Non-Linear Regression

Gaussian Mixture Regression (GMR)

Gaussian Mixture Regression (GMR): Principle

Given a **multimodal** dataset

$$x \in \mathbb{R}^{N_x}, y \in \mathbb{R}^{N_y}$$

x,y are random vectors with dimension N_x and N_y respectively

1) Estimate the ***joint*** density, *p(x,y)*
2) Estimate *y* given *x* by estimating the conditional density, *p(y|x)*

We need to decide on a probabilistic model for p(x,y).
GMR assumes that p(x,y) is modeled by a mixture of Gaussians

$$p(x,y) = \sum_{k=1}^{K} \alpha_k \cdot p\left(x,y;\mu^k,\Sigma^k\right), \qquad \text{with } p\left(x,y;\mu^k,\Sigma^k\right) = N\left(\mu^k,\Sigma^k\right)$$

μ^k, Σ^k : mean and covariance matrix of Gaussian component k.

GMR: with a single Gaussian component (k=1)

$p(x,y) \sim \mathcal{N}(\mu,\Sigma)$; μ and Σ are maximum likelihood estimates

$\mu = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}$ vector with dimensionality $N_x + N_y$ $\quad \mu_x = \frac{1}{M}\sum_{m=1}^{M} x^m$, $\mu_y = \frac{1}{M}\sum_{m=1}^{M} y^m$

$\Sigma = \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_y \end{bmatrix}$ symetric matrix with dimensionality $\left(N_x + N_y\right)\times\left(N_x + N_y\right)$

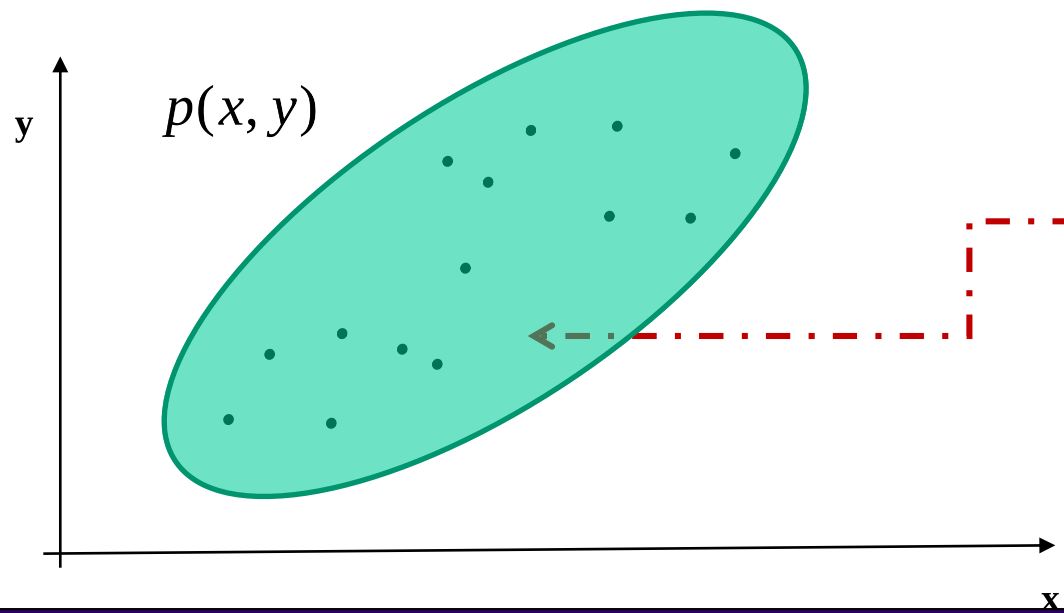
$$\Sigma_x = \frac{1}{M}\sum_{m=1}^{M}\left(x^m - \mu_x\right)\left(x^m - \mu_x\right)^T, \Sigma_{xy} = \frac{1}{M}\sum_{m=1}^{M}\left(x^m - \mu_x\right)\left(y^m - \mu_y\right)^T$$

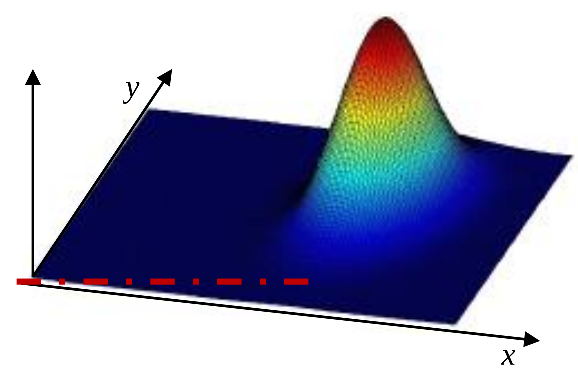
GMR: with a single Gaussian component (k=1)

