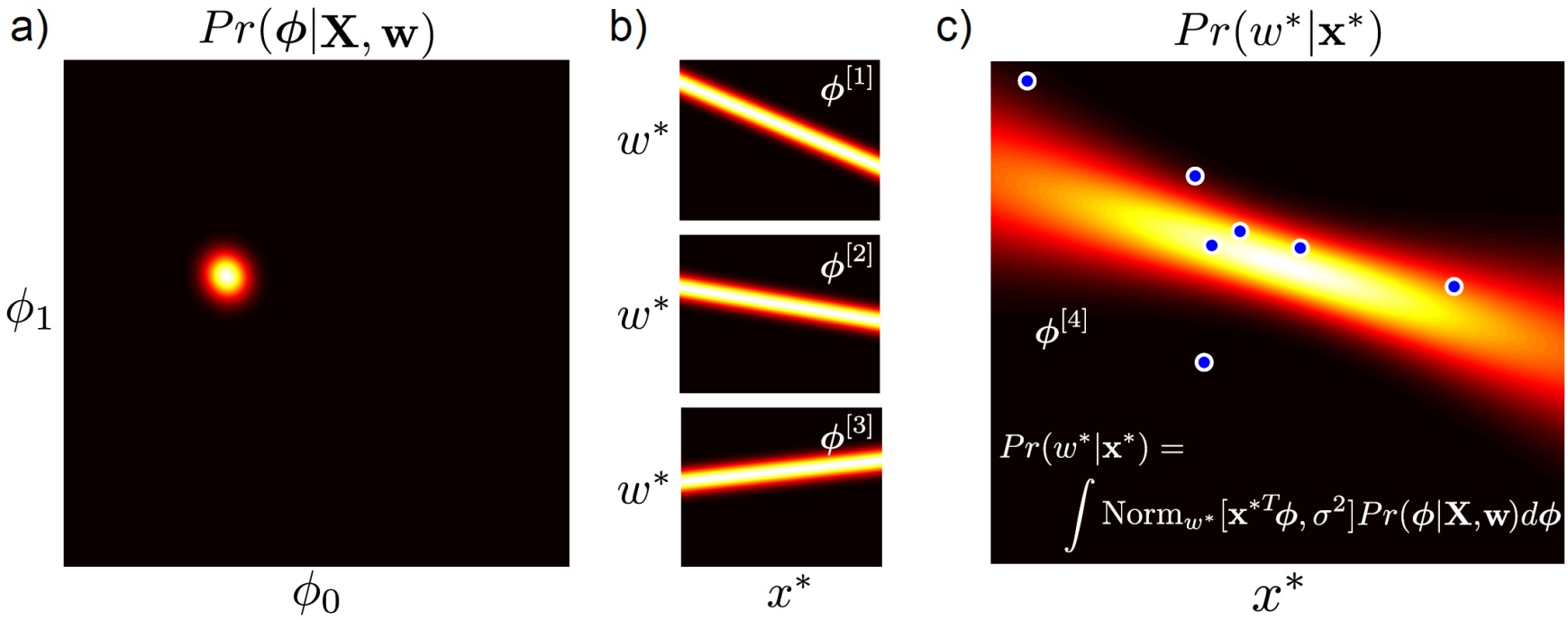
$$Pr(\phi|\mathbf{X}, \mathbf{w}) = \frac{Pr(\mathbf{w}|\mathbf{X}, \phi)Pr(\phi)}{Pr(\mathbf{w}|\mathbf{X})}$$

Simon J.D. Prince

Bayesian Regression (Inference)

Given a new observed vector **x***, compute:

$$Pr(w^*|\mathbf{x}^*, \mathbf{X}, \mathbf{w}) = \int Pr(w^*|\mathbf{x}^*, \boldsymbol{\phi}) Pr(\boldsymbol{\phi}|\mathbf{X}, \mathbf{w}) d\boldsymbol{\phi}$$



Simon J.D. Prince

Bayesian Regression

(We concentrate on $\boldsymbol{\phi}$ – come back to σ^2 later!)

Likelihood

$$Pr(\mathbf{w}|\mathbf{X}, \boldsymbol{\phi}) = \mathrm{Norm}_{\mathbf{w}}[\mathbf{X}^T\boldsymbol{\phi}, \sigma^2\mathbf{I}]$$

Prior

$$Pr(\boldsymbol{\phi}) = \mathrm{Norm}_{\boldsymbol{\phi}}[\mathbf{0}, \sigma_p^2\mathbf{I}]$$

Bayes' rule

$$Pr(\boldsymbol{\phi}|\mathbf{X}, \mathbf{w}) = \frac{Pr(\mathbf{w}|\mathbf{X}, \boldsymbol{\phi})Pr(\boldsymbol{\phi})}{Pr(\mathbf{w}|\mathbf{X})}$$

Simon J.D. Prince

Bayes' rule

Change variable for likelihood

$$Pr(\mathbf{w}|\mathbf{X},\boldsymbol{\phi}) = \mathrm{Norm}_{\mathbf{w}}[\mathbf{X}^T\boldsymbol{\phi}, \sigma^2\mathbf{I}]$$

using

$$\mathrm{Norm}_{\mathbf{x}}[\mathbf{A}\mathbf{y}+\mathbf{b},\boldsymbol{\Sigma}] = \kappa \cdot \mathrm{Norm}_{\mathbf{y}}[\mathbf{A}'\mathbf{x}+\mathbf{b}',\boldsymbol{\Sigma}'],$$

$$\begin{aligned}\boldsymbol{\Sigma}' &= (\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{A})^{-1}\\ \mathbf{A}' &= (\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{A})^{-1}\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\\ \mathbf{b}' &= -(\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{A})^{-1}\mathbf{A}^T\boldsymbol{\Sigma}^{-1}\mathbf{b}.\end{aligned}$$

gives

$$\mathrm{Norm}_{\mathbf{w}}[\mathbf{X}^T\phi, \sigma^2\mathbf{I}] = \mathrm{Norm}_{\phi}\left[\left(\frac{1}{\sigma^2}\mathbf{X}\mathbf{X}^T\right)^{-1}\frac{1}{\sigma^2}\mathbf{X}\mathbf{w}, \left(\frac{1}{\sigma^2}\mathbf{X}\mathbf{X}^T\right)^{-1}\right]$$

Bayes' rule

Using

$$\mathrm{Norm}_{\mathbf{x}}[\mathbf{a},\mathbf{A}]\mathrm{Norm}_{\mathbf{x}}[\mathbf{b},\mathbf{B}] = \kappa \cdot \mathrm{Norm}_{\mathbf{x}}\left[\left(\mathbf{A}^{-1}+\mathbf{B}^{-1}\right)^{-1}\left(\mathbf{A}^{-1}\mathbf{a}+\mathbf{B}^{-1}\mathbf{b}\right),\left(\mathbf{A}^{-1}+\mathbf{B}^{-1}\right)^{-1}\right]$$

gives

$$\mathrm{Norm}_{\phi}\left[\left(\frac{1}{\sigma^2}\mathbf{X}\mathbf{X}^T\right)^{-1}\frac{1}{\sigma^2}\mathbf{X}\mathbf{w},\left(\frac{1}{\sigma^2}\mathbf{X}\mathbf{X}^T\right)^{-1}\right]\mathrm{Norm}_{\phi}\left[\mathbf{0},\sigma_p^2\mathbf{I}\right] =$$

$$\mathrm{Norm}_{\phi}\left[\left(\frac{1}{\sigma^2}\mathbf{X}\mathbf{X}^T+\frac{1}{\sigma_p^2}\mathbf{I}\right)^{-1}\left(\frac{1}{\sigma^2}\mathbf{X}\mathbf{X}^T\left(\frac{1}{\sigma^2}\mathbf{X}\mathbf{X}^T\right)^{-1}\frac{1}{\sigma^2}\mathbf{X}\mathbf{w}\right),\left(\frac{1}{\sigma^2}\mathbf{X}\mathbf{X}^T+\frac{1}{\sigma_p^2}\mathbf{I}\right)^{-1}\right] =$$

$$\mathrm{Norm}_{\phi}\left[\frac{1}{\sigma^2}\left(\frac{1}{\sigma^2}\mathbf{X}\mathbf{X}^T+\frac{1}{\sigma_p^2}\mathbf{I}\right)^{-1}\mathbf{X}\mathbf{w},\left(\frac{1}{\sigma^2}\mathbf{X}\mathbf{X}^T+\frac{1}{\sigma_p^2}\mathbf{I}\right)^{-1}\right]$$

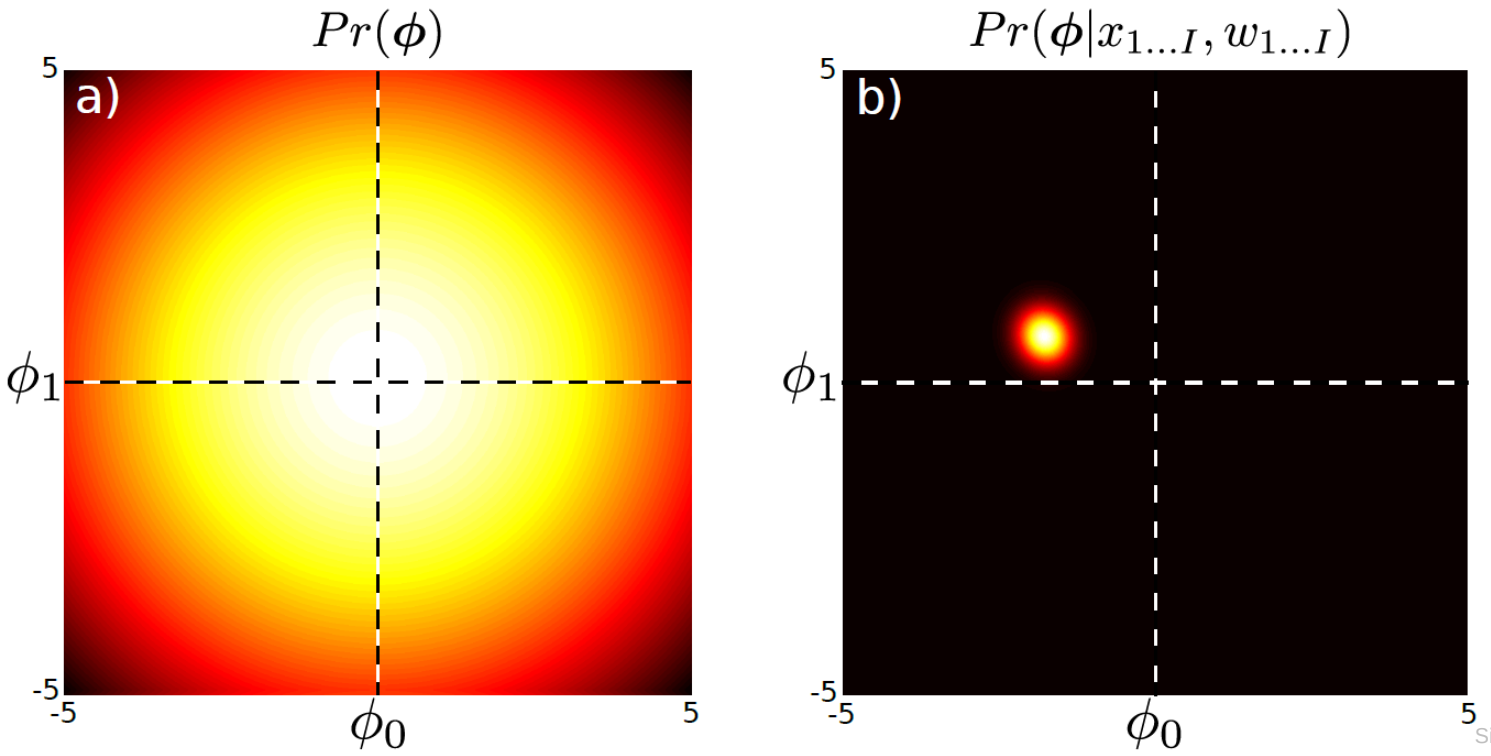
$$Pr(\boldsymbol{\phi}|\mathbf{X},\mathbf{w}) = \mathrm{Norm}_{\boldsymbol{\phi}}\left[\frac{1}{\sigma^2}\mathbf{A}^{-1}\mathbf{X}\mathbf{w},\mathbf{A}^{-1}\right] \quad \mathbf{A} = \frac{1}{\sigma^2}\mathbf{X}\mathbf{X}^T+\frac{1}{\sigma_p^2}\mathbf{I}$$

Posterior Dist. over Parameters

$$Pr(\boldsymbol{\phi}|\mathbf{X},\mathbf{w}) = \text{Norm}_{\boldsymbol{\phi}}\left[\frac{1}{\sigma^2}\mathbf{A}^{-1}\mathbf{X}\mathbf{w}, \mathbf{A}^{-1}\right]$$

where

$$\mathbf{A} = \frac{1}{\sigma^2}\mathbf{X}\mathbf{X}^T + \frac{1}{\sigma_p^2}\mathbf{I}$$



Simon J.D. Prince

Inference

Given a new observed vector **x***, compute Pr(*w**|**x***,**X**,**w**)

Using

$$Pr(\boldsymbol{\phi}|\mathbf{X},\mathbf{w}) = \text{Norm}_{\boldsymbol{\phi}}\left[\frac{1}{\sigma^2}\mathbf{A}^{-1}\mathbf{X}\mathbf{w}, \mathbf{A}^{-1}\right]$$

$$Pr(w^*|x^*,\phi) = \text{Norm}_{w^*}[\phi^T x^*, \sigma^2]$$

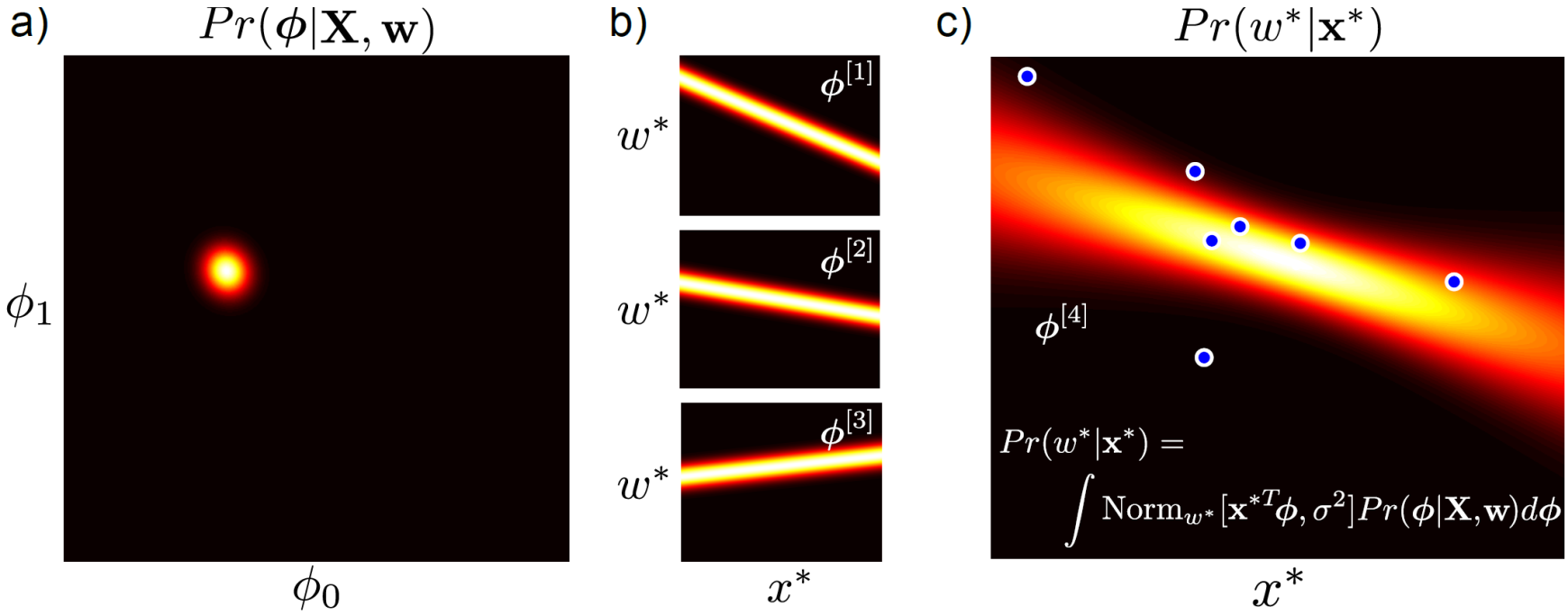
We get

$$\begin{aligned} Pr(w^*|\mathbf{x}^*,\mathbf{X},\mathbf{w}) &= \int Pr(w^*|\mathbf{x}^*,\boldsymbol{\phi})Pr(\boldsymbol{\phi}|\mathbf{X},\mathbf{w})d\boldsymbol{\phi} \\ &= \int \text{Norm}_{w^*}[\boldsymbol{\phi}^T\mathbf{x}^*, \sigma^2]\text{Norm}_{\boldsymbol{\phi}}\left[\frac{1}{\sigma^2}\mathbf{A}^{-1}\mathbf{X}\mathbf{w}, \mathbf{A}^{-1}\right] d\boldsymbol{\phi} \\ &= \text{Norm}_{w^*}\left[\frac{1}{\sigma^2}\mathbf{x}^{*T}\mathbf{A}^{-1}\mathbf{X}\mathbf{w}, \mathbf{x}^{*T}\mathbf{A}^{-1}\mathbf{x}^* + \sigma^2\right]. \end{aligned}$$

