Evidence maximisation, α marginalisation, priors

H. Koeppl

Christian Doppler Laboratory for Nonlinear Signal Processing,

Graz University of Technology, Austria

Talk overview

- Introduction, maximum likelihood estimation (MLE)
- MLE $\longrightarrow$ Bayesian-, MAP-estimate
- Bayesian estimation, regularisation
- Hierarchical Bayesian models
- Approximative scheme: α-marginalisation
- Approximative scheme: Evidence procedure
- Illustrative example (Mathematica)
- Automatic relevance determination
- Priors
- Conclusion

MLE review through example [1|5]

Suppose one measures a noisy output z of a functional relation $f : \mathbb{R}^n \rightarrow \mathbb{R}$, $y = f(\boldsymbol{x})$. We assume:

$$z = y + \epsilon$$

Goal: reconstruct the functional relation from $\{\boldsymbol{x}, z\}$
Synonyms: interpolation, nonlinear regression, supervised learning
Ansatz for $f(\boldsymbol{x})$ (no model missmatch):

$$y = \hat{f}(\boldsymbol{x}, \boldsymbol{w}) = \sum_{i=1}^{M-1} w_i K(\boldsymbol{c}_i, \boldsymbol{x}) + w_0$$

MLE review through example [2|5]

Why this model class?

AREAS OF APPLICATION:

- Straight line fitting, curve fitting
- Multivariate linear regression
- Discrete-time integral models (e.g. FIR)
- Kernel regression methods (e.g. RBF networks)
- Regression using orthogonal functions

MLE review through example [3|5]

Assume Gaussian noise

$$\epsilon \sim \mathcal{N}(0, \sigma^2) \text{ with } \sigma^2 \text{known}$$

$$p(\epsilon) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(z-y)^2}{2\sigma^2}\right]$$

Likelihood function

$$l(\boldsymbol{w}|z) \equiv p(z|\boldsymbol{w}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(z-\hat{f}(\boldsymbol{x},\boldsymbol{w}))^2}{2\sigma^2}\right]$$

MLE review through example [4|5]

More measurements $\{\boldsymbol{x}_i, z_i\}_{k=1}^{N}$,
White noise:

$$p(z_1, \ldots, z_N|\boldsymbol{w}) = \prod_{i=1}^{N} p(z_i|\boldsymbol{w})$$

Thus

$$p(\boldsymbol{z}|\boldsymbol{w}) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left[-\frac{\sum_{i=1}^{N}(z_i - \hat{f}(\boldsymbol{x}_i, \boldsymbol{w}))^2}{2\sigma^2}\right]$$

MLE:

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmax}}\, p(\boldsymbol{z}|\boldsymbol{w}) = \underset{\boldsymbol{w}}{\operatorname{argmin}}(-\ln p(\boldsymbol{z}|\boldsymbol{w}))$$

MLE review through example [5|5]

Linear problem:

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmin}} \|\boldsymbol{z} - \boldsymbol{\Phi}\boldsymbol{w}\|^2 \text{ with } \boldsymbol{\Phi} \in \mathbb{R}^{N \times M}$$

Maximum likelihood estimate:

$$\hat{\boldsymbol{w}}_{MLE} = (\boldsymbol{\Phi}^T \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T \boldsymbol{z}$$

MLE $\longrightarrow$ Bayes, MAP [1|3]

$$p(\boldsymbol{z}|\boldsymbol{w}) = \frac{p(\boldsymbol{z},\boldsymbol{w})}{p(\boldsymbol{w})} = \frac{p(\boldsymbol{w}|\boldsymbol{z})p(\boldsymbol{z})}{p(\boldsymbol{w})}$$

Thus

$$p(\boldsymbol{w}|\boldsymbol{z}) = \frac{p(\boldsymbol{z}|\boldsymbol{w})p(\boldsymbol{w})}{p(\boldsymbol{z})}$$