1) We first estimate the ***joint* density**, *p(x,y)*, across pairs of datapoints *with a single Gauss distribution.*
2) *Then, we compute p(y|x) in order to estimate y for a query point x.*

$$p(x,y) = p(y\,|\,x)\,p(x) \Rightarrow p(y\,|\,x) = \frac{p(x,y)}{\boxed{p(x)}} \qquad p(x) = \int p(x,y)dy$$

Marginal distribution





2D projection of a Gauss function
Ellipse contour, e.g. 2 std deviations

GMR: with a single Gaussian component (k=1)

The marginals and the conditionals of a joint Gauss distribution are also Gauss distributions[1]

Joint Density

$$p(x,y) \sim \mathcal{N}(x,y \mid \mu,\Sigma)$$

$$\mu = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_y \end{bmatrix}$$

Marginals

$$p(x) \sim \mathcal{N}(x \mid \mu_x,\Sigma_x)$$

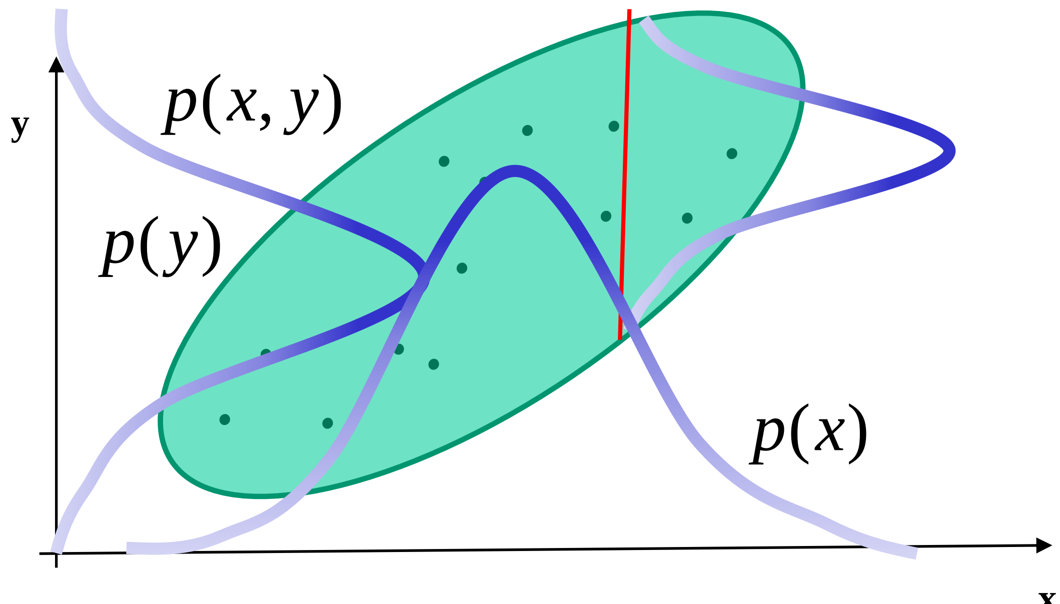
$$p(y) \sim \mathcal{N}(y \mid \mu_y,\Sigma_y)$$