Simon J.D. Prince

Inference

$$
\begin{aligned}
Pr(w^*|\mathbf{x}^*, \mathbf{X}, \mathbf{w}) &= \int Pr(w^*|\mathbf{x}^*, \boldsymbol{\phi}) Pr(\boldsymbol{\phi}|\mathbf{X}, \mathbf{w}) d\boldsymbol{\phi} \\
&= \int \text{Norm}_{w^*}[\boldsymbol{\phi}^T \mathbf{x}^*, \sigma^2] \text{Norm}_{\boldsymbol{\phi}}\left[\frac{1}{\sigma^2}\mathbf{A}^{-1}\mathbf{X}\mathbf{w}, \mathbf{A}^{-1}\right] d\boldsymbol{\phi} \\
&= \text{Norm}_{w^*}\left[\frac{1}{\sigma^2}\mathbf{x}^{*T}\mathbf{A}^{-1}\mathbf{X}\mathbf{w}, \mathbf{x}^{*T}\mathbf{A}^{-1}\mathbf{x}^* + \sigma^2\right].
\end{aligned}
$$



Simon J.D. Prince

Practical Issue

Problem: In high dimensions, the matrix **A** may be too big to invert

$$\mathbf{A} = \frac{1}{\sigma^2}\mathbf{X}\mathbf{X}^T + \frac{1}{\sigma_p^2}\mathbf{I}$$

Solution: Re-express using Matrix Inversion Lemma

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

$$\mathbf{A}^{-1} = \left(\frac{1}{\sigma^2}\mathbf{X}\mathbf{X}^T + \frac{1}{\sigma_p^2}\mathbf{I}_D\right)^{-1} = \sigma_p^2\mathbf{I}_D - \sigma_p^2\mathbf{X}\left(\mathbf{X}^T\mathbf{X} + \frac{\sigma^2}{\sigma_p^2}\mathbf{I}_I\right)^{-1}\mathbf{X}^T$$

Simon J.D. Prince

Recall: Matrix inversion relations

UNIVERSITÄT BONN

Consider the $d \times d$ matrix $\mathbf{A}$, the $k \times k$ matrix $\mathbf{C}$ and the $k \times d$ matrix $\mathbf{B}$ where $\mathbf{A}$ and $\mathbf{C}$ are symmetric, positive definite matrices. The following equality holds:

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

Proof:

$$\begin{aligned} \mathbf{B}^T\mathbf{C}^{-1}\mathbf{B}\mathbf{A}\mathbf{B}^T + \mathbf{B}^T &= \mathbf{B}^T + \mathbf{B}^T\mathbf{C}^{-1}\mathbf{B}\mathbf{A}\mathbf{B}^T \\ \mathbf{B}^T\mathbf{C}^{-1}(\mathbf{B}\mathbf{A}\mathbf{B}^T + \mathbf{C}) &= (\mathbf{A}^{-1} + \mathbf{B}^T\mathbf{C}^{-1}\mathbf{B})\mathbf{A}\mathbf{B}^T. \end{aligned}$$

Taking the inverse of both sides we get

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

as required.

Recall: Matrix inversion relations

Consider the $d \times d$ matrix $\mathbf{A}$, the $k \times k$ matrix $\mathbf{C}$ and the $k \times d$ matrix $\mathbf{B}$ where $\mathbf{A}$ and $\mathbf{C}$ are symmetric, positive definite matrices. The following equality holds:

$$(\mathbf{A}^{-1} + \mathbf{B}^T\mathbf{C}^{-1}\mathbf{B})^{-1} = \mathbf{A} - \mathbf{A}\mathbf{B}^T(\mathbf{B}\mathbf{A}\mathbf{B}^T + \mathbf{C})^{-1}\mathbf{B}\mathbf{A}. \quad \text{(C.61)}$$

This is sometimes known as the *matrix inversion lemma*.

Proof:

$$\begin{aligned}
&(\mathbf{A}^{-1} + \mathbf{B}^T\mathbf{C}^{-1}\mathbf{B})^{-1} \\
&\quad = (\mathbf{A}^{-1} + \mathbf{B}^T\mathbf{C}^{-1}\mathbf{B})^{-1}(\mathbf{I} + \mathbf{B}^T\mathbf{C}^{-1}\mathbf{B}\mathbf{A} - \mathbf{B}^T\mathbf{C}^{-1}\mathbf{B}\mathbf{A}) \\
&\quad = (\mathbf{A}^{-1} + \mathbf{B}^T\mathbf{C}^{-1}\mathbf{B})^{-1}\left((\mathbf{A}^{-1} + \mathbf{B}^T\mathbf{C}^{-1}\mathbf{B})\mathbf{A} - \mathbf{B}^T\mathbf{C}^{-1}\mathbf{B}\mathbf{A}\right) \\
&\quad = \mathbf{A} - (\mathbf{A}^{-1} + \mathbf{B}^T\mathbf{C}^{-1}\mathbf{B})^{-1}\mathbf{B}^T\mathbf{C}^{-1}\mathbf{B}\mathbf{A}. \quad \text{(C.62)}
\end{aligned}$$

We know:

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

Therefore:

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

Practical Issue

If the number of training samples I is less than feature dimension D, the equation

$$Pr(w^*|\mathbf{x}^*, \mathbf{X}, \mathbf{w}) = \text{Norm}_{w^*}\left[\frac{1}{\sigma^2}\mathbf{x}^{*T}\mathbf{A}^{-1}\mathbf{X}\mathbf{w}, \mathbf{x}^{*T}\mathbf{A}^{-1}\mathbf{x}^* + \sigma^2\right].$$

can be reformulated using

$$\mathbf{A}^{-1} = \left(\frac{1}{\sigma^2}\mathbf{X}\mathbf{X}^T + \frac{1}{\sigma_p^2}\mathbf{I}_D\right)^{-1} = \sigma_p^2\mathbf{I}_D - \sigma_p^2\mathbf{X}\left(\mathbf{X}^T\mathbf{X} + \frac{\sigma^2}{\sigma_p^2}\mathbf{I}_I\right)^{-1}\mathbf{X}^T$$

Final expression: inverses are (I x I) , not (D x D)

$$Pr(w^*|\mathbf{x}^*, \mathbf{X}, \mathbf{w}) =$$
$$\text{Norm}_{w^*}\left[\frac{\sigma_p^2}{\sigma^2}\mathbf{x}^{*T}\mathbf{X}\mathbf{w} - \frac{\sigma_p^2}{\sigma^2}\mathbf{x}^{*T}\mathbf{X}\left(\mathbf{X}^T\mathbf{X} + \frac{\sigma^2}{\sigma_p^2}\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{X}\mathbf{w},\right.$$
$$\left.\sigma_p^2\mathbf{x}^{*T}\mathbf{x}^* - \sigma_p^2\mathbf{x}^{*T}\mathbf{X}\left(\mathbf{X}^T\mathbf{X} + \frac{\sigma^2}{\sigma_p^2}\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{x}^* + \sigma^2\right]$$

Simon J.D. Prince

Fitting Variance

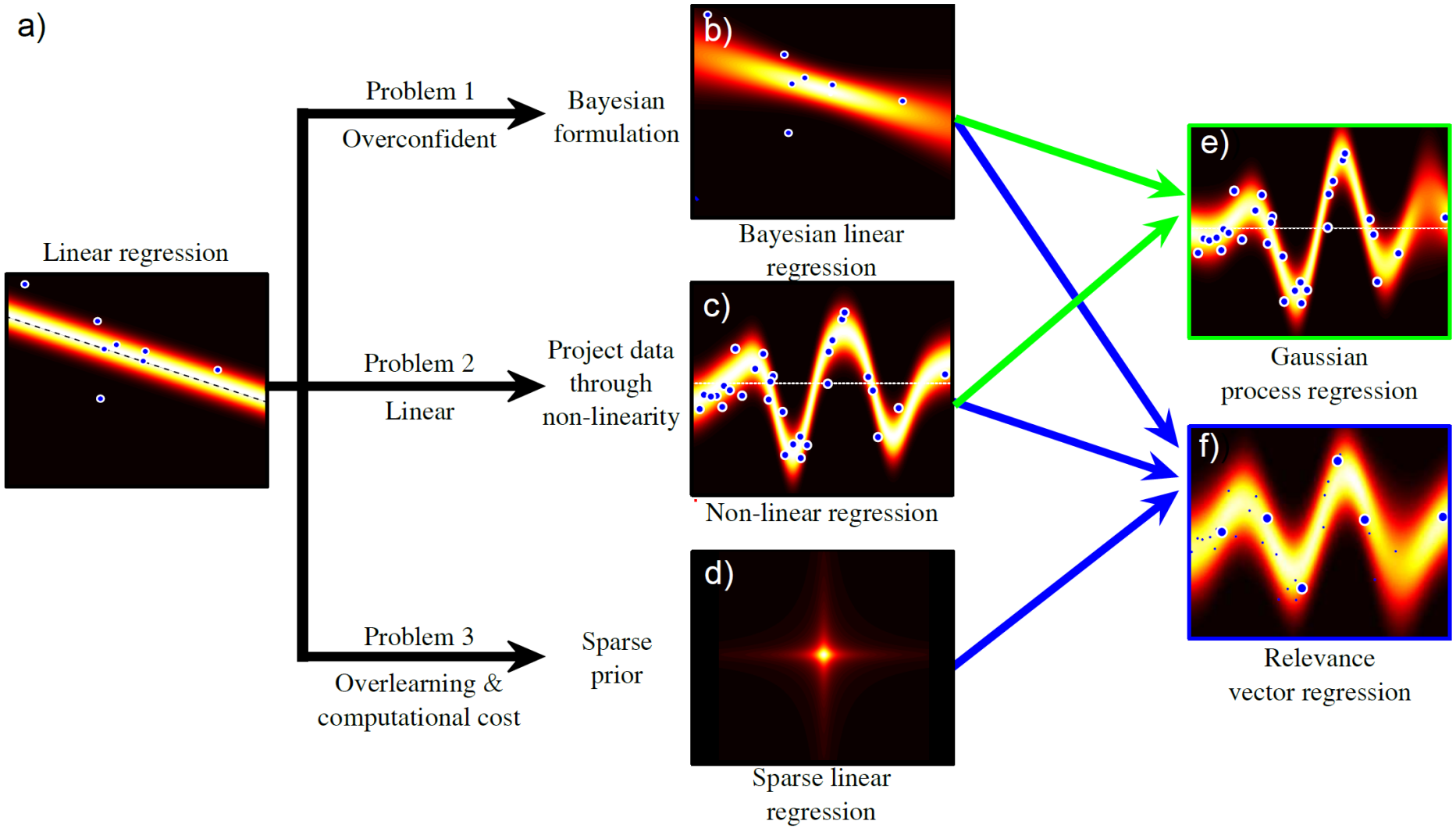
- We'll fit the variance with maximum likelihood
- Optimize the marginal likelihood

$$
\begin{aligned}
Pr(\mathbf{w}|\mathbf{X},\sigma^2) &= \int Pr(\mathbf{w}|\mathbf{X},\boldsymbol{\phi},\sigma^2)Pr(\boldsymbol{\phi})\, d\boldsymbol{\phi} \\
&= \int \mathrm{Norm}_{\mathbf{w}}[\mathbf{X}^T\boldsymbol{\phi},\sigma^2\mathbf{I}]\mathrm{Norm}_{\boldsymbol{\phi}}[\mathbf{0},\sigma_p^2\mathbf{I}]\, d\boldsymbol{\phi} \\
&= \mathrm{Norm}_{\mathbf{w}}[\mathbf{0},\sigma_p^2\mathbf{X}^T\mathbf{X}+\sigma^2\mathbf{I}]
\end{aligned}
$$

- σ^2 can be estimated by grid-search to get maximum of log likelihood or nonlinear optimization

Simon J.D. Prince

Regression Models



Simon J.D. Prince

Non-Linear Regression

UNIVERSITÄT BONN

GOAL:

Keep the math of linear regression, but extend to more general functions

KEY IDEA:

You can make a non-linear function from a linear weighted sum of non-linear basis functions

Simon J.D. Prince

Non-linear regression

Linear regression:

$$Pr(w_i|\mathbf{x}_i, \boldsymbol{\theta}) = \text{Norm}_{w_i}\left[\boldsymbol{\phi}^T\mathbf{x}_i, \sigma^2\right]$$

Non-Linear regression:

$$Pr(w_i|\mathbf{x}_i, \boldsymbol{\theta}) = \text{Norm}_{w_i}[\boldsymbol{\phi}^T\mathbf{z}_i, \sigma^2]$$

where $\mathbf{z}_i = \mathbf{f}[\mathbf{x}_i]$

In other words, create z by evaluating x against basis functions, then linearly regress against z.

Simon J.D. Prince

Example: polynomial regression

$$Pr(w_i|x_i) = \text{Norm}_{w_i}[\phi_0 + \phi_1 x_i + \phi_2 x_i^2 + \phi_3 x_i^3, \sigma^2].$$

A special case of

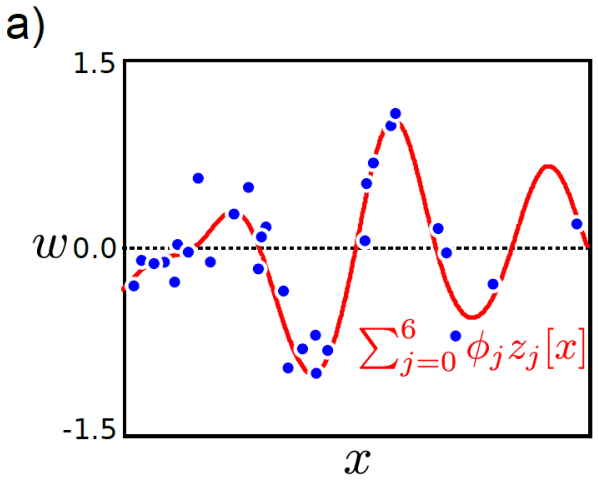
$$Pr(w_i|\mathbf{x}_i) = \text{Norm}_{w_i}[\boldsymbol{\phi}^T \mathbf{z}_i, \sigma^2]$$