Posterior $\propto$ Likelihood $\times$ Prior

Bayesian estimate:

$$\hat{\boldsymbol{w}} = \mathsf{E}_{\boldsymbol{w}|\boldsymbol{z}}[\boldsymbol{w}] = \int_{\boldsymbol{W}} \boldsymbol{w}p(\boldsymbol{w}|\boldsymbol{z})d\boldsymbol{w}$$

MLE $\longrightarrow$ Bayes, MAP [2|3]

When does a posteriori and likelihood coincide?

$$p(\boldsymbol{w}|\boldsymbol{z}) = p(\boldsymbol{z}|\boldsymbol{w}) \text{ if } p(\boldsymbol{w})/p(\boldsymbol{z}) = 1$$

- $p(\boldsymbol{z})$ is not a function of $\boldsymbol{w}$ thus $p(\boldsymbol{w}) = const$.
- flat prior distribution $\rightarrow$ no preferences

What is MAP ?

Maximum a posteriori estimate:

$$\hat{\boldsymbol{w}}_{MAP} = \underset{\boldsymbol{w}}{\operatorname{argmax}}\, p(\boldsymbol{w}|\boldsymbol{z})$$

MLE $\longrightarrow$ Bayes, MAP [3|3]

When does a posteriori estimate and MAP coincide?
$p(\boldsymbol{w}|\boldsymbol{z})$ is unimodal and symmetric

When does a posteriori estimate and MLE coincide?

flat prior + $p(\boldsymbol{w}|\boldsymbol{z})$ is unimodal and symmetric

Our first Bayesian model [1|5]

$$p(\boldsymbol{w}|\boldsymbol{z}) = \frac{p(\boldsymbol{z}|\boldsymbol{w})p(\boldsymbol{w})}{\int_{\boldsymbol{W}} p(\boldsymbol{z}|\boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w}}$$

All we need is to specify a prior $p(\boldsymbol{w})$
One choice:

$$p(\boldsymbol{w}) = (2\pi/\alpha)^{-\frac{M}{2}} \exp(-\frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w})$$

α is a known and fixed parameter

Our first Bayesian model [2|5]

How to compute the posterior $p(\boldsymbol{w}|\boldsymbol{z})$?
Can we do the normalisation integral? (Yes, but ...)

$$\int_{\boldsymbol{W}} p(\boldsymbol{z}|\boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w} = C \int_{\mathbb{R}^M} \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{z} - \boldsymbol{\Phi}\boldsymbol{w}\|^2 - \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w}\right) d\boldsymbol{w}$$

First check the integrand. What is the minimiser of

$$\frac{1}{2\sigma^2}(\boldsymbol{z} - \boldsymbol{\Phi}\boldsymbol{w})^T(\boldsymbol{z} - \boldsymbol{\Phi}\boldsymbol{w}) + \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w}$$

$$\frac{\partial}{\partial \boldsymbol{w}} : \quad -\frac{1}{\sigma^2}\boldsymbol{\Phi}^T(\boldsymbol{z} - \boldsymbol{\Phi}\boldsymbol{w}) + \alpha\boldsymbol{w}$$

Our first Bayesian model [3|5]

STILL: How to compute the posterior $p(\boldsymbol{w}|\boldsymbol{z})$?

$$\underbrace{\left(\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \alpha\sigma^2\boldsymbol{I}\right)}_{\equiv\boldsymbol{\Sigma}}\boldsymbol{w}_m = \boldsymbol{\Phi}^T\boldsymbol{z}$$

$$\boldsymbol{w}_m = (\boldsymbol{\Phi}^T\boldsymbol{\Phi} + \alpha\sigma^2\boldsymbol{I})^{-1}\boldsymbol{\Phi}^T\boldsymbol{z}$$