Conditional

$$p(y \mid x) \sim \mathcal{N}(y;\mu_{y|x},\Sigma_{y|x})$$

$$\mu_{y|x} = \mu_y + \Sigma_{yx}\Sigma_x^{-1}(x-\mu_x)$$

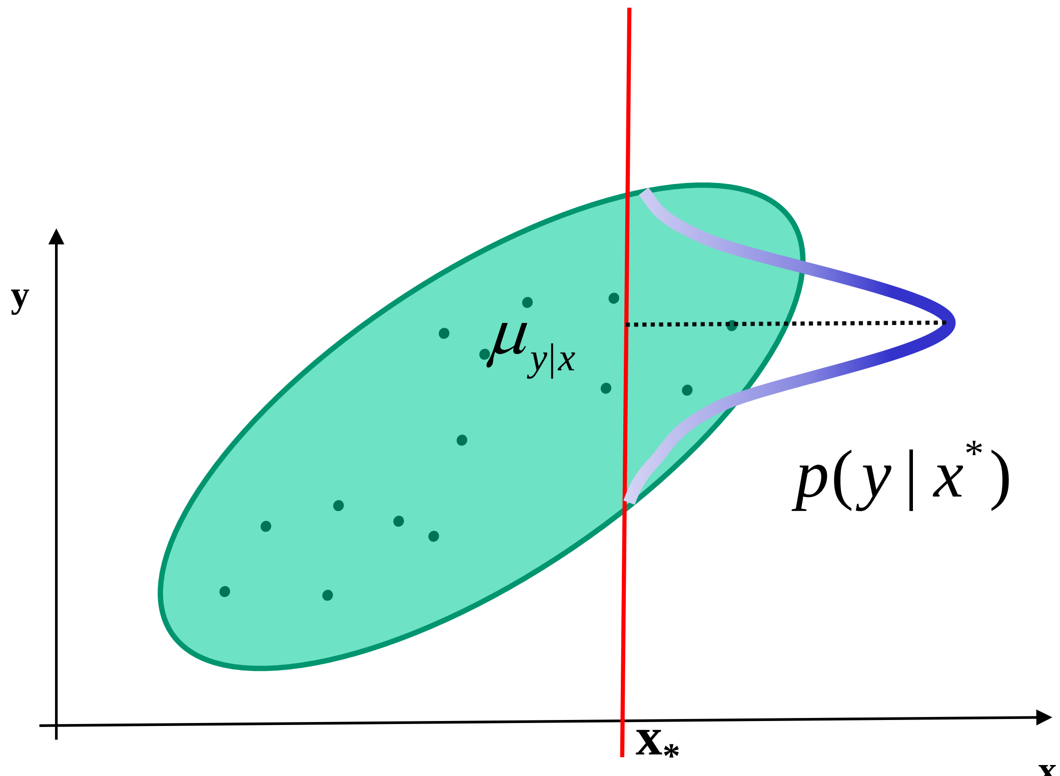
$$\Sigma_{y|x} = \Sigma_y - \Sigma_{yx}\Sigma_x^{-1}\Sigma_{xy}$$



GMR: with a single Gaussian component (k=1)

Given a query point x*, we compute the conditional.

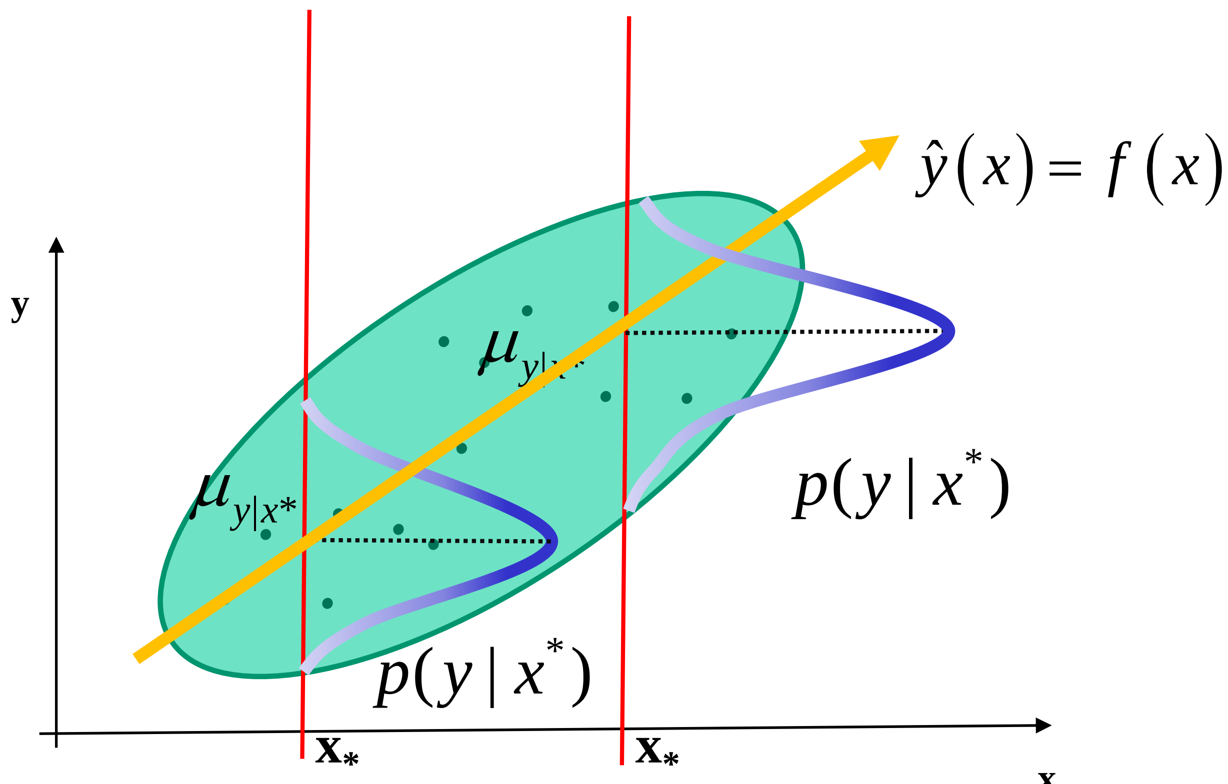
The mean of the conditional distribution is the model's prediction. $\hat{y}\left(x^{*}\right)=E\left\{p(y \mid x^{*})\right\}=\mu_{y|x^{*}}$