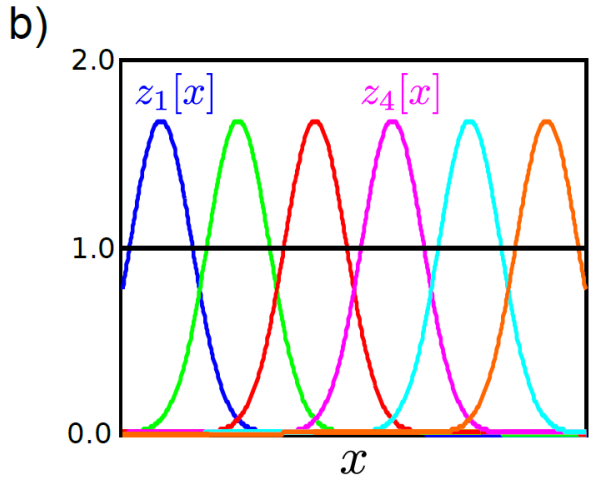
Where

$$\mathbf{z}_i = \begin{bmatrix} 1 \\ x_i \\ x_i^2 \\ x_i^3 \end{bmatrix}$$

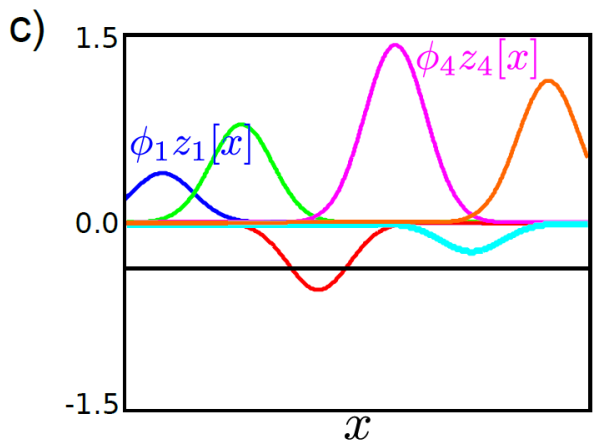
Simon J.D. Prince

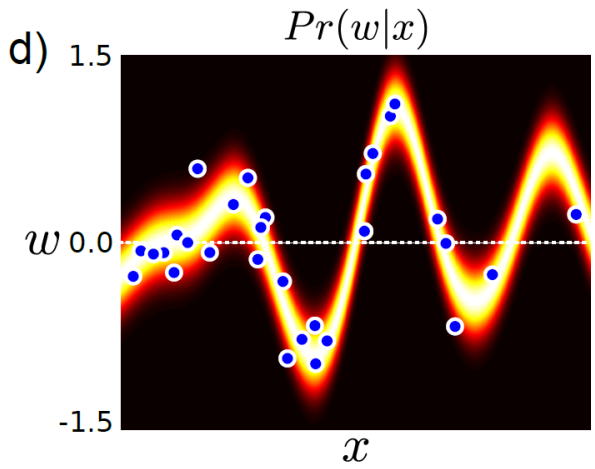
Radial basis functions





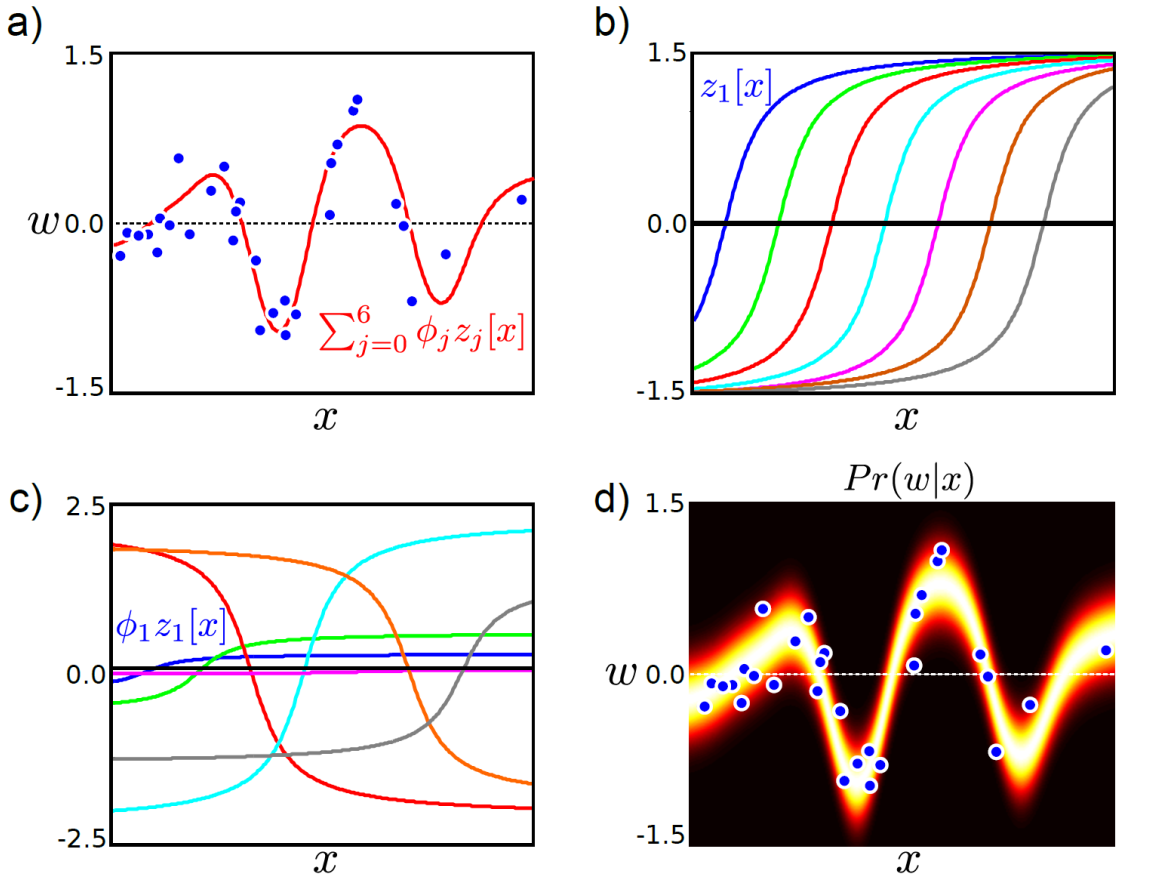
$$\mathbf{z}_i = \begin{bmatrix} 1 \\ \exp\left[-(x_i-\alpha_1)^2/\lambda\right] \\ \exp\left[-(x_i-\alpha_2)^2/\lambda\right] \\ \exp\left[-(x_i-\alpha_3)^2/\lambda\right] \\ \exp\left[-(x_i-\alpha_4)^2/\lambda\right] \\ \exp\left[-(x_i-\alpha_5)^2/\lambda\right] \\ \exp\left[-(x_i-\alpha_6)^2/\lambda\right] \end{bmatrix}$$





Simon J.D. Prince

Arc Tan Functions



$$\mathbf{z}_i = \begin{bmatrix} \arctan[\lambda x_i - \alpha_1] \\ \arctan[\lambda x_i - \alpha_2] \\ \arctan[\lambda x_i - \alpha_3] \\ \arctan[\lambda x_i - \alpha_4] \\ \arctan[\lambda x_i - \alpha_5] \\ \arctan[\lambda x_i - \alpha_6] \\ \arctan[\lambda x_i - \alpha_7] \end{bmatrix}$$

Simon J.D. Prince

Non-linear regression

Linear regression:

$$Pr(w_i|\mathbf{x}_i, \boldsymbol{\theta}) = \text{Norm}_{w_i}\left[\boldsymbol{\phi}^T\mathbf{x}_i, \sigma^2\right]$$

Non-Linear regression:

$$Pr(w_i|\mathbf{x}_i, \boldsymbol{\theta}) = \text{Norm}_{w_i}[\boldsymbol{\phi}^T\mathbf{z}_i, \sigma^2]$$

where $\mathbf{z}_i = \mathbf{f}[\mathbf{x}_i]$

In other words, create z by evaluating x against basis functions, then linearly regress against z.

Simon J.D. Prince

Maximum Likelihood

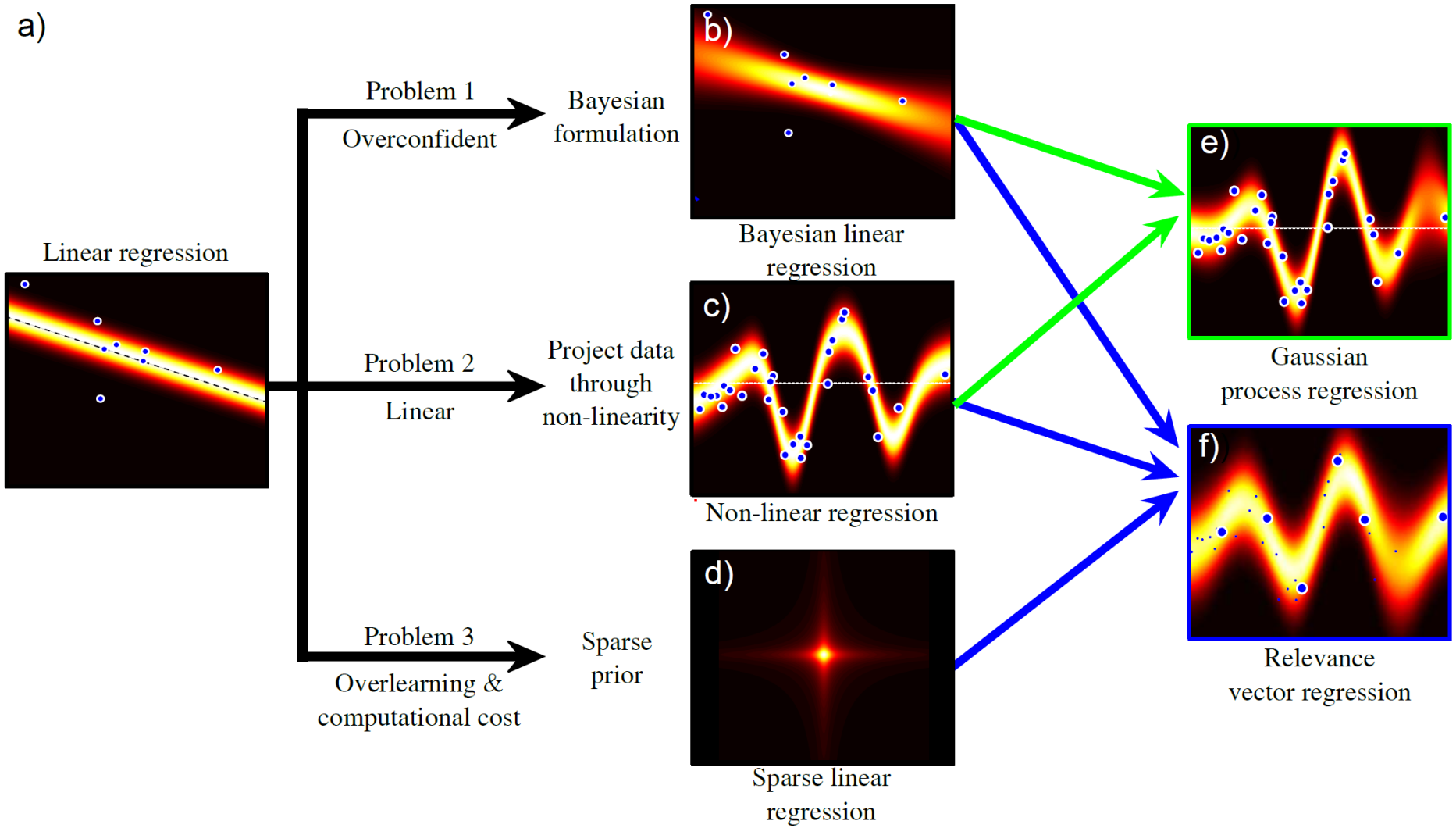
Same as linear regression, but substitute in **Z** for **X**:

$$\hat{\boldsymbol{\phi}} = (\mathbf{Z}\mathbf{Z}^T)^{-1}\mathbf{Z}\mathbf{w}$$

$$\hat{\sigma}^2 = \frac{(\mathbf{w}-\mathbf{Z}^T\boldsymbol{\phi})^T(\mathbf{w}-\mathbf{Z}^T\boldsymbol{\phi})}{I}$$

Simon J.D. Prince

Regression Models



Simon J.D. Prince

Bayesian Approach

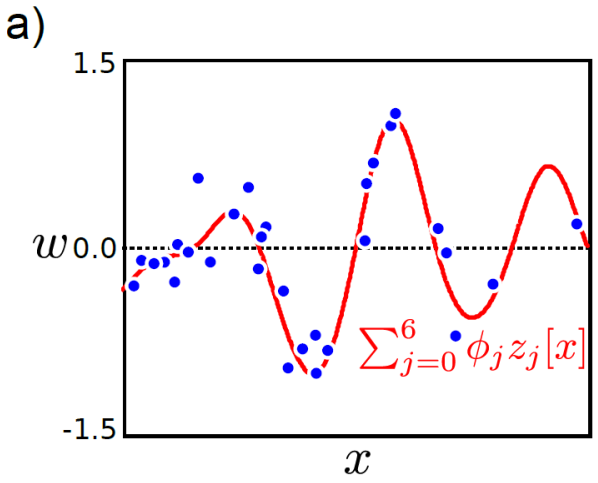
Learn σ^2 from marginal likelihood as before

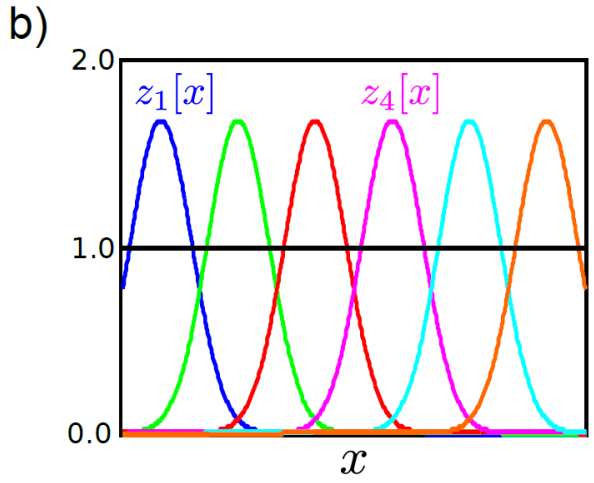
Final predictive distribution:

$$Pr(w^*|\mathbf{z}^*, \mathbf{X}, \mathbf{w}) = \text{Norm}_w \left[\frac{\sigma_p^2}{\sigma^2}\mathbf{z}^{*T}\mathbf{Z}\mathbf{w} - \frac{\sigma_p^2}{\sigma^2}\mathbf{z}^{*T}\mathbf{Z}\left(\mathbf{Z}^T\mathbf{Z} + \frac{\sigma^2}{\sigma_p^2}\mathbf{I}\right)^{-1}\mathbf{Z}^T\mathbf{Z}\mathbf{w}, \right.$$
$$\left. \sigma_p^2\mathbf{z}^{*T}\mathbf{z}^* - \sigma_p^2\mathbf{z}^{*T}\mathbf{Z}\left(\mathbf{Z}^T\mathbf{Z} + \frac{\sigma^2}{\sigma_p^2}\mathbf{I}\right)^{-1}\mathbf{Z}^T\mathbf{z}^* + \sigma^2 \right]$$

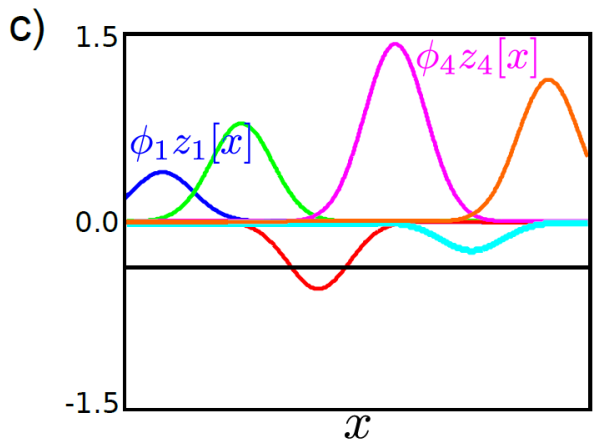
Simon J.D. Prince

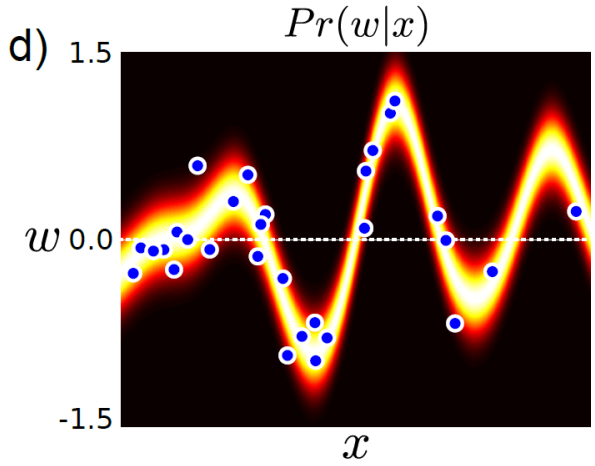
Radial basis functions





$$\mathbf{z}_i = \begin{bmatrix} 1 \\ \exp\left[-(x_i-\alpha_1)^2/\lambda\right] \\ \exp\left[-(x_i-\alpha_2)^2/\lambda\right] \\ \exp\left[-(x_i-\alpha_3)^2/\lambda\right] \\ \exp\left[-(x_i-\alpha_4)^2/\lambda\right] \\ \exp\left[-(x_i-\alpha_5)^2/\lambda\right] \\ \exp\left[-(x_i-\alpha_6)^2/\lambda\right] \end{bmatrix}$$

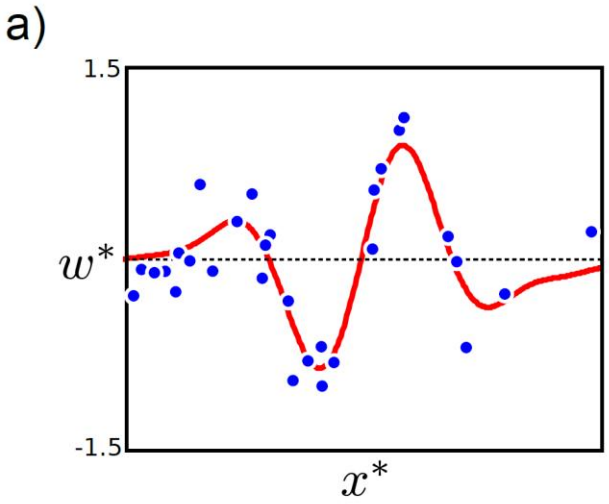


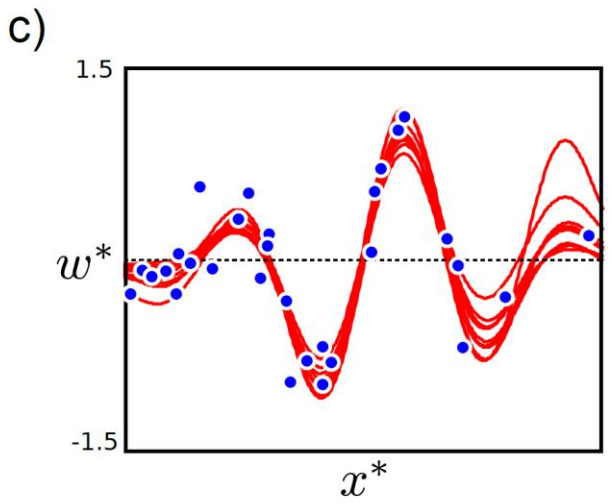


Constant variance σ^2

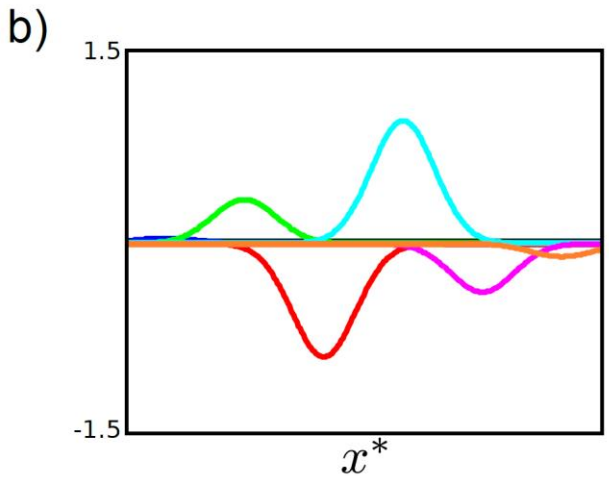
Simon J.D. Prince

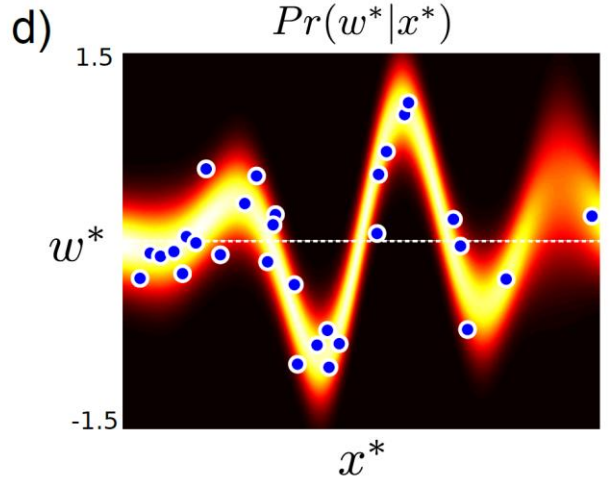
Radial basis functions - Bayesian





$$\mathbf{z}_i = \begin{bmatrix} 1 \\ \exp\left[-(x_i-\alpha_1)^2/\lambda\right] \\ \exp\left[-(x_i-\alpha_2)^2/\lambda\right] \\ \exp\left[-(x_i-\alpha_3)^2/\lambda\right] \\ \exp\left[-(x_i-\alpha_4)^2/\lambda\right] \\ \exp\left[-(x_i-\alpha_5)^2/\lambda\right] \\ \exp\left[-(x_i-\alpha_6)^2/\lambda\right] \end{bmatrix}$$