We now try to rewrite the exponent of the integrand

$$\frac{1}{2\sigma^2}(\boldsymbol{z} - \boldsymbol{\Phi}\boldsymbol{w})^T(\boldsymbol{z} - \boldsymbol{\Phi}\boldsymbol{w}) + \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w}$$

$$\frac{1}{2\sigma^2}(\boldsymbol{z}^T\boldsymbol{z} - 2\boldsymbol{w}^T\boldsymbol{\Phi}^T\boldsymbol{z} + \boldsymbol{w}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\boldsymbol{w}) + \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} + \frac{1}{2\sigma^2}\boldsymbol{w}_m^T\boldsymbol{\Sigma}\boldsymbol{w}_m - \frac{1}{2\sigma^2}\boldsymbol{w}_m^T\boldsymbol{\Sigma}\boldsymbol{w}$$

Our first Bayesian model [4|5]

STILL: How to compute the posterior $p(\boldsymbol{w}|\boldsymbol{z})$?

$$\frac{1}{2\sigma^2}(\boldsymbol{z} - \boldsymbol{\Phi}\boldsymbol{w})^T(\boldsymbol{z} - \boldsymbol{\Phi}\boldsymbol{w}) + \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w}$$

can be written as

$$\frac{1}{2\sigma^2}(\boldsymbol{z}^T\boldsymbol{z} - \boldsymbol{w}_m^T\boldsymbol{\Sigma}\boldsymbol{w}_m) + \frac{1}{2\sigma^2}(\boldsymbol{w} - \boldsymbol{w}_m)^T\boldsymbol{\Sigma}(\boldsymbol{w} - \boldsymbol{w}_m)$$

Thus

$$p(\boldsymbol{w}|\boldsymbol{z}) \propto p(\boldsymbol{z}|\boldsymbol{w})p(\boldsymbol{w}) \propto \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{w} - \boldsymbol{w}_m)^T\boldsymbol{\Sigma}(\boldsymbol{w} - \boldsymbol{w}_m)\right)$$

Our first Bayesian model [5|5]

STILL: How to compute the posterior $p(\boldsymbol{w}|\boldsymbol{z})$?

$$p(\boldsymbol{w}|\boldsymbol{z}) \propto \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{w}-\boldsymbol{w}_m)^T\boldsymbol{\Sigma}(\boldsymbol{w}-\boldsymbol{w}_m)\right)$$

- Multivariate Gaussian with covariance $\boldsymbol{\Sigma}^{-1}$
- Centered at $\boldsymbol{w}_m$ thus $\boldsymbol{w}_m \equiv \boldsymbol{w}_{MAP}$
- and $\mathrm{E}_{\boldsymbol{w}|\boldsymbol{z}}[\boldsymbol{w}] = \int_{\boldsymbol{W}} \boldsymbol{w} p(\boldsymbol{w}|\boldsymbol{z}) d\boldsymbol{w} \equiv \boldsymbol{w}_{MAP}$

What to take home?

The regularised linear least squares problem (Tikhonov)

$$\min_{\boldsymbol{w}} \left\{ \|\boldsymbol{z} - \boldsymbol{\Phi}\boldsymbol{w}\|^2 + \lambda \|\boldsymbol{w}\|^2 \right\}$$

with the solution

$$\hat{\boldsymbol{w}} = (\boldsymbol{\Phi}^T \boldsymbol{\Phi} + \lambda \boldsymbol{I})^{-1} \boldsymbol{\Phi}^T \boldsymbol{z}$$

can be seen as the Bayesian problem with Gaussian likelihood and Gaussian prior, where $\lambda = \alpha\sigma^2$.