GMR: with a single Gaussian component (k=1)

Given a query point x*, we compute the conditional.

The mean of the conditional distribution is the model's prediction. $\hat{y}\left(x^*\right) = E\left\{p(y \mid x^*)\right\} = \mu_{y|x^*}$

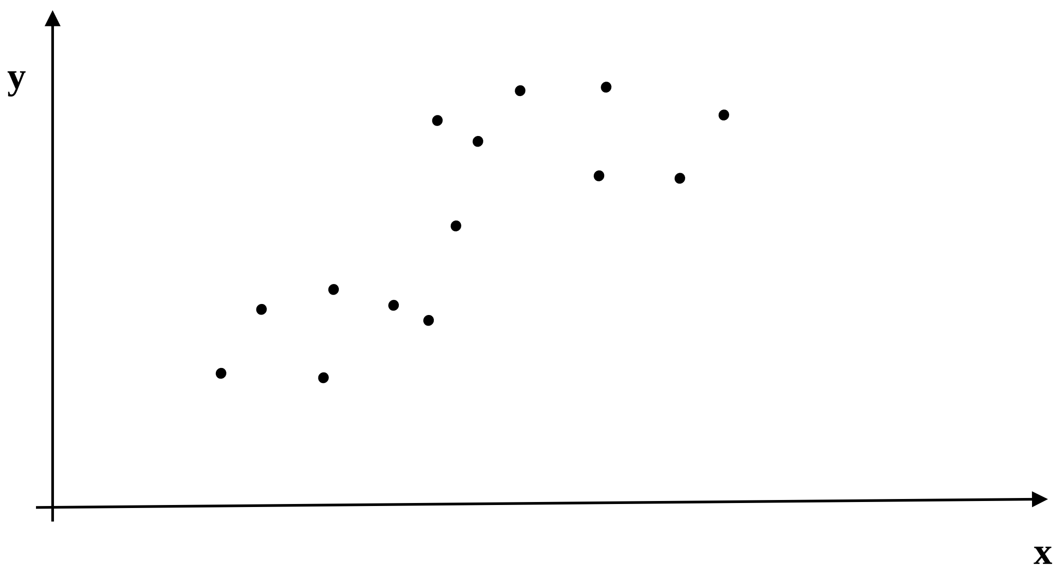


GMR: with multiple Gaussian components (k>1)

1) Estimate the ***joint* density**, *p(x,y)*, across pairs of datapoints *using GMM.*

$$p(x,y)=\sum_{k=1}^{K}\alpha_k \cdot p\left(x,y;\mu^k,\Sigma^k\right), \qquad \text{with } p\left(x,y;\mu^k,\Sigma^k\right)=N\left(\mu^k,\Sigma^k\right)$$

μ^i,Σ^i : mean and covariance matrix of Gaussian k.