Simon J.D. Prince

The Kernel Trick

Notice that the final equation doesn't need the data itself, but just dot products between data items of the form $\mathbf{z}_i^T\mathbf{z}_j$

$$Pr(w^*|\mathbf{z}^*,\mathbf{X},\mathbf{w}) =$$
$$\text{Norm}_w\left[\frac{\sigma_p^2}{\sigma^2}\mathbf{z}^{*T}\mathbf{Z}\mathbf{w}-\frac{\sigma_p^2}{\sigma^2}\mathbf{z}^{*T}\mathbf{Z}\left(\mathbf{Z}^T\mathbf{Z}+\frac{\sigma^2}{\sigma_p^2}\mathbf{I}\right)^{-1}\mathbf{Z}^T\mathbf{Z}\mathbf{w},\right.$$
$$\left.\sigma_p^2\mathbf{z}^{*T}\mathbf{z}^*-\sigma_p^2\mathbf{z}^{*T}\mathbf{Z}\left(\mathbf{Z}^T\mathbf{Z}+\frac{\sigma^2}{\sigma_p^2}\mathbf{I}\right)^{-1}\mathbf{Z}^T\mathbf{z}^*+\sigma^2\right]$$

So, we take data $\mathbf{x}_i$ and $\mathbf{x}_j$ pass through non-linear function to create $\mathbf{z}_i$ and $\mathbf{z}_j$ and then take dot products of different $\mathbf{z}_i^T\mathbf{z}_j$

Simon J.D. Prince

The Kernel Trick

So, we take data $\mathbf{x}_i$ and $\mathbf{x}_j$ pass through non-linear function to create $\mathbf{z}_i$ and $\mathbf{z}_j$ and then take dot products of different $\mathbf{z}_i^T\mathbf{z}_j$

Key idea:

Define a “kernel” function that does all of this together.

- Takes data $\mathbf{x}_i$ and $\mathbf{x}_j$
- Returns a value for dot product $\mathbf{z}_i^T\mathbf{z}_j$

If we choose this function carefully, then it will correspond to some underlying $\mathbf{z}=\mathbf{f}[\mathbf{x}]$.

Never compute $\mathbf{z}$ explicitly - can be very high or infinite dimension

Simon J.D. Prince

Gaussian Process Regression

UNIVERSITÄT BONN

Before

$$Pr(w^*|\mathbf{z}^*, \mathbf{X}, \mathbf{w}) =$$

$$\text{Norm}_w \left[\frac{\sigma_p^2}{\sigma^2}\mathbf{z}^{*T}\mathbf{Z}\mathbf{w} - \frac{\sigma_p^2}{\sigma^2}\mathbf{z}^{*T}\mathbf{Z}\left(\mathbf{Z}^T\mathbf{Z} + \frac{\sigma^2}{\sigma_p^2}\mathbf{I}\right)^{-1}\mathbf{Z}^T\mathbf{Z}\mathbf{w}, \right.$$

$$\left. \sigma_p^2\mathbf{z}^{*T}\mathbf{z}^* - \sigma_p^2\mathbf{z}^{*T}\mathbf{Z}\left(\mathbf{Z}^T\mathbf{Z} + \frac{\sigma^2}{\sigma_p^2}\mathbf{I}\right)^{-1}\mathbf{Z}^T\mathbf{z}^* + \sigma^2 \right]$$

After

$$Pr(w^*|\mathbf{x}^*, \mathbf{X}, \mathbf{w}) =$$

$$\text{Norm}_{w^*} \left[\frac{\sigma_p^2}{\sigma^2}\mathbf{K}[\mathbf{x}^*, \mathbf{X}]\mathbf{w} - \frac{\sigma_p^2}{\sigma^2}\mathbf{K}[\mathbf{x}^*, \mathbf{X}]\left(\mathbf{K}[\mathbf{X}, \mathbf{X}] + \frac{\sigma^2}{\sigma_p^2}\mathbf{I}\right)^{-1}\mathbf{K}[\mathbf{X}, \mathbf{X}]\mathbf{w}, \right.$$

$$\left. \sigma_p^2\mathbf{K}[\mathbf{x}^*, \mathbf{x}^*] - \sigma_p^2\mathbf{K}[\mathbf{x}^*, \mathbf{X}]\left(\mathbf{K}[\mathbf{X}, \mathbf{X}] + \frac{\sigma^2}{\sigma_p^2}\mathbf{I}\right)^{-1}\mathbf{K}[\mathbf{X}, \mathbf{x}^*] + \sigma^2 \right]$$

where the notation $\mathbf{K}[\mathbf{X}, \mathbf{X}]$ represents a matrix of dot products where element (i, j) is given by $\mathrm{k}[\mathbf{x}_i, \mathbf{x}_j]$.

Simon J.D. Prince

Why use k(x, x') instead of (φ(x), φ(x'))?

UNIVERSITÄT BONN

1) Memory usage:

- Storing $\varphi(x_1), \ldots, \varphi(x_n)$ requires $O(nm)$ memory.
- Storing $k(x_1, x_1), \ldots, k(x_n, x_n)$ requires $O(n^2)$ memory.

2) Speed:

- We might find an expression for $k(x_i, x_j)$ that is faster to calculate than forming $\varphi(x_i)$ and then $\langle \varphi(x_i), \varphi(x_j) \rangle$.

  Example: comparing angles $(x \in [0, 2\pi])$

$$\varphi : x \mapsto (\cos(x), \sin(x)) \in \mathbb{R}^2$$

$$\begin{aligned} \langle \varphi(x_i), \varphi(x_j) \rangle &= \langle\ (\cos(x_i), \sin(x_i)), (\cos(x_j), \sin(x_j))\ \rangle \\ &= \cos(x_i)\cos(x_j) + \sin(x_i)\sin(x_j)\ \ = \cos(x_i - x_j) \end{aligned}$$

  Equivalently, but faster, without φ:

$$k(x_i, x_j) := \cos(x_i - x_j)$$

Example Kernels

- linear

$$k[\mathbf{x}_i, \mathbf{x}_j] = \mathbf{x}_i^T \mathbf{x}_j,$$

- degree p polynomial

$$k[\mathbf{x}_i, \mathbf{x}_j] = (\mathbf{x}_i^T \mathbf{x}_j + 1)^p,$$

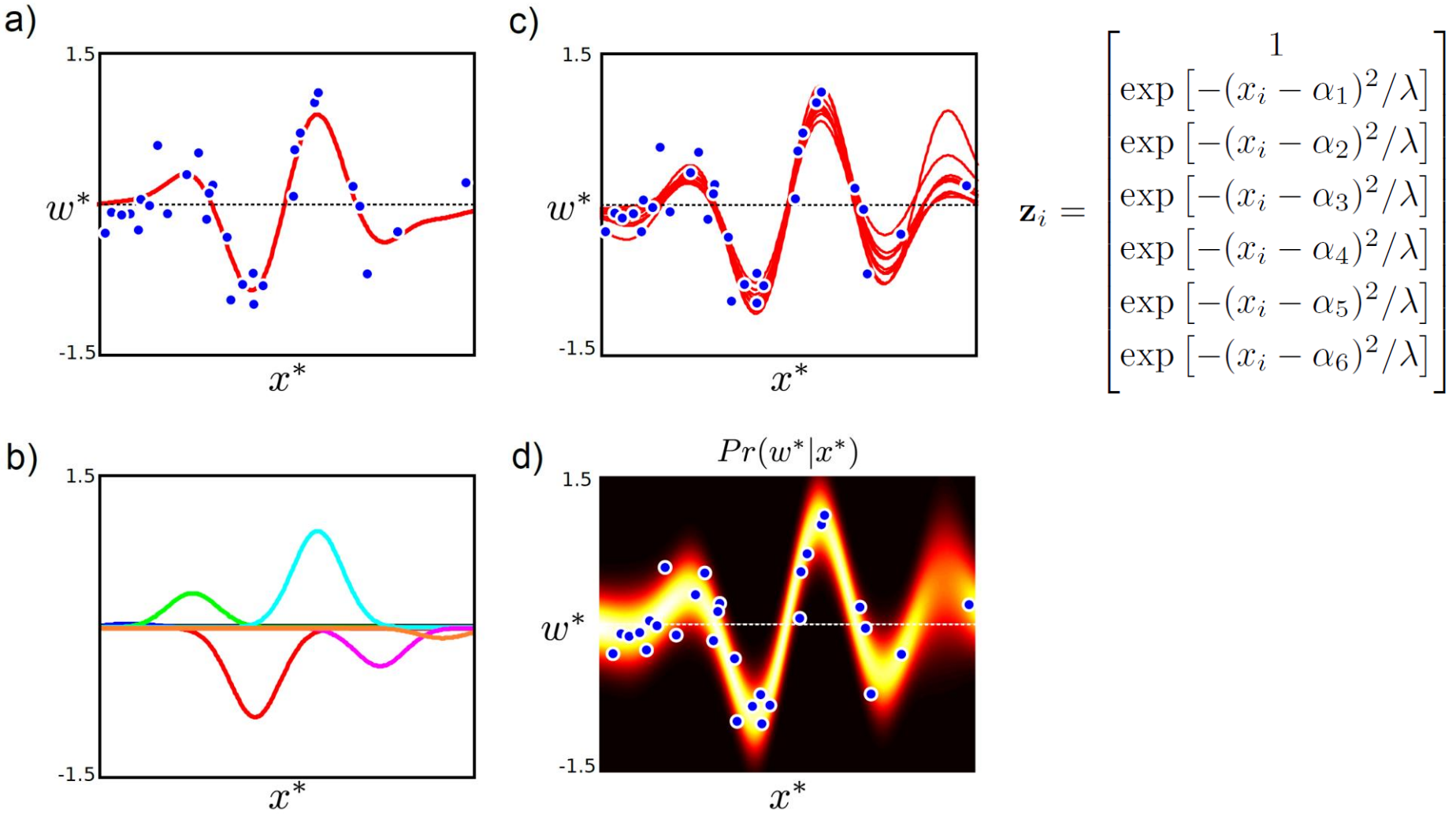
- radial basis function (RBF) or Gaussian

$$k[\mathbf{x}_i, \mathbf{x}_j] = \exp\left[-0.5\left(\frac{(\mathbf{x}_i - \mathbf{x}_j)^T(\mathbf{x}_i - \mathbf{x}_j)}{\lambda^2}\right)\right].$$

(Equivalent to having an infinite number of radial basis functions at every position in space.)

Simon J.D. Prince

Radial basis functions - Bayesian



Simon J.D. Prince

RBF Kernel Fits

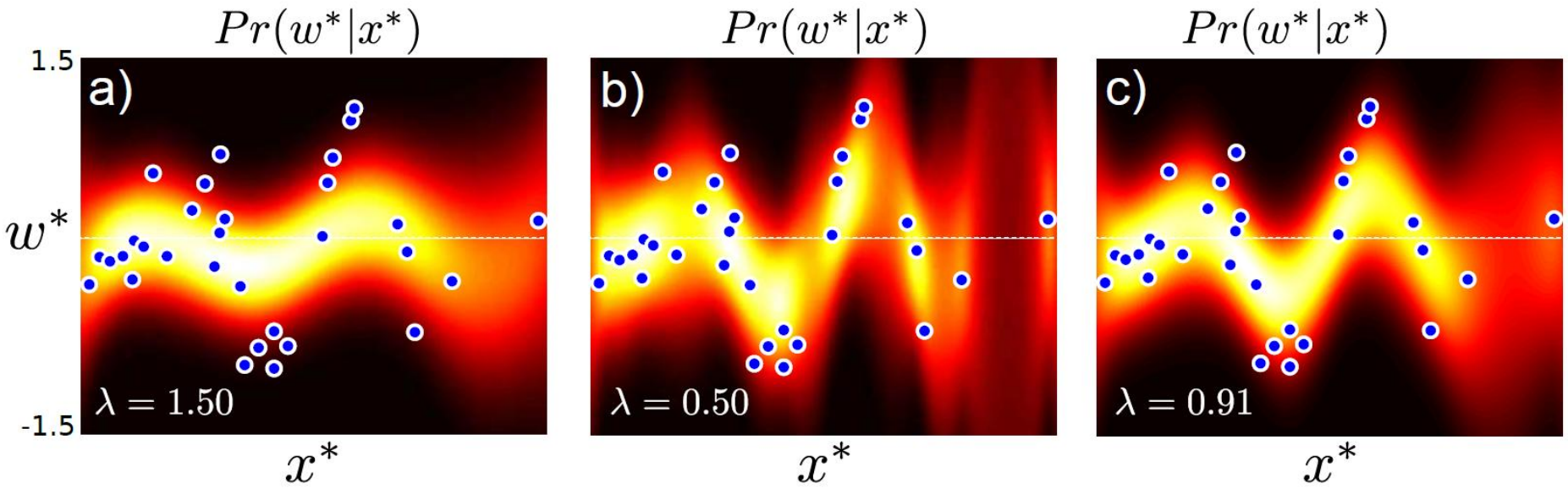


Figure 8.9 Gaussian process regression using an RBF kernel a) When the length scale parameter λ is large, the function is too smooth. b) For small values of the length parameter the model does not successfully interpolate between the examples. c) The regression using the maximum likelihood length scale parameter is neither too smooth nor disjointed.

$$k[\mathbf{x}_i, \mathbf{x}_j] = \exp\left[-0.5\left(\frac{\mathbf{x}_i - \mathbf{x}_j}{\lambda}\right)^2\right]$$

Simon J.D. Prince

Fitting Variance

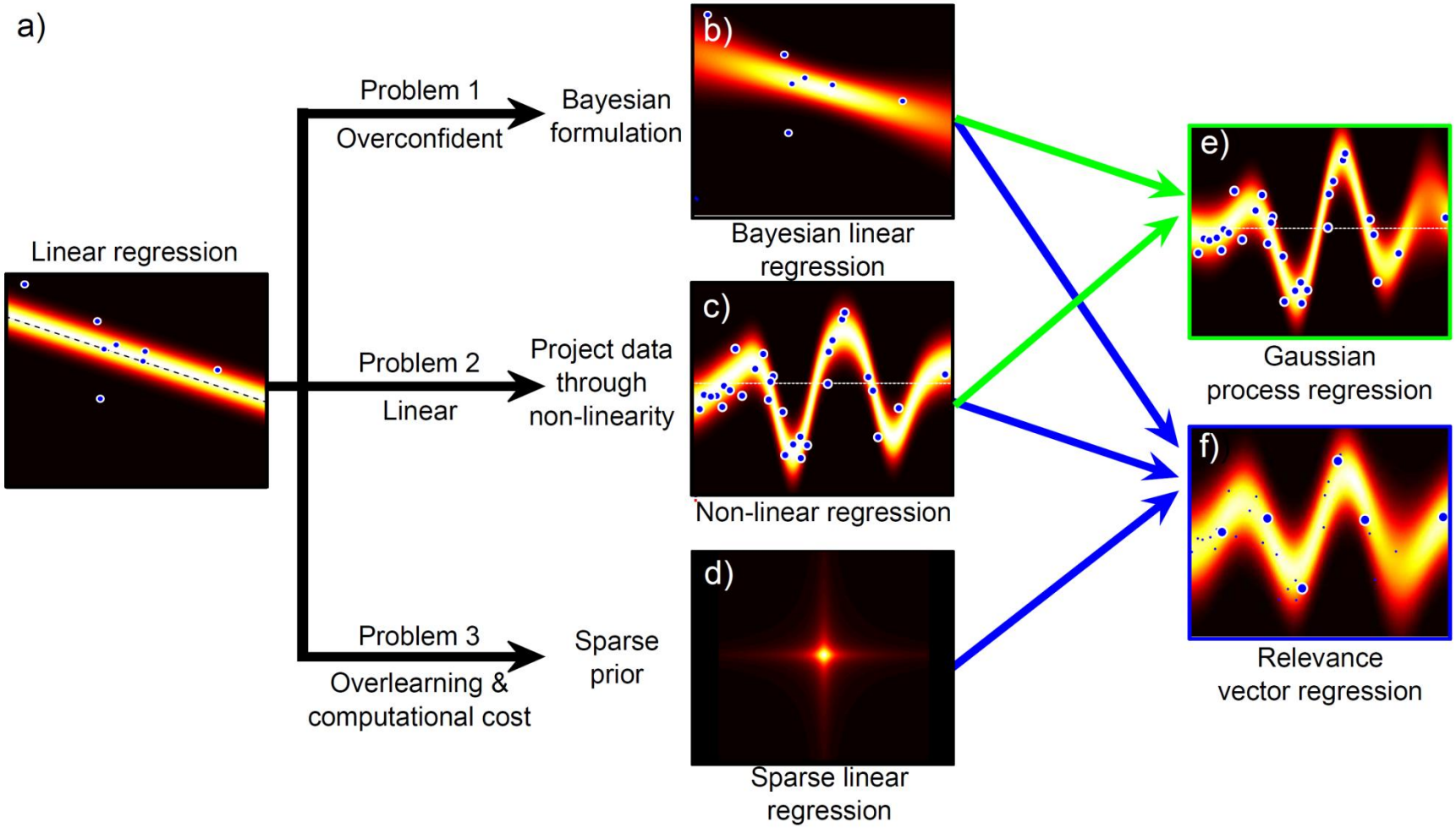
- We'll fit the variance with maximum likelihood
- Optimize the marginal likelihood

$$\begin{aligned} Pr(\mathbf{w}|\mathbf{X}, \sigma^2) &= \int Pr(\mathbf{w}|\mathbf{X}, \boldsymbol{\phi}, \sigma^2) Pr(\boldsymbol{\phi}) d\boldsymbol{\phi} \\ &= \text{Norm}_{\mathbf{w}}[\mathbf{0}, \sigma_p^2 \mathbf{K}[\mathbf{X}, \mathbf{X}] + \sigma^2 \mathbf{I}] \end{aligned}$$