Extending our model [1|4]

What if we do not known σ^2 and α ?
Bayesian answer: define prior distributions $p(\sigma^2)$, $p(\alpha)$
Our old $\boldsymbol{w}$-prior

$$p(\boldsymbol{w}) = (2\pi/\alpha)^{-\frac{M}{2}} \exp(-\frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w})$$

gets

$$p(\boldsymbol{w}|\alpha) = (2\pi/\alpha)^{-\frac{M}{2}} \exp(-\frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w})$$

and

$$p(\boldsymbol{w}, \alpha) = p(\boldsymbol{w}|\alpha)p(\alpha)$$

Extending our model [2|4]

Marginalisation of $p(\boldsymbol{w}, \alpha)$ with respect to α is our new $\boldsymbol{w}$-prior

$$p(\boldsymbol{w}) = \int p(\boldsymbol{w}|\alpha)p(\alpha)d\alpha$$

What about the noise variance σ^2 ?
The old likelihood function $p(\boldsymbol{z}|\boldsymbol{w})$ gets

$$p(\boldsymbol{z}|\boldsymbol{w}, \sigma^2)$$

and

$$p(\boldsymbol{z}, \sigma^2|\boldsymbol{w}) = p(\boldsymbol{z}|\boldsymbol{w}, \sigma^2)p(\sigma^2)$$

Extending our model [3|4]

Marginalisation of $p(\boldsymbol{z}, \sigma^2|\boldsymbol{w})$ with respect to σ^2 is our new likelihood

$$p(\boldsymbol{z}|\boldsymbol{w}) = \int p(\boldsymbol{z}|\boldsymbol{w}, \sigma^2)p(\sigma^2)d\sigma^2$$

The posterior computes again to

$$p(\boldsymbol{w}|\boldsymbol{z}) = \frac{p(\boldsymbol{z}|\boldsymbol{w})p(\boldsymbol{w})}{\int_{\boldsymbol{W}} p(\boldsymbol{z}|\boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w}}$$

where likelihood and $\boldsymbol{w}$-prior are in general non-Gaussian.
THE POSTERIOR IS IN GENERAL NOT GAUSSIAN
Even if can find the posterior analytically we have to compute

$$\mathsf{E}_{\boldsymbol{w}|\boldsymbol{z}}[\boldsymbol{w}]$$

Extending our model [4|4]

INTERMEDIATE SUMMARY: Perfect Bayesian estimation

- Define a likelihood function from the model with a parameter σ^2 ("hyperparameter", "nuisance-parameter")
- Define a prior over weights with the hyperparameter α
- Integrate over σ^2 to get the actual likelihood
- Integrate over α to get the actual prior
- Compute the normalisation integral of likelihood $\times$ prior
- Get the posterior
- Compute the expectation for weights $\hat{\boldsymbol{w}} = \mathsf{E}_{\boldsymbol{w}|\boldsymbol{z}}[\boldsymbol{w}]$

MODELS WITH HYPERPRIORS: HIERARCHICAL BAYES

Approximate Solutions [1|2]

Suppose we choose this new $p(\alpha)$ and $p(\sigma^2)$ such that

- we can do the integrals for true likelihood and $\boldsymbol{w}$-prior
- we can find the posteriori analytically
- we can not integrate the posterior for $\mathrm{E}_{\boldsymbol{w}|\boldsymbol{z}}[\boldsymbol{w}]$

One Solution:

- Find $\boldsymbol{w}_{MAP}$s for posterior $p(\boldsymbol{w}|\boldsymbol{z})$ (in general multimodal)
- approximate at each $\boldsymbol{w}_{MAP}$ a Gaussian
- decide heuristically which one is right

[Buntine, Weigend 1994] ("MAP method" for further reference)

"α-marginalisation approximation"

Approximate Solutions [2|2]

Disadvantages of MAP method:

- Always goes for the peak
- Does not care about the probability mass
- This discrepancy gets amplified for high dimensions of $\boldsymbol{w}$-space. (usual)

DIFFERENT APPROACH:
"Evidence procedure", "type II maximum likelihood"

- Find good values for σ^2 and α
- Freeze values of the hyperparameters
- Posterior is given by our first simple model: GAUSSIAN
- How: Maximise the evidence of the hyperparameters given the data: $\max_{\alpha,\sigma^2} p(\alpha, \sigma^2|\boldsymbol{z})$

The Evidence procedure [1|8]