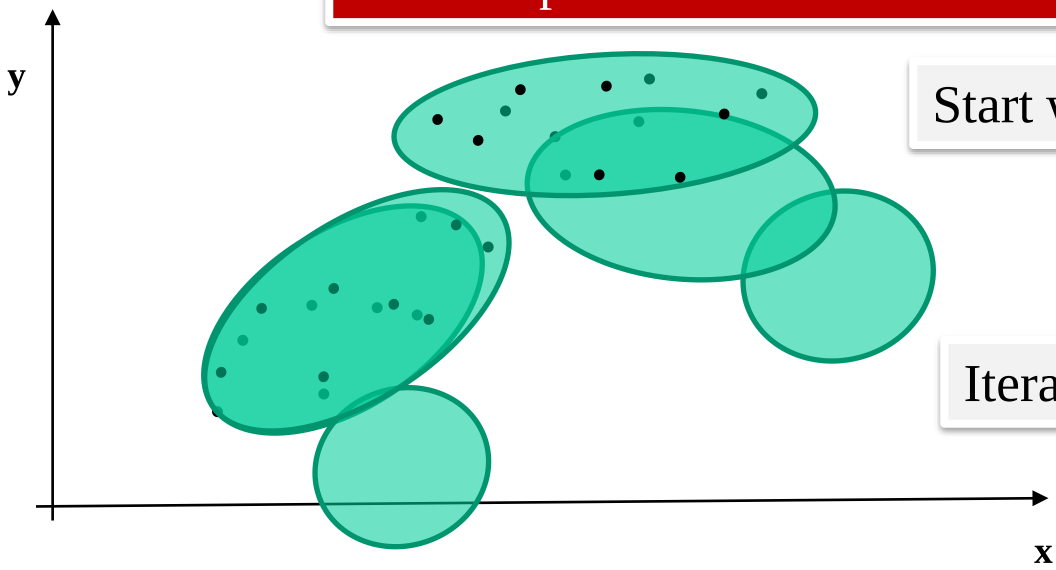
GMR: with multiple Gaussian components (k>1)

1) Estimate the *joint* density, *p(x,y)*, across pairs of datapoints *using GMM.*

$$p(x,y)=\sum_{k=1}^{K}\alpha_k \cdot p\left(x,y;\boxed{\mu^k,\Sigma^k}\right), \qquad \text{with } p\left(x,y;\mu^k,\Sigma^k\right)=N\left(\mu^k,\Sigma^k\right)$$

μ^i,Σ^i : mean and covariance matrix of Gaussian k.

Parameters are learned through Expectation-maximization.
Iterative procedure.



Start with random initialization.

$K=2$

Iterate until convergence

Estimating the joint distribution

1) Estimate the *joint density*, *p(x,y)*, across pairs of datapoints *using GMM.*

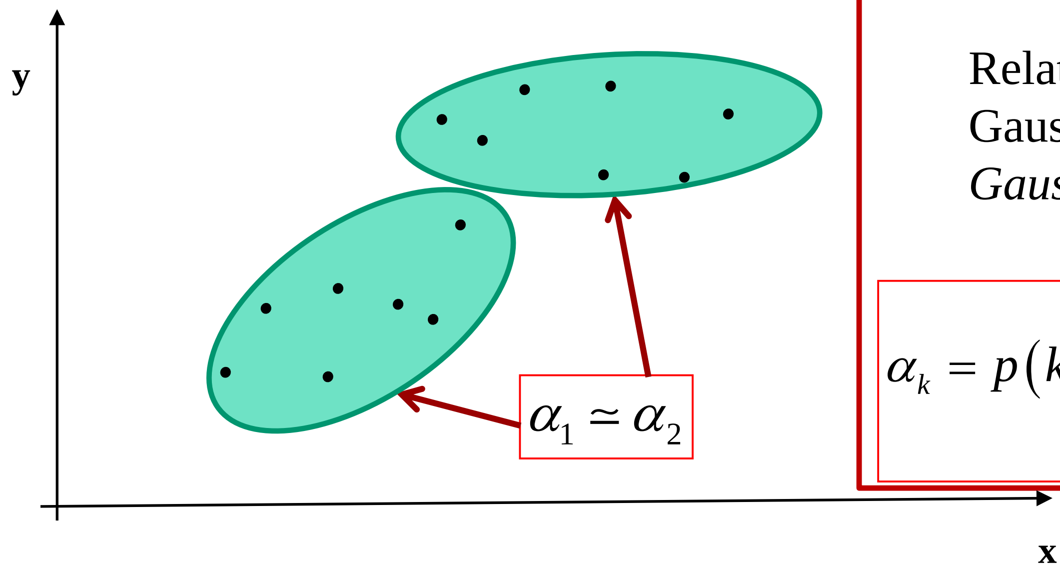
$$p(x,y)=\sum_{k=1}^{K}\alpha_k\, p\left(x,y;\mu^k,\Sigma^k\right), \qquad \text{with } p\left(x,y;\mu^k,\Sigma^k\right)=N\left(\mu^k,\Sigma^k\right)$$

μ^i,Σ^i : mean and covariance matrix of Gaussian k.