- Have to use non-linear optimization

Simon J.D. Prince

Regression Models



Simon J.D. Prince

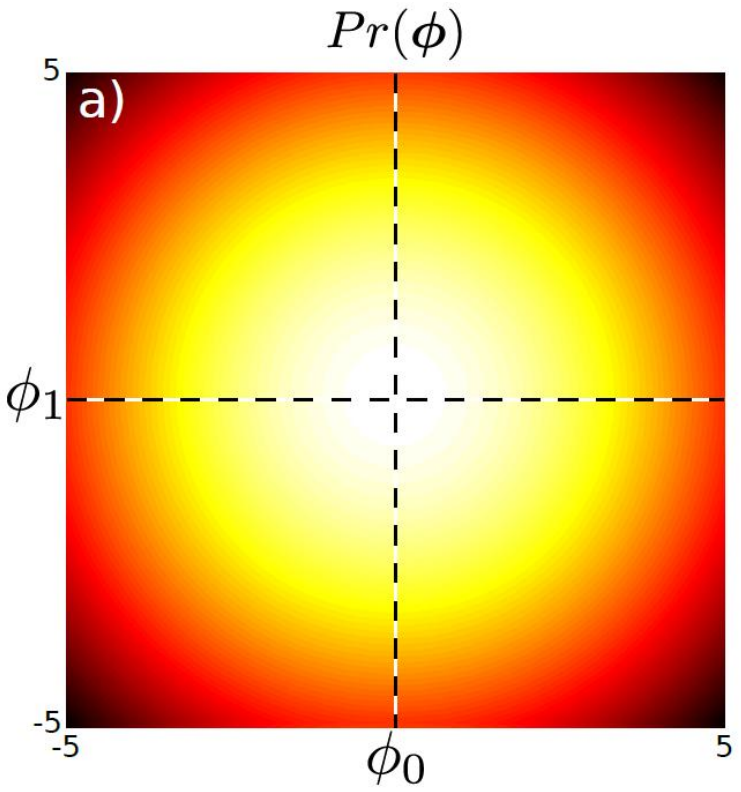
Sparse Linear Regression

Perhaps not every dimension of the data **x** is informative

A sparse solution forces some of the coefficients in ϕ to be zero

Method:

– apply a different prior on ϕ that encourages sparsity



Sparse Linear Regression

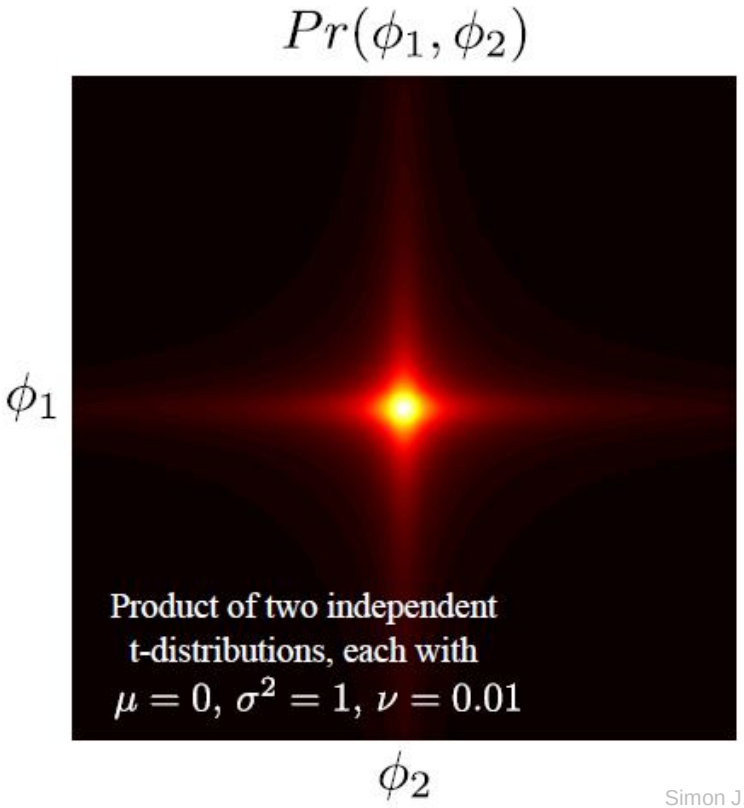
Perhaps not every dimension of the data **x** is informative

A sparse solution forces some of the coefficients in ϕ to be zero

Method:

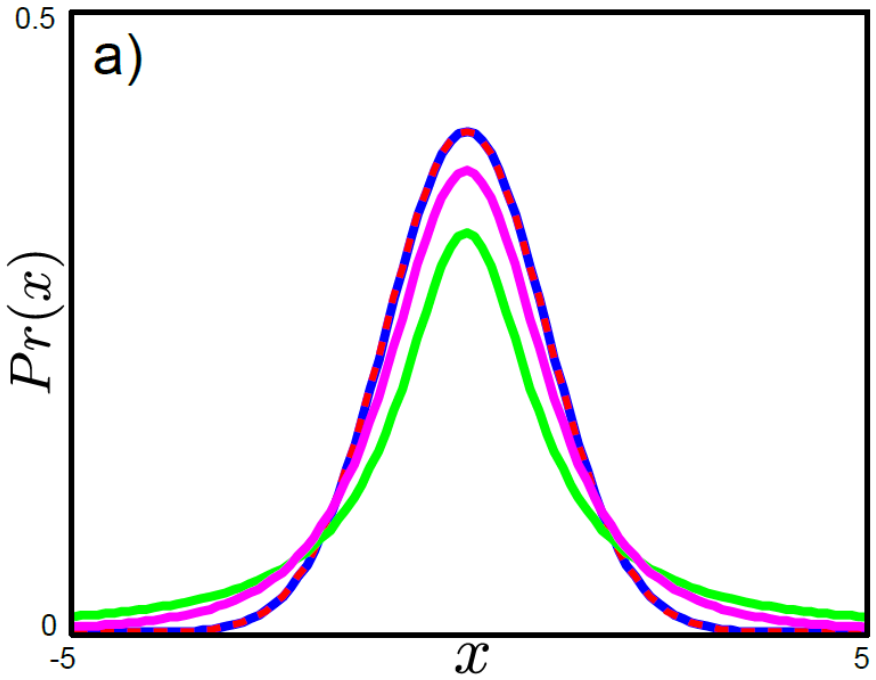
– apply a different prior on ϕ that encourages sparsity

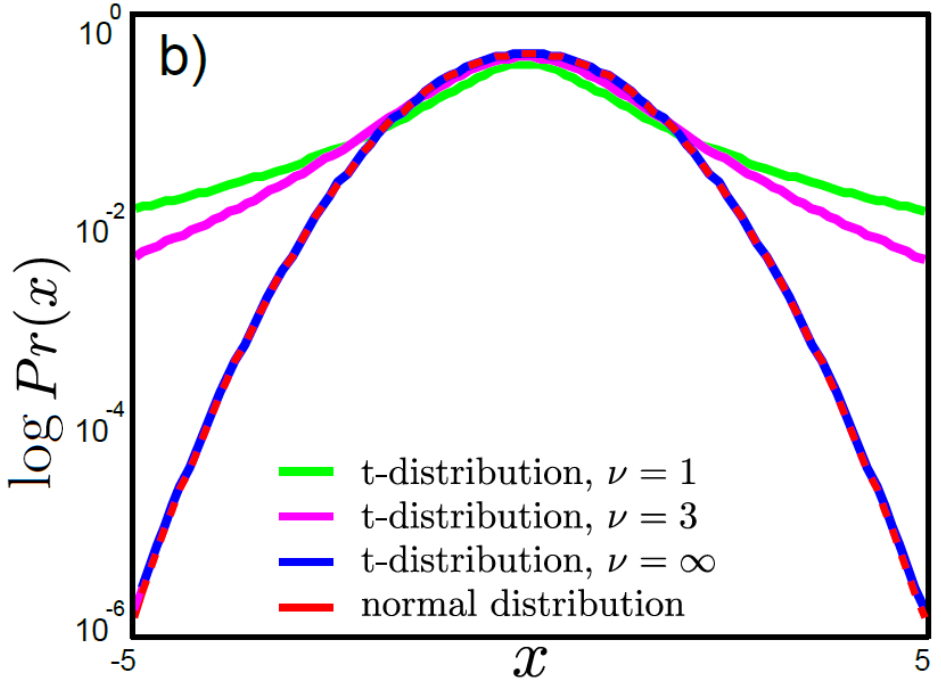
– product of t-distributions



Simon J.D. Prince

Student t-distributions





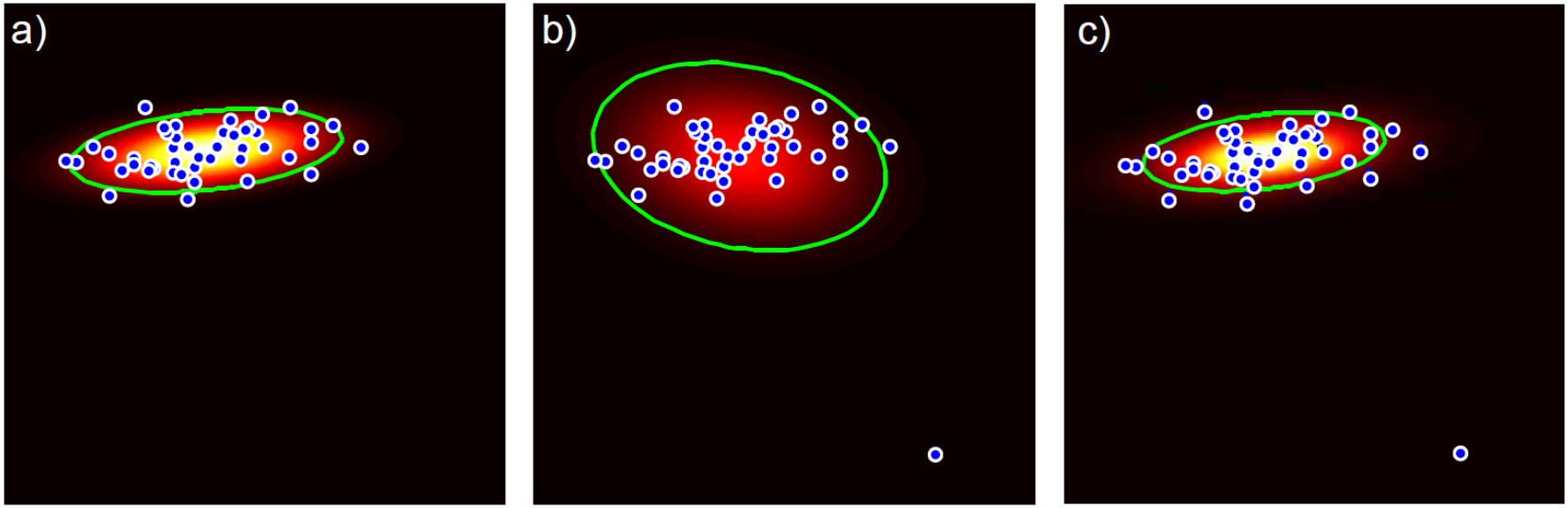
$$Pr(x) = \text{Stud}_{\mathbf{x}}\left[\mu, \sigma^2, \nu\right]$$

$$= \frac{\Gamma\left[\frac{\nu+1}{2}\right]}{\sqrt{\nu\pi\sigma^2}\Gamma\left[\frac{\nu}{2}\right]}\left(1+\frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}}$$

Simon J.D. Prince

Student t-distributions motivation

The normal distribution is not very robust – a single outlier can completely throw it off because the tails fall off so fast...



Normal distribution

Normal distribution
w/ one extra datapoint!

t-distribution

Simon J.D. Prince

Student t-distributions

UNIVERSITÄT BONN

Univariate student t-distribution

$$\begin{aligned} Pr(x) &= \text{Stud}_{\mathbf{x}}\left[\mu, \sigma^2, \nu\right] \\ &= \frac{\Gamma\left[\frac{\nu+1}{2}\right]}{\sqrt{\nu\pi\sigma^2}\Gamma\left[\frac{\nu}{2}\right]}\left(1+\frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}} \end{aligned}$$

Multivariate student t-distribution

$$\begin{aligned} Pr(\mathbf{x}) &= \text{Stud}_{\mathbf{x}}\left[\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu\right] \\ &= \frac{\Gamma\left[\frac{\nu+D}{2}\right]}{(\nu\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}\Gamma\left[\frac{\nu}{2}\right]}\left(1+\frac{(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}{\nu}\right)^{-\frac{\nu+D}{2}} \end{aligned}$$

Simon J.D. Prince

t-distribution as a marginalization

Define hidden variable h

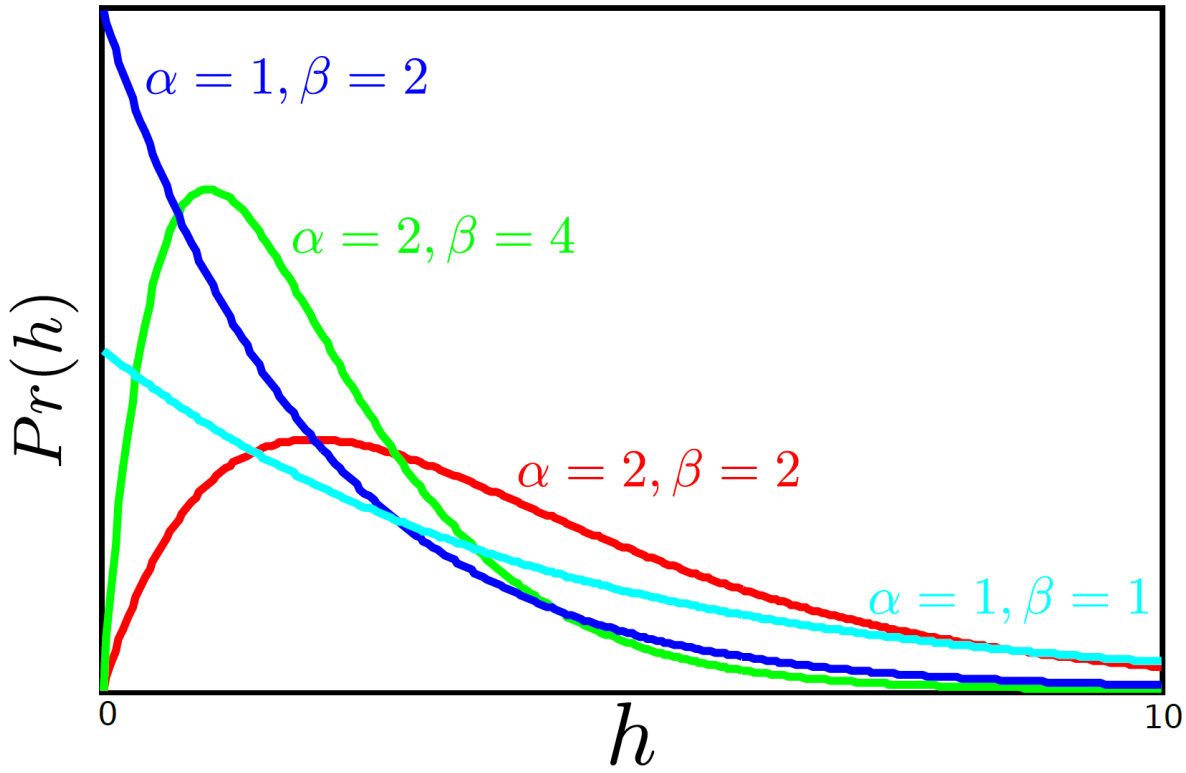
$$\begin{aligned} Pr(\mathbf{x}|h) &= \text{Norm}_x[\boldsymbol{\mu}, \boldsymbol{\Sigma}/h] \\ Pr(h) &= \text{Gam}_h[\nu/2, \nu/2] \end{aligned}$$

Can be expressed as a marginalization (Expectation-maximization)

$$\begin{aligned} Pr(\mathbf{x}) = \int Pr(\mathbf{x}, h)dh &= \int Pr(\mathbf{x}|h)Pr(h)dh \\ &= \int \text{Norm}_x[\boldsymbol{\mu}, \boldsymbol{\Sigma}/h]\text{Gam}_h[\nu/2, \nu/2]dh \\ &= \text{Stud}_{\mathbf{x}}[\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu]. \end{aligned}$$

Simon J.D. Prince

Gamma distribution



$$\mathrm{Gam}_h[\alpha, \beta] = \frac{\beta^\alpha}{\Gamma[\alpha]} \exp[-\beta h] h^{\alpha-1}$$