Expand the posterior

$$p(\boldsymbol{w}|\boldsymbol{z},\alpha,\sigma^2) = \frac{p(\boldsymbol{w},\boldsymbol{z},\alpha,\sigma^2)}{p(\boldsymbol{z},\alpha,\sigma^2)} = \frac{p(\boldsymbol{z}|\boldsymbol{w},\alpha,\sigma^2)p(\boldsymbol{w},\alpha,\sigma^2)}{p(\boldsymbol{z}|\alpha,\sigma^2)p(\alpha,\sigma^2)}$$

$$= \frac{p(\boldsymbol{z}|\boldsymbol{w},\alpha,\sigma^2)p(\boldsymbol{w}|\alpha,\sigma^2)p(\alpha,\sigma^2)}{p(\boldsymbol{z}|\alpha,\sigma^2)p(\alpha,\sigma^2)} = \frac{p(\boldsymbol{z}|\boldsymbol{w},\sigma^2)p(\boldsymbol{w}|\alpha)}{p(\boldsymbol{z}|\alpha,\sigma^2)}$$

Suppose the evidence procedure gives us values α_{ev} and σ^2_{ev}: The posterior is then given

$$p(\boldsymbol{w}|\boldsymbol{z},\alpha_{ev},\sigma^2_{ev}) = \frac{p(\boldsymbol{z}|\boldsymbol{w},\sigma^2_{ev})p(\boldsymbol{w}|\alpha_{ev})}{p(\boldsymbol{z}|\alpha_{ev},\sigma^2_{ev})}$$

How does it relate to the correct posterior, with marginalised hyperparameters?

The Evidence procedure [2|8]

How does it relate to the correct posterior, with marginalised hyperparameters?

Consider the marginalisation:

$$p(\boldsymbol{w}, \boldsymbol{z}) = \int p(\boldsymbol{w}, \boldsymbol{z}, \alpha, \sigma^2) d\alpha d\sigma^2$$

then

$$\begin{aligned} p(\boldsymbol{w}|\boldsymbol{z}) &= \frac{1}{p(z)} \int p(\boldsymbol{w}|\boldsymbol{z}, \alpha, \sigma^2) p(\boldsymbol{z}, \alpha, \sigma^2) d\alpha d\sigma^2 \\ &= \frac{1}{p(z)} \int p(\boldsymbol{w}|\boldsymbol{z}, \alpha, \sigma^2) p(\alpha, \sigma^2|\boldsymbol{z}) p(\boldsymbol{z}) d\alpha d\sigma^2 \\ &= \int p(\boldsymbol{w}|\boldsymbol{z}, \alpha, \sigma^2) p(\alpha, \sigma^2|\boldsymbol{z}) d\alpha d\sigma^2 \end{aligned}$$

The Evidence procedure [3|8]

$p(\alpha, \sigma^2 | \boldsymbol{z})$ is "the evidence for α, σ^2 in the data"
If $p(\alpha, \sigma^2 | \boldsymbol{z})$ is peaked at α_{ev}, σ^2_{ev} (ideally $\delta(\alpha_{ev}, \sigma^2_{ev})$)

$$p(\boldsymbol{w}|\boldsymbol{z}) \approx p(\boldsymbol{w}|\boldsymbol{z}, \alpha_{ev}, \sigma^2_{ev})$$

How to find the peak of $p(\alpha, \sigma^2 | \boldsymbol{z})$ for our model?

$$p(\alpha, \sigma^2 | \boldsymbol{z}) \propto p(\boldsymbol{z}|\alpha, \sigma^2) p(\alpha) p(\sigma^2)$$

If we assume flat hyperpriors we only have to find the maximum of

$$p(\boldsymbol{z}|\alpha, \sigma^2) = \int_{\boldsymbol{W}} p(\boldsymbol{z}|\boldsymbol{w}, \sigma^2) p(\boldsymbol{w}|\boldsymbol{\alpha}) d\boldsymbol{w}$$