Mixing Coefficients $\sum_{k=1}^{K}\alpha_k=1$

Relative importance of each Gaussian *k (measure how well the Gaussian explains the dataset):*

$$\alpha_k=p(k)\cong\frac{1}{M}\sum_{i=1}^{M}\frac{p\left(x^i,y^i;\mu^k,\Sigma^k\right)}{\sum_l p\left(x^i,y^i;\mu^l,\Sigma^l\right)}$$



Estimating the marginal distribution

2) Estimate conditional $p(y|x) = \frac{p(x,y)}{\boxed{p(x)}}$

Marginal distribution

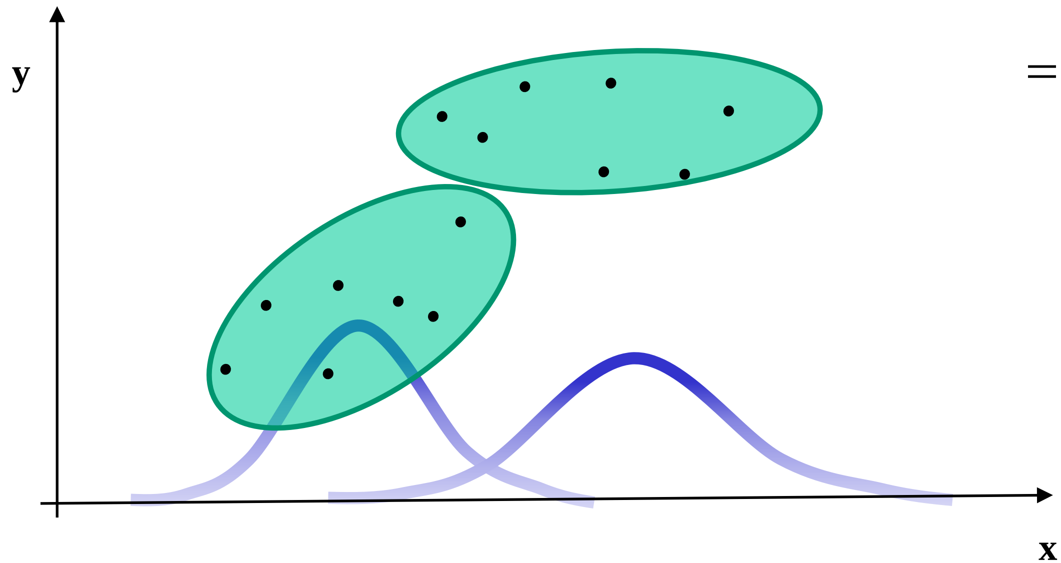
$$p(x) = \int p(x,y)dy =$$

$$= \int \sum_{k=1}^{K} \alpha_k \cdot N\left(x, y; \mu^k, \Sigma^k\right) dy$$

$$= \sum_{k=1}^{K} \alpha_k \cdot \int N\left(x, y; \mu^k, \Sigma^k\right) dy$$

$$= \boxed{\sum_{k=1}^{K} \alpha_k \cdot N(x; \mu_x^k, \Sigma_x^k)}$$

The marginal is again a GMM



Estimating the conditional distribution

2) Estimate conditional $p(y|x) = \frac{\boxed{p(x,y)}}{p(x)}$

$$p(x,y) = \sum_{k=1}^{K} \alpha_k \cdot p\left(x, y; \mu^k, \Sigma^k\right) = \sum_{k=1}^{K} \alpha_k \cdot p\left(x; \mu_x^k, \Sigma_x^k\right) p\left(y \,|\, x; \mu_{y|x}^k, \Sigma_{y|x}^k\right)$$

The covariance matrix of each Gauss function k can be decomposed into blocks of matrices

$$\Sigma^k = \begin{bmatrix} \Sigma_{xx}^k & \Sigma_{xy}^k \\ \Sigma_{yx}^k & \Sigma_{yy}^k \end{bmatrix},$$

with Σ_{xx}^k and Σ_{yy}^k the covariance matrices on x and y and Σ_{xy}^k the crosscovariance matrix.

$$p(y|x) = \frac{\sum_{k=1}^{K} \alpha_k \cdot p\left(x; \mu_x^k, \Sigma_x^k\right) p\left(y \,|\, x; \mu_{y|x}^k, \Sigma_{y|x}^k\right)}{\sum_{k=1}^{K} \alpha_k \cdot p(x; \mu_x^k, \Sigma_x^k)}$$

Set $\beta_k(x) = \frac{\alpha_k \cdot p\left(x; \mu_x^k, \Sigma_x^k\right)}{\sum_{k=1}^{K} \alpha_k \cdot p(x; \mu_x^k, \Sigma_x^k)}$