Simon J.D. Prince

t-distribution as a marginalization

Define hidden variable h

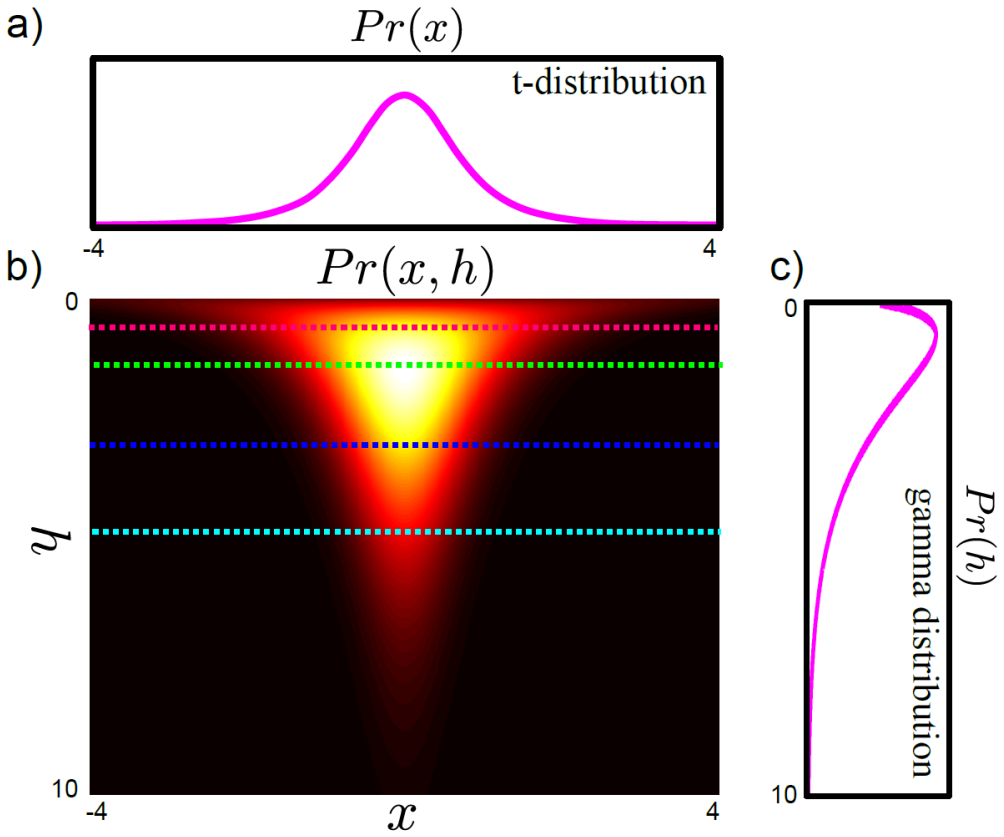
$$\begin{aligned} Pr(\mathbf{x}|h) &= \text{Norm}_x[\boldsymbol{\mu}, \boldsymbol{\Sigma}/h] \\ Pr(h) &= \text{Gam}_h[\nu/2, \nu/2] \end{aligned}$$

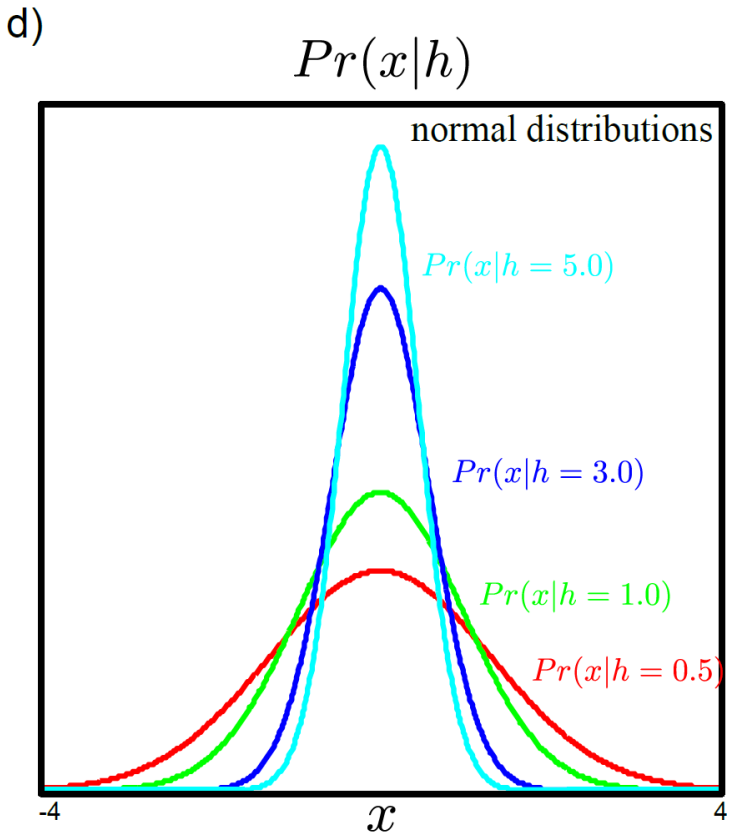
Things to note:

- Again this provides a method to sample from the t-distribution
- Variable h has a clear interpretation:
  - Each datum drawn from a Gaussian, mean $\boldsymbol{\mu}$
  - Covariance depends inversely on h
- Can think of this as an infinite mixture (sum becomes integral) of Gaussians w/ same mean, but different variances

Simon J.D. Prince

t-distribution





Simon J.D. Prince

Sparse Linear Regression

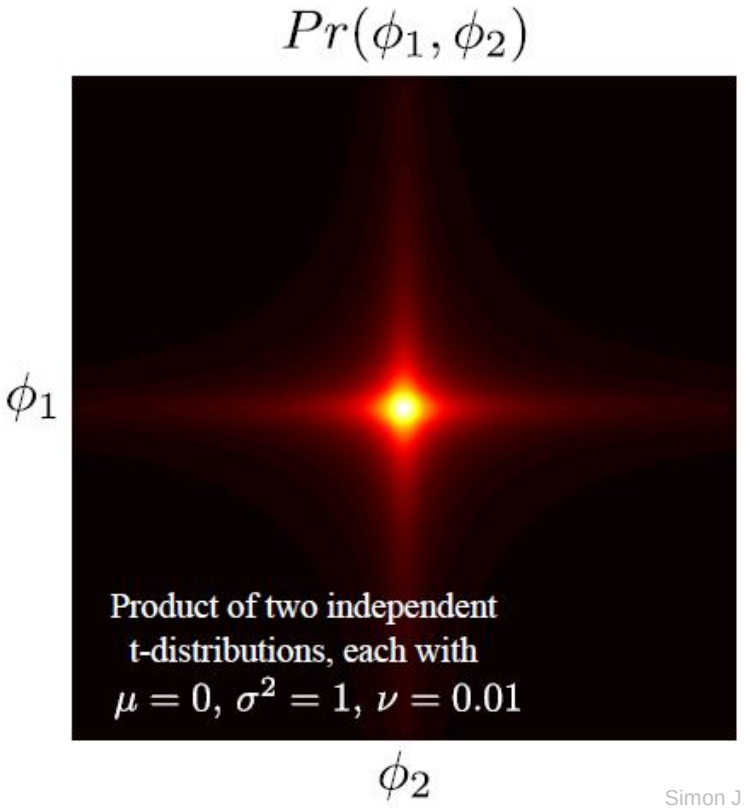
Perhaps not every dimension of the data **x** is informative

A sparse solution forces some of the coefficients in ϕ to be zero

Method:

– apply a different prior on ϕ that encourages sparsity

– product of t-distributions



Simon J.D. Prince

Sparse Linear Regression

Apply product of t-distributions to parameter vector

$$\begin{aligned} Pr(\boldsymbol{\phi}) &= \prod_{d=1}^{D} \text{Stud}_{\phi_d}[0, 1, \nu] \\ &= \prod_{d=1}^{D} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{\phi_d^2}{\nu}\right)^{-(\nu+1)/2} \end{aligned}$$

As before, we use

$$Pr(\boldsymbol{\phi}|\mathbf{X}, \mathbf{w}, \sigma^2) = \frac{Pr(\mathbf{w}|\mathbf{X}, \boldsymbol{\phi}, \sigma^2) Pr(\boldsymbol{\phi})}{Pr(\mathbf{w}|\mathbf{X}, \sigma^2)}$$

Cannot compute posterior in closed from.

Simon J.D. Prince

Sparse Linear Regression

To make progress, write as marginal of joint distribution

$$
\begin{aligned}
Pr(\boldsymbol{\phi}) &= \prod_{d=1}^{D} \int \mathrm{Norm}_{\phi_d}[0, 1/h_d]\mathrm{Gam}_{h_d}[\nu/2, \nu/2]\, dh_d \\
&= \int \mathrm{Norm}_{\boldsymbol{\phi}}[0, \mathbf{H}^{-1}] \prod_{d=1}^{D} \mathrm{Gam}_{h_d}[\nu/2, \nu/2]\, d\mathbf{H},
\end{aligned}
$$

Diagonal matrix with hidden variables $\{h_d\}$ on diagonal

Simon J.D. Prince

Sparse Linear Regression

UNIVERSITÄT BONN

Substituting in the prior

$$
\begin{aligned}
Pr(\mathbf{w}|\mathbf{X},\sigma^2) &= \int Pr(\mathbf{w},\boldsymbol{\phi}|\mathbf{X},\sigma^2)\, d\boldsymbol{\phi} \\
&= \int Pr(\mathbf{w}|\mathbf{X},\boldsymbol{\phi},\sigma^2) Pr(\boldsymbol{\phi})\, d\boldsymbol{\phi} \\
&= \int \text{Norm}_{\mathbf{w}}[\mathbf{X}^T\boldsymbol{\phi},\sigma^2\mathbf{I}] \int \text{Norm}_{\boldsymbol{\phi}}[0,\mathbf{H}^{-1}] \prod_{d=1}^{D} \text{Gam}_{h_d}[\nu/2,\nu/2]\, d\mathbf{H} d\boldsymbol{\phi} \\
&= \int\int \text{Norm}_{\mathbf{w}}[\mathbf{X}^T\boldsymbol{\phi},\sigma^2\mathbf{I}] \text{Norm}_{\boldsymbol{\phi}}[0,\mathbf{H}^{-1}] \prod_{d=1}^{D} \text{Gam}_{h_d}[\nu/2,\nu/2]\, d\mathbf{H} d\boldsymbol{\phi} \\
&= \int \text{Norm}_{\mathbf{w}}[\mathbf{0},\mathbf{X}^T\mathbf{H}^{-1}\mathbf{X}+\sigma^2\mathbf{I}] \prod_{d=1}^{D} \text{Gam}_{h_d}[\nu/2,\nu/2]\, d\mathbf{H}.
\end{aligned}
$$

Still cannot compute, but can approximate

$$
Pr(\mathbf{w}|\mathbf{X},\sigma^2) \approx \max_{\mathbf{H}} \left[\text{Norm}_{\mathbf{w}}[\mathbf{0},\mathbf{X}^T\mathbf{H}^{-1}\mathbf{X}+\sigma^2\mathbf{I}] \prod_{d=1}^{D} \text{Gam}_{h_d}[\nu/2,\nu/2] \right]
$$

Simon J.D. Prince

Sparse Linear Regression

UNIVERSITÄT BONN

$$Pr(\mathbf{w}|\mathbf{X},\sigma^2) \approx \max_{\mathbf{H}} \left[\text{Norm}_{\mathbf{w}}[\mathbf{0}, \mathbf{X}^T\mathbf{H}^{-1}\mathbf{X} + \sigma^2\mathbf{I}] \prod_{d=1}^{D} \text{Gam}_{h_d}[\nu/2, \nu/2] \right]$$

To fit the model, update variance σ^2 and hidden variables $\{h_d\}$.

- Initialize **H** and σ^2
- Iterate:
  - Update posterior:

$$Pr(\boldsymbol{\phi}|\mathbf{X},\mathbf{w}) = \text{Norm}_{\boldsymbol{\phi}}\left[\frac{1}{\sigma^2}\mathbf{A}^{-1}\mathbf{X}\mathbf{w}, \mathbf{A}^{-1}\right]$$

$$\begin{aligned} \boldsymbol{\mu} &= \frac{1}{\sigma^2}\mathbf{A}^{-1}\mathbf{X}\mathbf{w} \\ \boldsymbol{\Sigma} &= \mathbf{A}^{-1}, \end{aligned} \qquad \mathbf{A} = \frac{1}{\sigma^2}\mathbf{X}\mathbf{X}^T + \mathbf{H}$$

  - Estimate **H**
  - Update posterior
  - Estimate σ^2

Sparse Linear Regression

$$Pr(\mathbf{w}|\mathbf{X},\sigma^2) \approx \max_{\mathbf{H}}\left[\text{Norm}_{\mathbf{w}}[\mathbf{0}, \mathbf{X}^T\mathbf{H}^{-1}\mathbf{X}+\sigma^2\mathbf{I}]\prod_{d=1}^{D}\text{Gam}_{h_d}[\nu/2,\nu/2]\right]$$

To fit the model, update variance σ^2 and hidden variables $\{h_d\}$.

- Estimate H

$$h_d^{new} = \frac{1-h_d\Sigma_{dd}+\nu}{\mu_d^2+\nu}$$

- Estimate σ^2

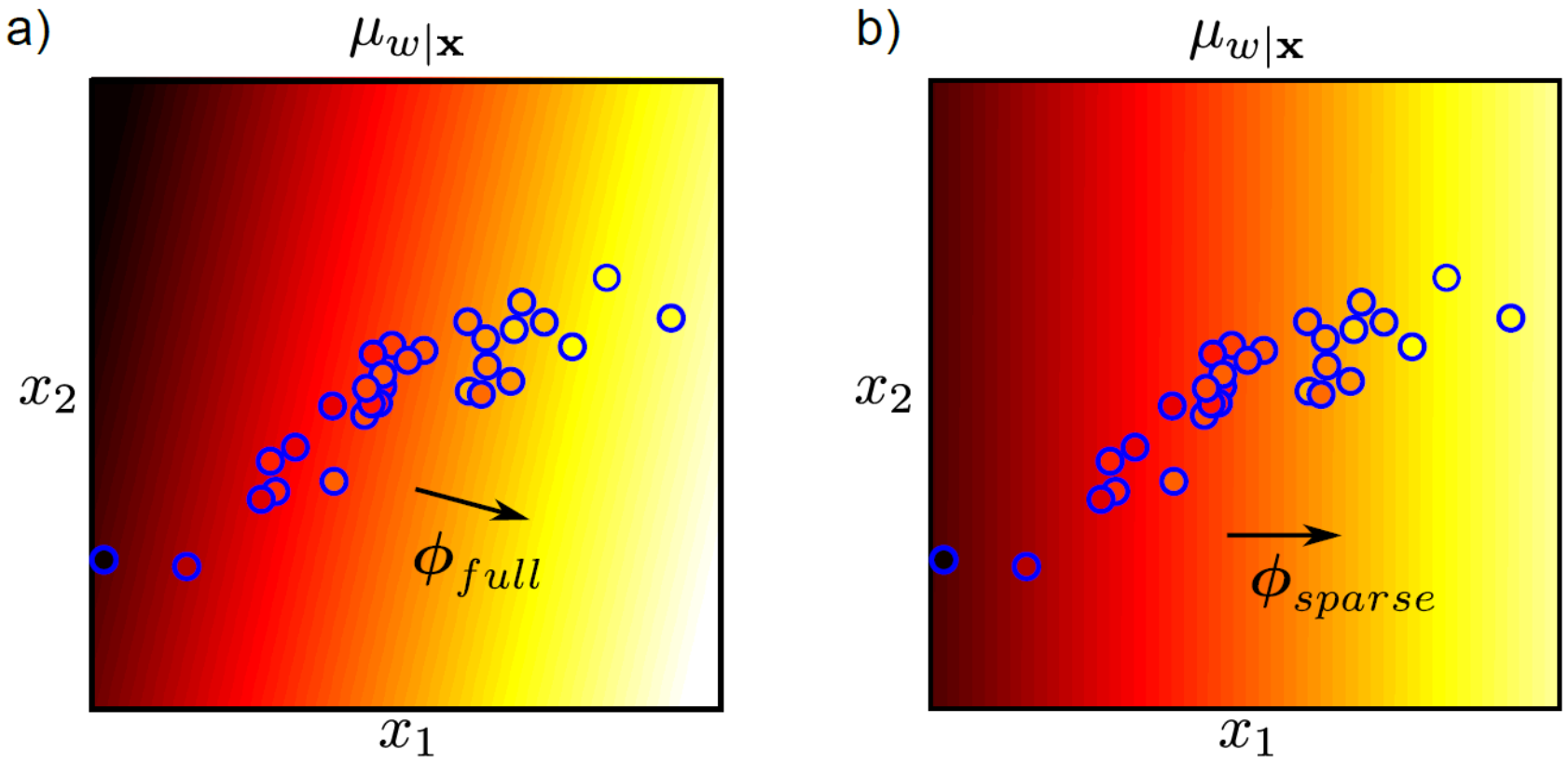
$$(\sigma^2)^{new} = \frac{1}{D-\sum_d(1-h_d\Sigma_{dd})}(\mathbf{w}-\mathbf{X}\boldsymbol{\mu})^T(\mathbf{w}-\mathbf{X}\boldsymbol{\mu})$$

where

$$\boldsymbol{\mu} = \frac{1}{\sigma^2}\mathbf{A}^{-1}\mathbf{X}\mathbf{w} \qquad \mathbf{A} = \frac{1}{\sigma^2}\mathbf{X}\mathbf{X}^T+\mathbf{H}$$
$$\boldsymbol{\Sigma} = \mathbf{A}^{-1},$$