The Evidence procedure [4|8]

I have seen you around...

$$p(\boldsymbol{z}|\boldsymbol{w},\sigma^2) = (2\pi\sigma^2)^{-\frac{N}{2}} \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{z} - \boldsymbol{\Phi}\boldsymbol{w}\|^2\right)$$

and

$$p(\boldsymbol{w}|\alpha) = (2\pi/\alpha)^{-\frac{M}{2}} \exp(-\frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w})$$

That is just the normalisation integral of the simple model! (fixed α, σ^2)

$$p(\boldsymbol{z}|\boldsymbol{w},\sigma^2)p(\boldsymbol{w}|\alpha) = \frac{1}{C}e^{-\frac{1}{2\sigma^2}(\boldsymbol{z}^T\boldsymbol{z}-\boldsymbol{w}_m^T\boldsymbol{\Sigma}\boldsymbol{w}_m+(\boldsymbol{w}-\boldsymbol{w}_m)^T\boldsymbol{\Sigma}(\boldsymbol{w}-\boldsymbol{w}_m))}$$

The Evidence procedure [5|8]

With the integral for multivariate Gaussian

$$\int_{\mathbb{R}^k} \exp(-\frac{1}{2}\boldsymbol{w}^T\boldsymbol{B}\boldsymbol{w})d\boldsymbol{w} = (2\pi)^{k/2}\frac{1}{\sqrt{|\boldsymbol{B}|}}$$

The "evidence integral", or "marginal likelihood" $p(\boldsymbol{z}|\alpha,\sigma^2)$ reads,

$$\int_{\boldsymbol{W}} p(\boldsymbol{z}|\boldsymbol{w},\sigma^2)p(\boldsymbol{w}|\boldsymbol{\alpha})d\boldsymbol{w} = \frac{1}{C'\sqrt{|\boldsymbol{\Sigma}|}}\exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{z}^T\boldsymbol{z} - \boldsymbol{w}_m\boldsymbol{\Sigma}\boldsymbol{w}_m)\right]$$

$$\{\alpha_{ev},\sigma^2_{ev}\} = \underset{\alpha,\sigma^2}{\operatorname{argmax}}\, p(\boldsymbol{z}|\alpha,\sigma^2) = \underset{\alpha,\sigma^2}{\operatorname{argmax}}(\ln(p(\boldsymbol{z}|\alpha,\sigma^2)))$$

The Evidence procedure [6|8]

No explicit solution for

$$\{\alpha_{ev}, \sigma^2_{ev}\} = \operatorname*{argmax}_{\alpha,\sigma^2} p(\boldsymbol{z}|\alpha, \sigma^2)$$

found but implicit expression can be used for re-estimation of the form

$$\alpha_{new} = g(\boldsymbol{w}_m, \alpha_{old}) \quad \text{and} \quad \sigma^2_{new} = h(\boldsymbol{w}_m, \sigma^2_{old})$$

As $\boldsymbol{w}_m = \boldsymbol{\Sigma}^{-1}(\alpha, \sigma^2)\boldsymbol{\Phi}^T\boldsymbol{z}$:

Concurrently update of $\{\alpha, \sigma^2\}$ and $\{\Sigma, \boldsymbol{w}_m\}$

The Evidence procedure [7|8]

SUMMARY OF EVIDENCE PROCEDURE

- It is an approximation to the true posterior. Works good if $p(\alpha, \sigma^2|\boldsymbol{z})$ has a clear peak
- The hyperparameter are fixed to the most probable values given the data
- For Gaussian likelihood + $\boldsymbol{w}$-prior and flat hyperpriors
  - ◇ efficient iterative scheme for $\{\alpha_{ev}, \sigma^2_{ev}\}$ (convergence to a local maximum of $p(\boldsymbol{z}|\alpha, \sigma^2)$)
  - ◇ posterior is approximated with one Gaussian

[Mackay 1992]