Estimating the conditional distribution

2) Estimate conditional $p(y \mid x) = \dfrac{p(x, y)}{p(x)}$

The conditional is again a GMM

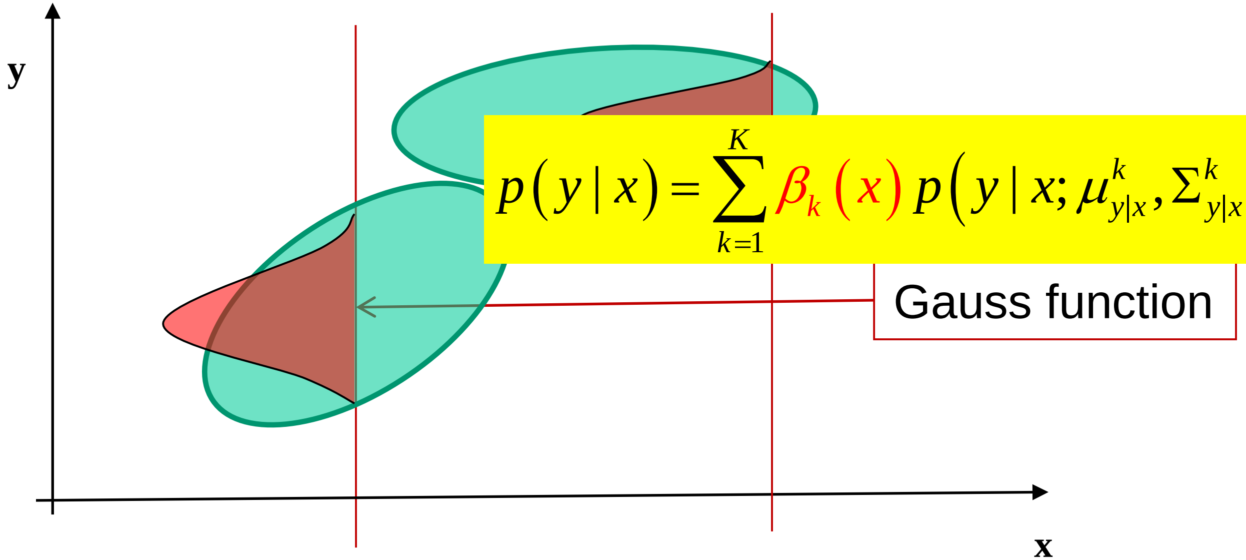
$$p\left(y \mid x\right) = \sum_{k=1}^{K} \beta_k\left(x\right) p\left(y \mid x; \mu_{y|x}^{k}, \Sigma_{y|x}^{k}\right)$$

Weight of the marginals

Set $\beta_k(x) = \dfrac{\alpha_k \cdot p\left(x; \mu_x^k, \Sigma_x^k\right)}{\sum_{k=1}^{K} \alpha_k \cdot p(x; \mu_x^k, \Sigma_x^k)}$

Estimating the conditional distribution

$$p\left(y \mid x; \mu_{y|x}^{k}, \Sigma_{y|x}^{k}\right)$$



$$p\left(y \mid x\right) = \sum_{k=1}^{K} \beta_k\left(x\right) p\left(y \mid x; \mu_{y|x}^{k}, \Sigma_{y|x}^{k}\right)$$

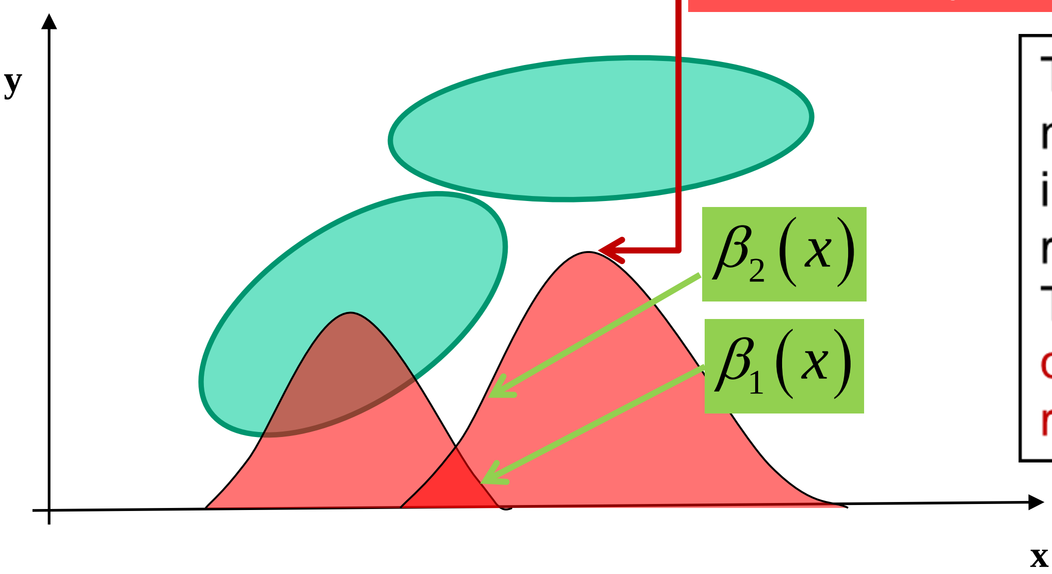
The expression changes depending on the query point