Simon J.D. Prince

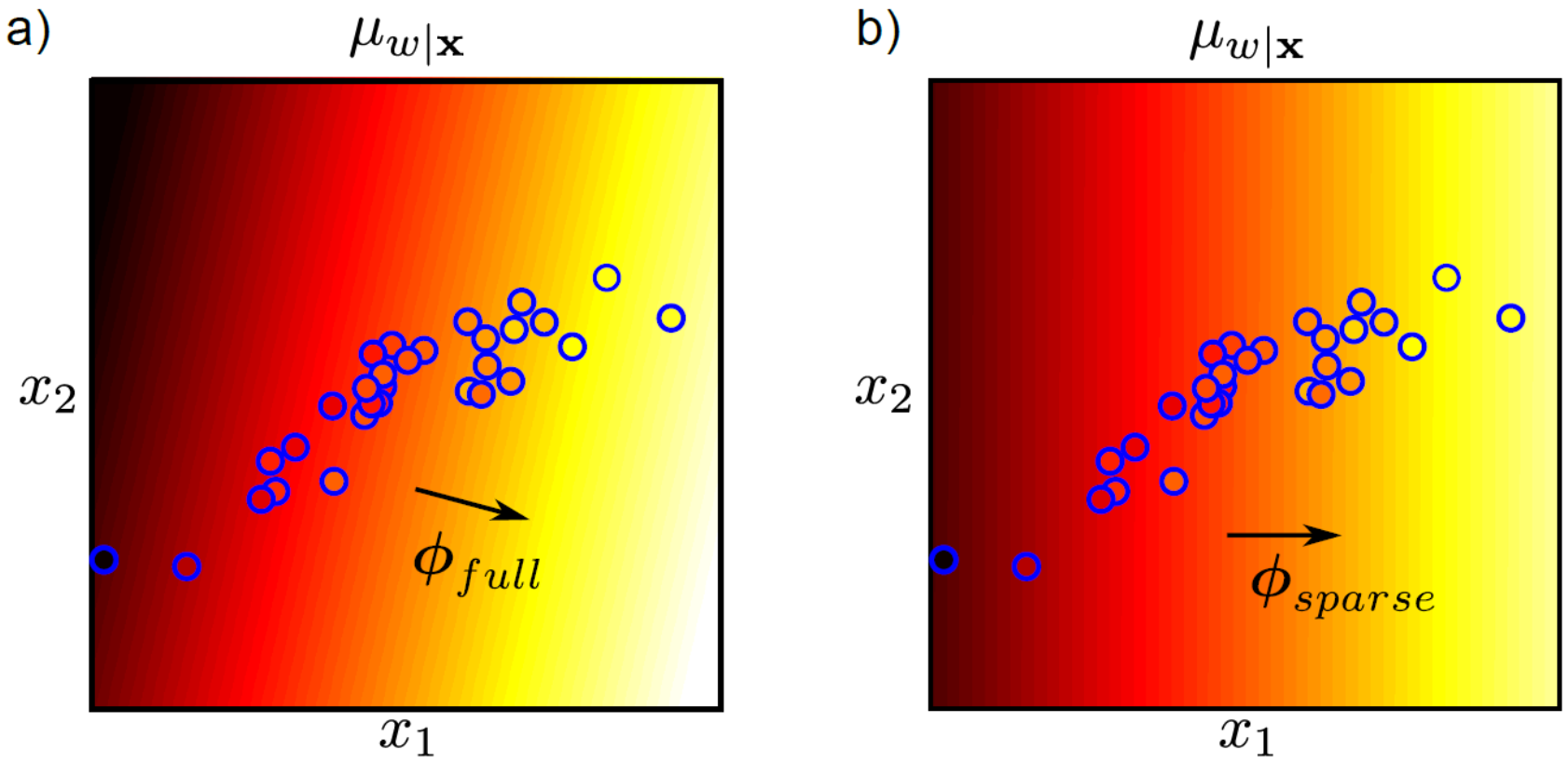
Sparse Linear Regression



After fitting, some of hidden variables become very big, implies prior tightly fitted around zero, can be eliminated from model

Simon J.D. Prince

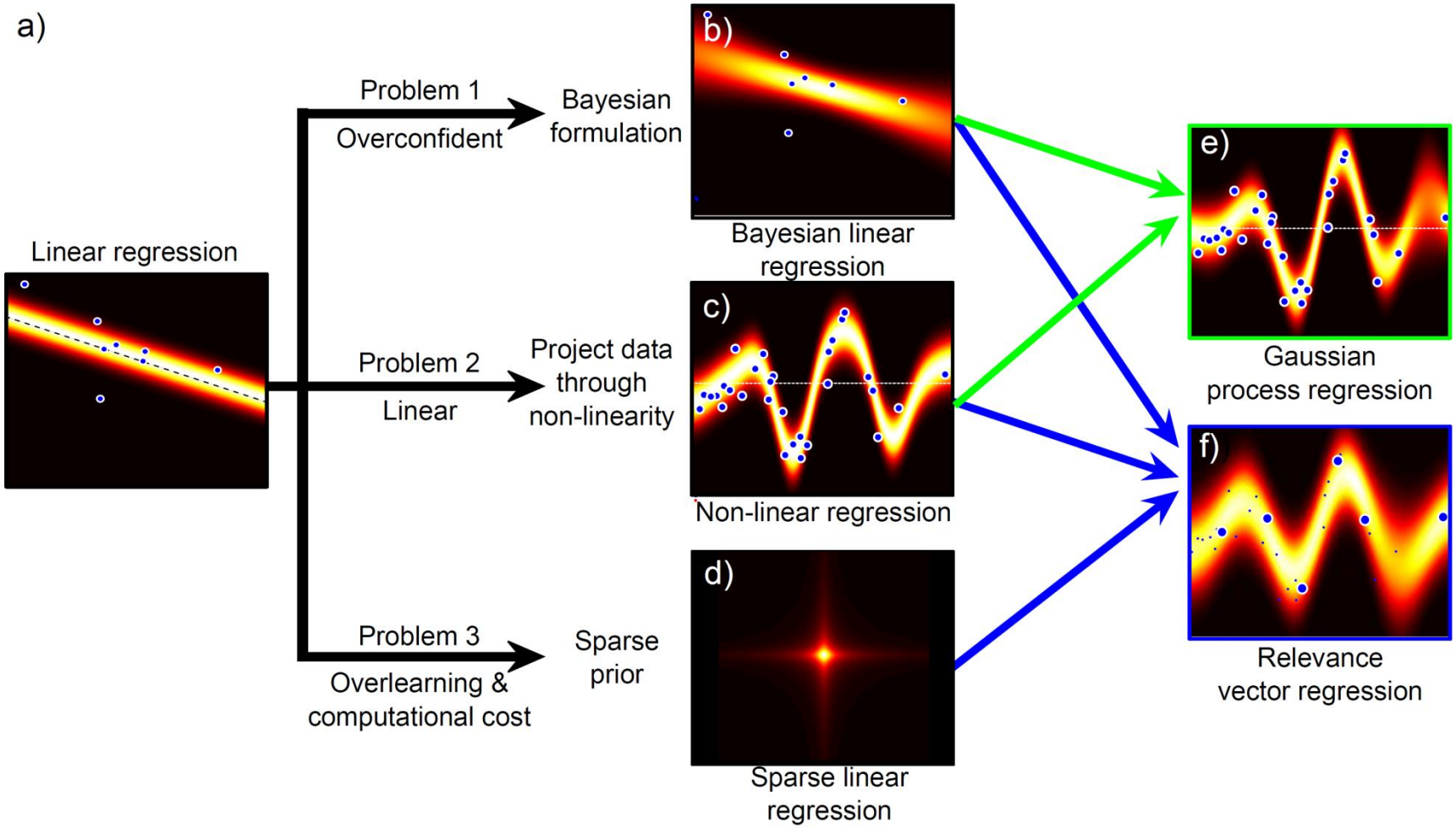
Sparse Linear Regression



Doesn't work for non-linear case as we need one hidden variable per dimension – becomes intractable with high dimensional transformation. To solve this problem, we move to the dual model.

Simon J.D. Prince

Regression Models



Simon J.D. Prince

Dual Linear Regression

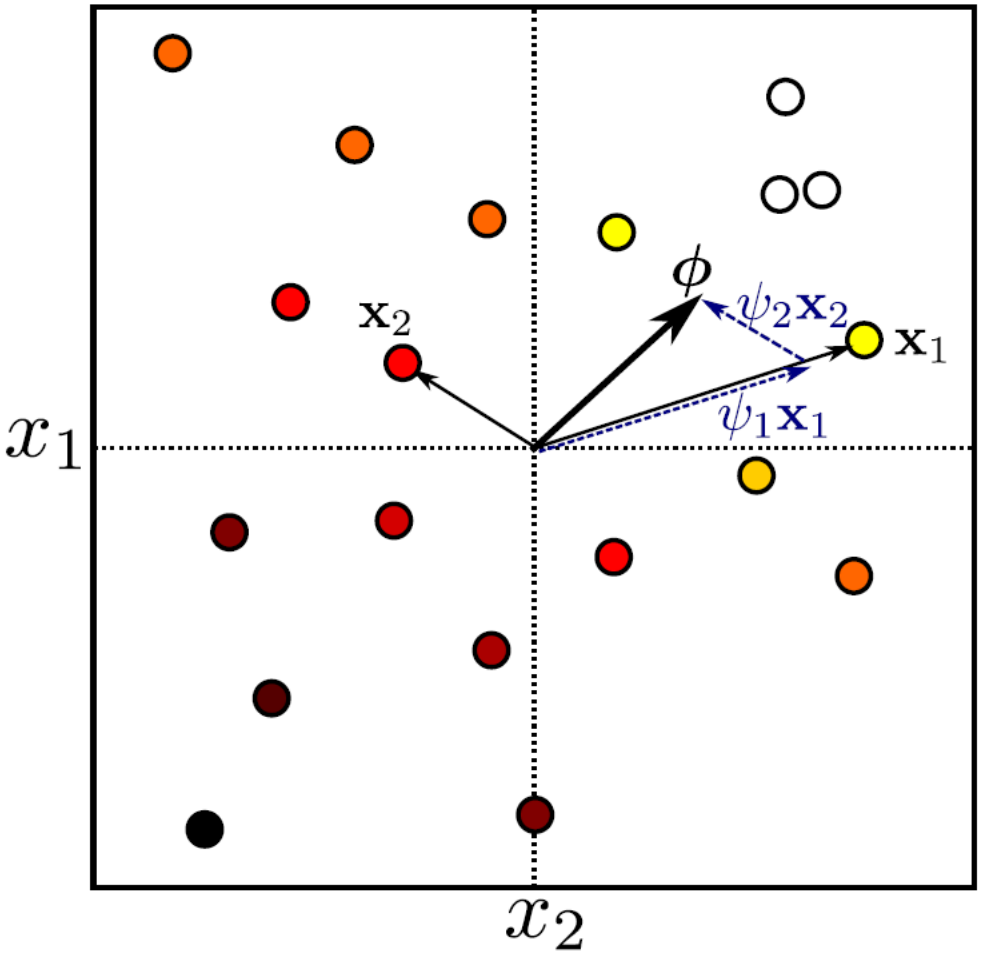
KEY IDEA:

Gradient Φ is just a vector in the data space

Can represent as a weighted sum of the data points

$$\phi = \mathbf{X}\psi$$

Now solve for Ψ. One parameter per training example.



Simon J.D. Prince

Dual Linear Regression

Original linear regression:

$$Pr(\mathbf{w}|\mathbf{X}) = \text{Norm}_{\mathbf{w}}[\mathbf{X}^T\boldsymbol{\phi}, \sigma^2\mathbf{I}]$$

Dual variables:

$$\boldsymbol{\phi} = \mathbf{X}\boldsymbol{\psi}$$

Dual linear regression:

$$Pr(\mathbf{w}|\mathbf{X}, \boldsymbol{\theta}) = \text{Norm}_{\mathbf{w}}\left[\mathbf{X}^T\mathbf{X}\boldsymbol{\psi}, \sigma^2\mathbf{I}\right]$$

Simon J.D. Prince

Maximum likelihood

Maximum likelihood solution:

$$\hat{\boldsymbol{\psi}}, \hat{\sigma}^2 = \underset{\boldsymbol{\psi},\sigma^2}{\operatorname{argmax}} \left[-\frac{I\log[2\pi]}{2} - \frac{I\log[\sigma]}{2} - \frac{(\mathbf{w}-\mathbf{X}^T\mathbf{X}\boldsymbol{\psi})^T(\mathbf{w}-\mathbf{X}^T\mathbf{X}\boldsymbol{\psi})}{2\sigma^2} \right]$$

Dual variables:

$$\begin{aligned} \hat{\boldsymbol{\psi}} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{w} \\ \hat{\sigma}^2 &= \frac{(\mathbf{w}-\mathbf{X}^T\mathbf{X}\boldsymbol{\psi})^T(\mathbf{w}-\mathbf{X}^T\mathbf{X}\boldsymbol{\psi})}{I} \end{aligned}$$

Same result as before:

$$\begin{aligned} \hat{\boldsymbol{\phi}} = \mathbf{X}\hat{\boldsymbol{\psi}} &= \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{w} \\ &= (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{w} \\ &= (\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{X}\mathbf{w}, \end{aligned}$$

Simon J.D. Prince

Bayesian case

$$Pr(\boldsymbol{\psi}) = \text{Norm}_{\boldsymbol{\psi}}[\mathbf{0}, \sigma_p^2\mathbf{I}]$$

$$Pr(\mathbf{w}|\mathbf{X}, \boldsymbol{\theta}) = \text{Norm}_{\mathbf{w}}\left[\mathbf{X}^T\mathbf{X}\boldsymbol{\psi}, \sigma^2\mathbf{I}\right]$$

Compute distribution over parameters:

$$Pr(\boldsymbol{\psi}|\mathbf{X}, \mathbf{w}, \sigma^2) = \frac{Pr(\mathbf{w}|\mathbf{X}, \boldsymbol{\psi}, \sigma^2)Pr(\boldsymbol{\psi})}{Pr(\mathbf{w}|\mathbf{X}, \sigma^2)}$$

Gives result:

$$Pr(\boldsymbol{\psi}|\mathbf{X}, \mathbf{w}, \sigma^2) = \text{Norm}_{\boldsymbol{\psi}}\left[\frac{1}{\sigma^2}\mathbf{A}^{-1}\mathbf{X}^T\mathbf{X}\mathbf{w}, \mathbf{A}^{-1}\right]$$

where

$$\mathbf{A} = \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}\mathbf{X}^T\mathbf{X} + \frac{1}{\sigma_p^2}\mathbf{I}$$

Simon J.D. Prince

Bayesian case

Predictive distribution:

$$\begin{aligned} Pr(w^*|\mathbf{x}^*, \mathbf{X}, \mathbf{w}) &= \int Pr(w^*|\mathbf{x}^*, \boldsymbol{\psi}) Pr(\boldsymbol{\psi}|\mathbf{X}, \mathbf{w})\, d\boldsymbol{\psi} \\ &= \text{Norm}_{w^*}\left[\frac{1}{\sigma^2}\mathbf{x}^{*T}\mathbf{X}\mathbf{A}^{-1}\mathbf{X}^T\mathbf{X}\mathbf{w}\ , \mathbf{x}^{*T}\mathbf{X}\mathbf{A}^{-1}\mathbf{X}^T\mathbf{x}^* + \sigma^2\right] \end{aligned}$$

where:

$$\mathbf{A} = \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}\mathbf{X}^T\mathbf{X} + \frac{1}{\sigma_p^2}\mathbf{I}$$

Notice that in both the maximum likelihood and Bayesian case depend on dot products $\mathbf{X}^T\mathbf{X}$. Can be kernelized!