The Evidence procedure [8|8]

Illustration, simple example

Extending evidence procedure [1|2]

AUTOMATIC RELEVANCE DETERMINATION

Introduce weight prior

$$p(\boldsymbol{w}|\boldsymbol{\alpha}) = \left[(2\pi)^{-\frac{M}{2}} \prod_{i=0}^{M} \alpha_i\right] \exp(-\frac{1}{2}\sum_{i=0}^{M} \alpha_i w_i^2)$$

Each weight gets a hyperparameter controlling its variance

What if one $\alpha_k \to \infty$ through applying evidence update rule?:

The weight w_k is centered at zero with variance $\to 0$

$\to w_k$ can be removed from the model!

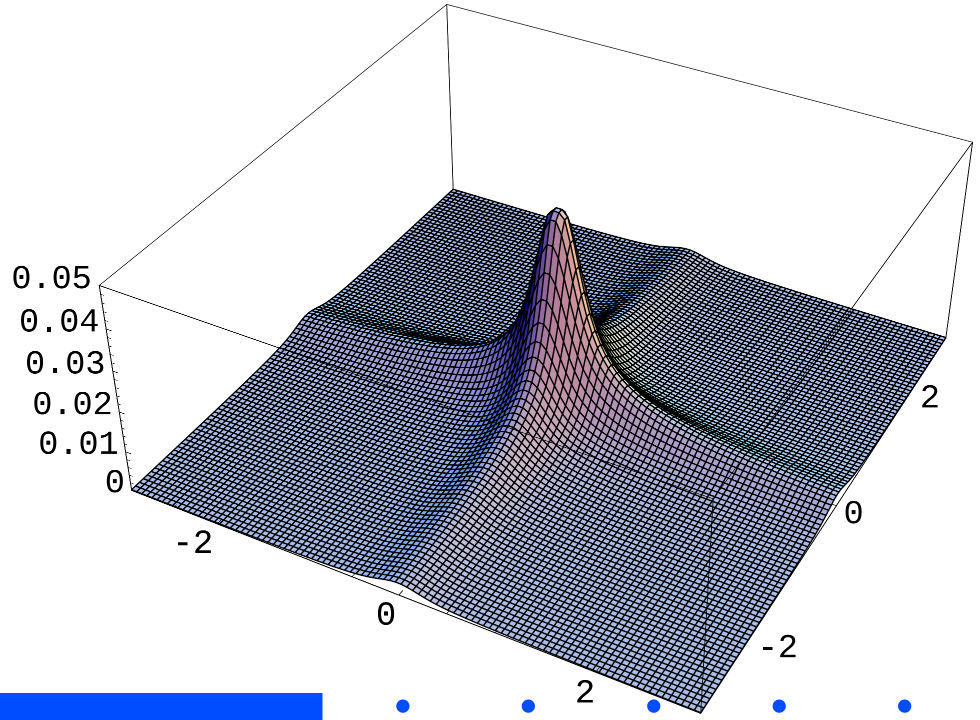
GENERATION OF HIGHLY SPARSE MODELS

[Mackay 1994, Neal 1996, Tipping 2001]

Extending evidence procedure [2|2]

What's the real w-prior (noninformative $p(\boldsymbol{\alpha})$ + Gaussian $p(\boldsymbol{w}|\boldsymbol{\alpha})$)?

$$p(\boldsymbol{w}) = \int_{D(\boldsymbol{\alpha})} p(\boldsymbol{w}|\boldsymbol{\alpha})p(\boldsymbol{\alpha})d\boldsymbol{\alpha}$$



Priors

- Noninformative Priors
- Improper Priors
- Invariance under group action
- Conjugate priors
- Maximum entropy priors

Conclusion

- Overview of MLE, Bayes, MAP for linear models
- Simple Bayesian models
- Hierarchical Bayesian models
- Approximative schemes
  - "α-marginalisation approximation"
  - Evidence procedure
- Automatic relevance determination