Simon J.D. Prince

Bayesian case (previous formulation)

UNIVERSITÄT BONN

Predictive distribution:

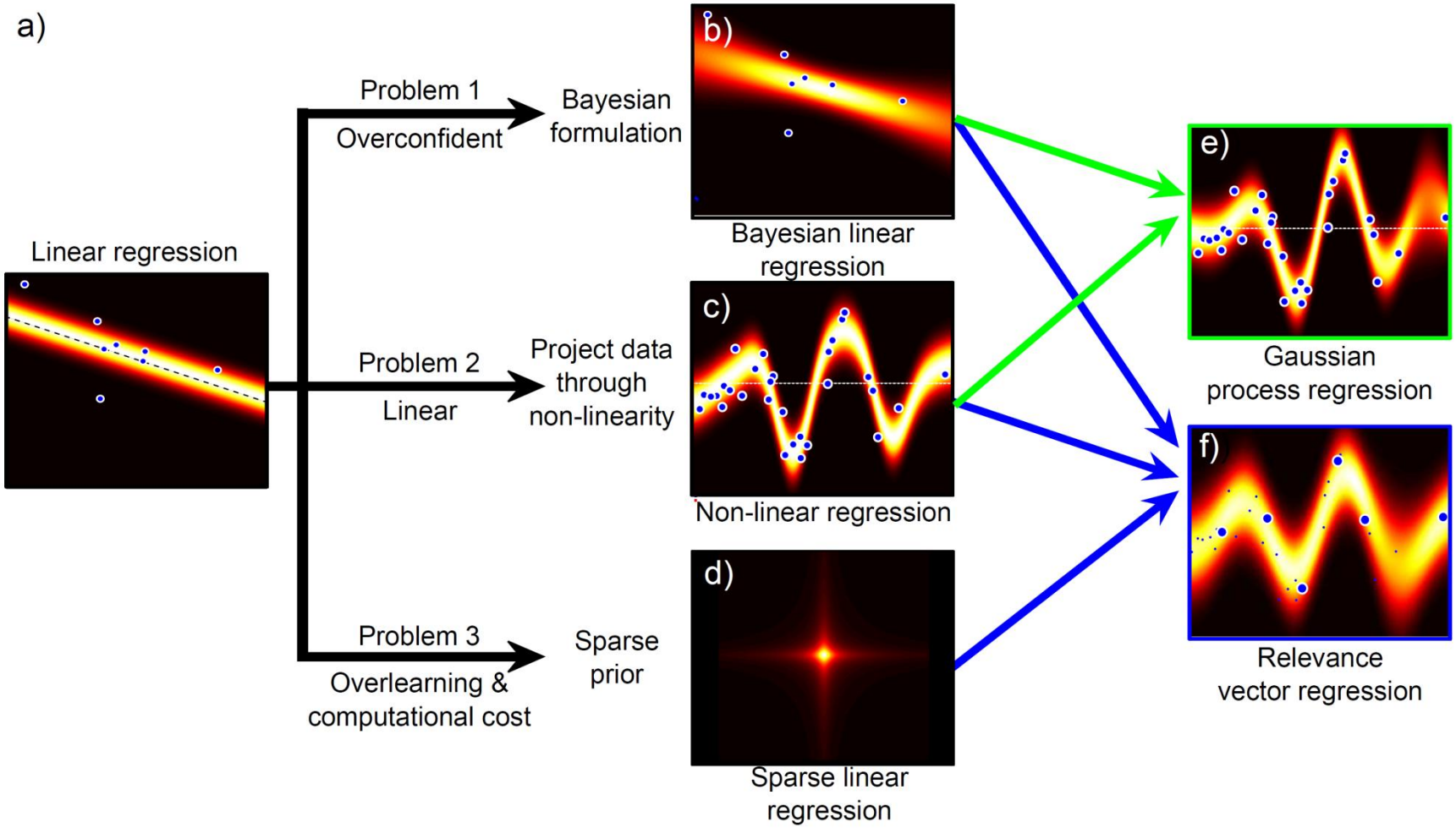
$$
\begin{aligned}
Pr(w^*|\mathbf{x}^*, \mathbf{X}, \mathbf{w}) &= \int Pr(w^*|\mathbf{x}^*, \boldsymbol{\phi}) Pr(\boldsymbol{\phi}|\mathbf{X}, \mathbf{w}) d\boldsymbol{\phi} \\
&= \int \mathrm{Norm}_{w^*}[\boldsymbol{\phi}^T \mathbf{x}^*, \sigma^2] \mathrm{Norm}_{\boldsymbol{\phi}}\left[\frac{1}{\sigma^2}\mathbf{A}^{-1}\mathbf{X}\mathbf{w}, \mathbf{A}^{-1}\right] d\boldsymbol{\phi} \\
&= \mathrm{Norm}_{w^*}\left[\frac{1}{\sigma^2}\mathbf{x}^{*T}\mathbf{A}^{-1}\mathbf{X}\mathbf{w}, \mathbf{x}^{*T}\mathbf{A}^{-1}\mathbf{x}^* + \sigma^2\right].
\end{aligned}
$$

where:

$$
\mathbf{A} = \frac{1}{\sigma^2}\mathbf{X}\mathbf{X}^T + \frac{1}{\sigma_p^2}\mathbf{I}
$$

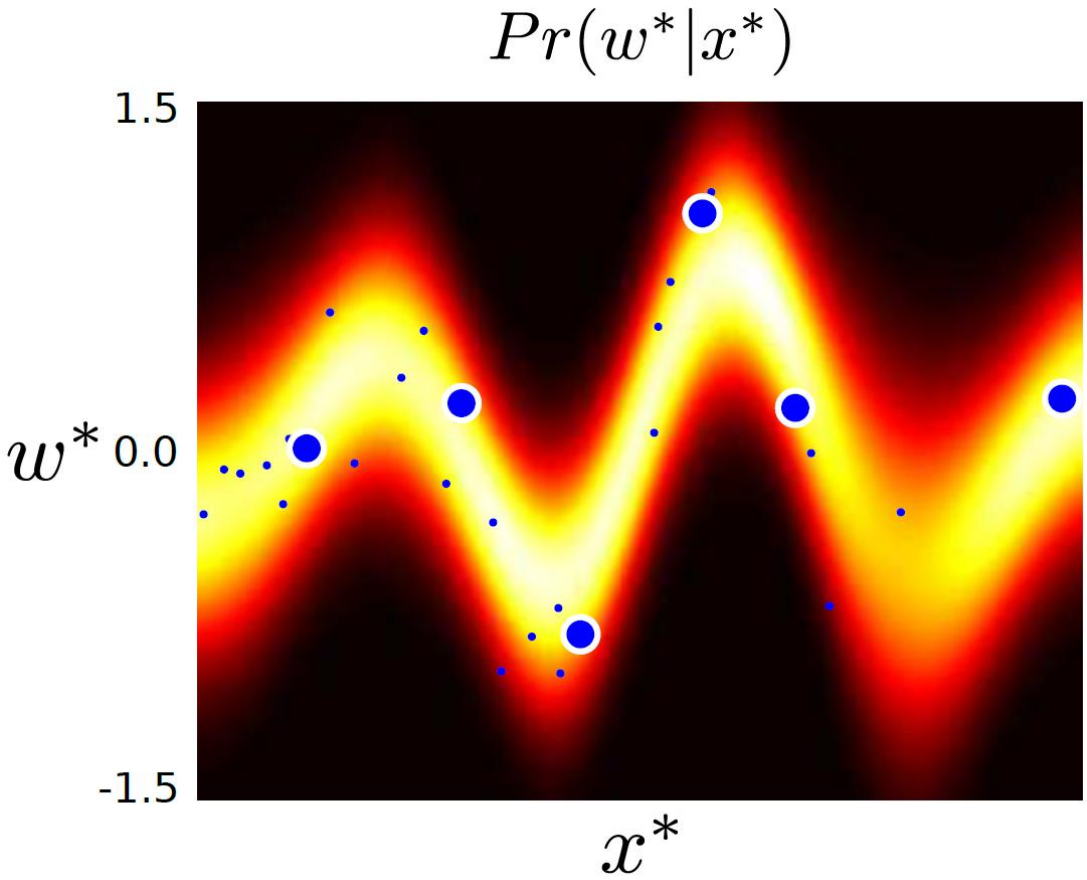
Simon J.D. Prince

Regression Models



Simon J.D. Prince

Relevance Vector Machine



Combines ideas of

- Dual regression (1 parameter per training example)
- Sparsity (most of the parameters are zero)

Model depends only sparsely on training data.

Simon J.D. Prince

Relevance Vector Machine

$$Pr(\mathbf{w}|\mathbf{X},\boldsymbol{\theta}) = \text{Norm}_{\mathbf{w}}\left[\mathbf{X}^T\mathbf{X}\boldsymbol{\psi}, \sigma^2\mathbf{I}\right]$$
$$Pr(\boldsymbol{\psi}) = \prod_{i=1}^{I} \text{Stud}_{\boldsymbol{\psi}_i}[0,1,\nu]$$

Using same approximations as for sparse model we get the problem:

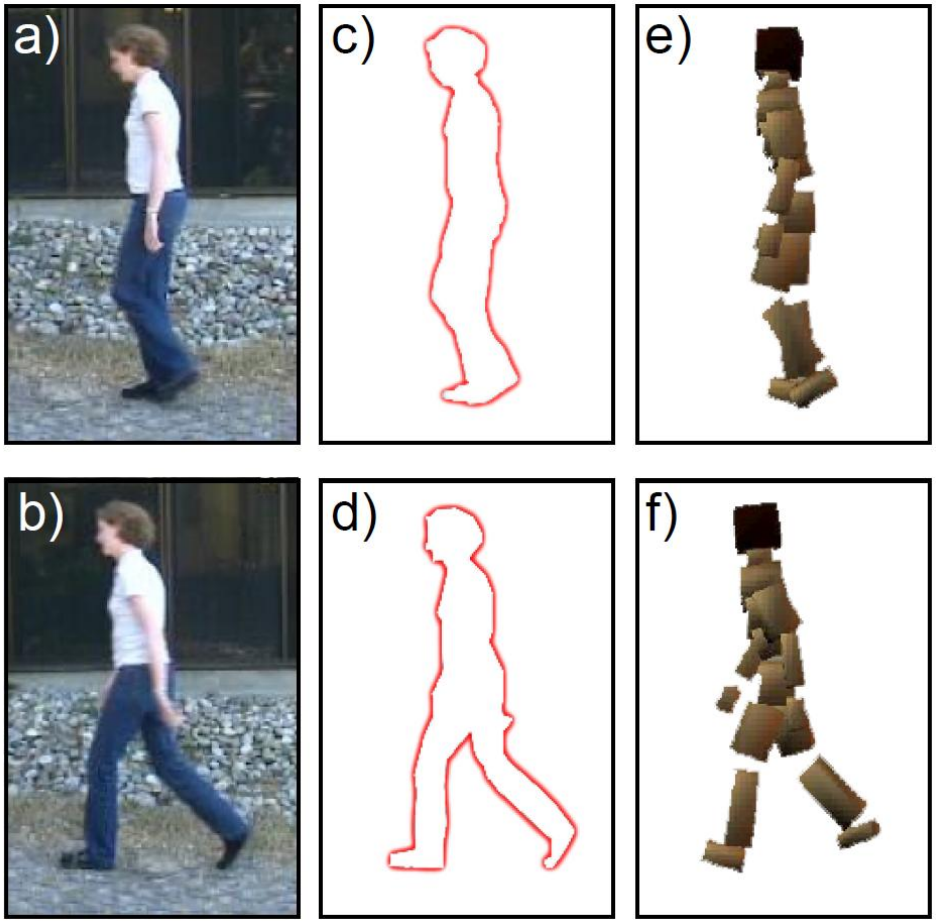
$$Pr(\mathbf{w}|\mathbf{X},\sigma^2) \approx \max_{\mathbf{H}}\left[\text{Norm}_{\mathbf{w}}[\mathbf{0}, \mathbf{X}^T\mathbf{X}\mathbf{H}^{-1}\mathbf{X}^T\mathbf{X} + \sigma^2\mathbf{I}]\prod_{i=1}^{I}\text{Gam}_{h_i}[\nu/2,\nu/2]\right]$$

To solve, update variance σ^2 and hidden variables $\{h_d\}$ alternately.

Notice that this only depends on dot-products and so can be kernelized

Simon J.D. Prince

Body Pose Regression

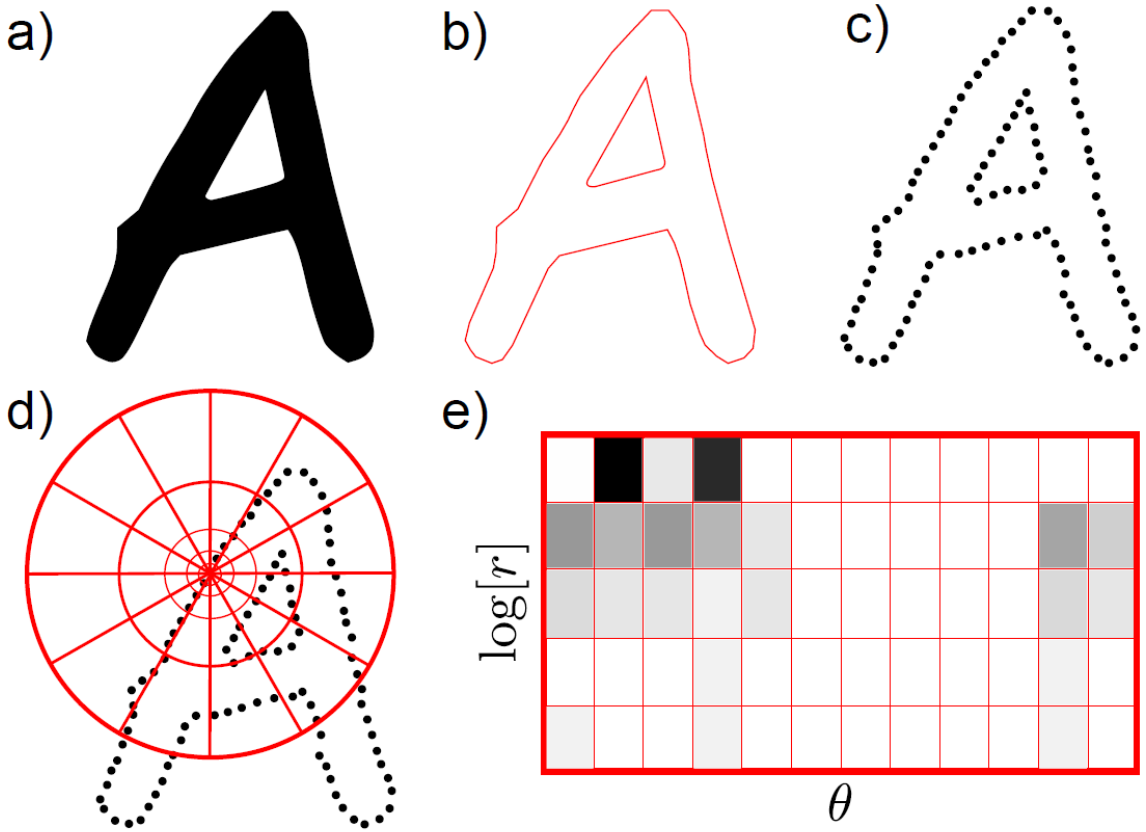


Encode silhouette as 100x1 vector, encode body pose as 55 x1 vector. Learn relationship

[A. Agarwal and B. Triggs. **3D Human Pose from Silhouettes by Relevance Vector Regression**. CVPR 2004]

Simon J.D. Prince

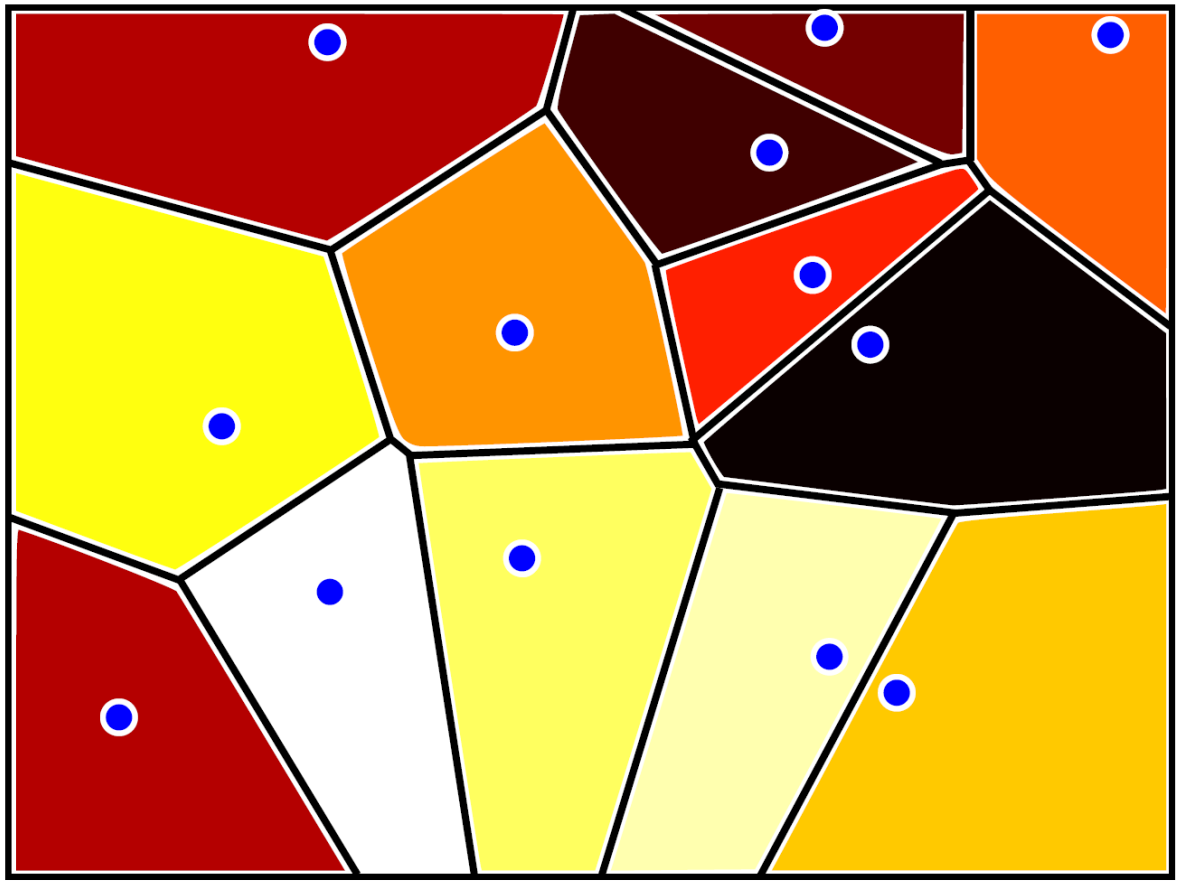
Shape Context



Returns 60 x 1 vector for each of 400 points around the silhouette

Simon J.D. Prince

Codebook



Cluster 60D space (based on all training data) into 100 vectors
Assign each 60x1 vector to closest cluster (Voronoi partition)
Final data vector is 100x1 histogram over distribution of assignments

Simon J.D. Prince

Results

- 2636 training examples, solution depends on only 6% of these
- 6 degree average error

Simon J.D. Prince


UNIVERSITÄT BONN