Estimating the conditional distribution

$$p(y|x) = \sum_{k=1}^{K} \beta_k(x)\, p\left(y | x; \mu_{y|x}^k, \Sigma_{y|x}^k\right)$$

with $$\beta_k(x) = \frac{\alpha_k \cdot p\left(x; \mu_x^k, \Sigma_x^k\right)}{\sum_{k=1}^{K} \alpha_k \cdot p(x; \mu_x^k, \Sigma_x^k)}$$

Influence of each marginal is modulated by α



The factors β give a measure of the relative importance of each K regressive model.
They are computed at each query point → weighted regression

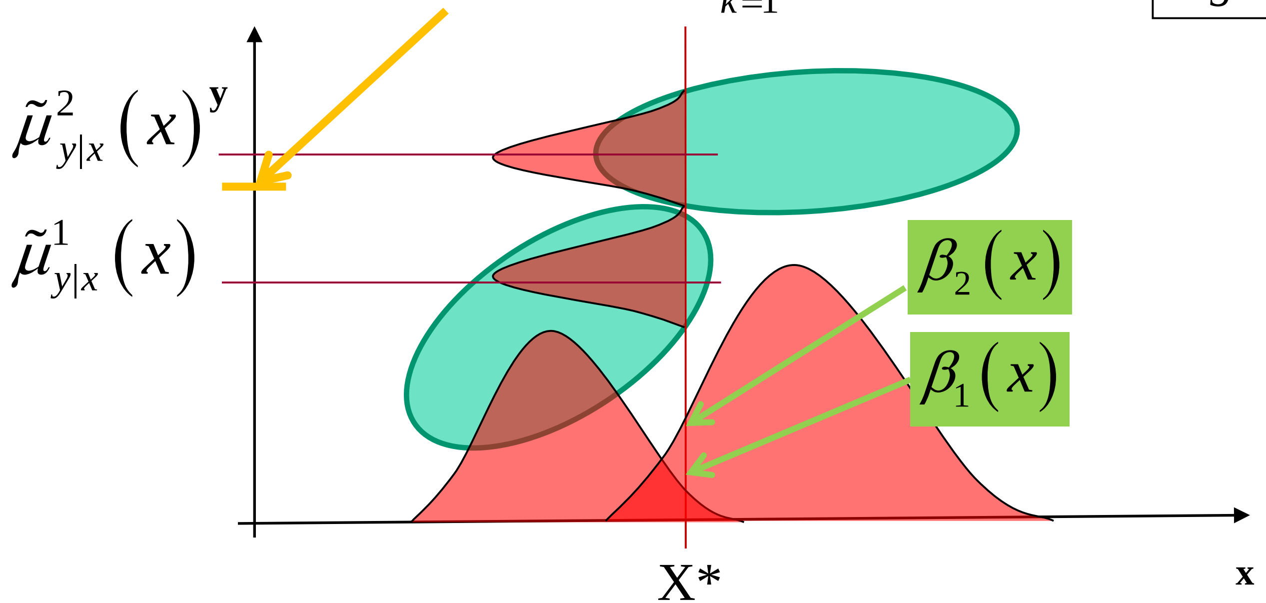
Computing the regression

The GMR prediction at a query point x* is obtained by computing *E{p(y|x)}*

$$y = E\{p(y|x)\} = \sum_{k=1}^{K} \beta_k(x) \underbrace{\left(\mu_y^k + \Sigma_{yx}^k \left(\Sigma_{xx}^k\right)^{-1} \left(x - \mu_x^k\right) \right)}_{\tilde{\mu}_{y|x}^k(x)}$$

$$\Rightarrow \boxed{y} = \sum_{k=1}^{K} \beta_k(x) \cdot \tilde{\mu}_{y|x}^k(x)$$

Linear combination of *K local regressive* models



Computing the variance

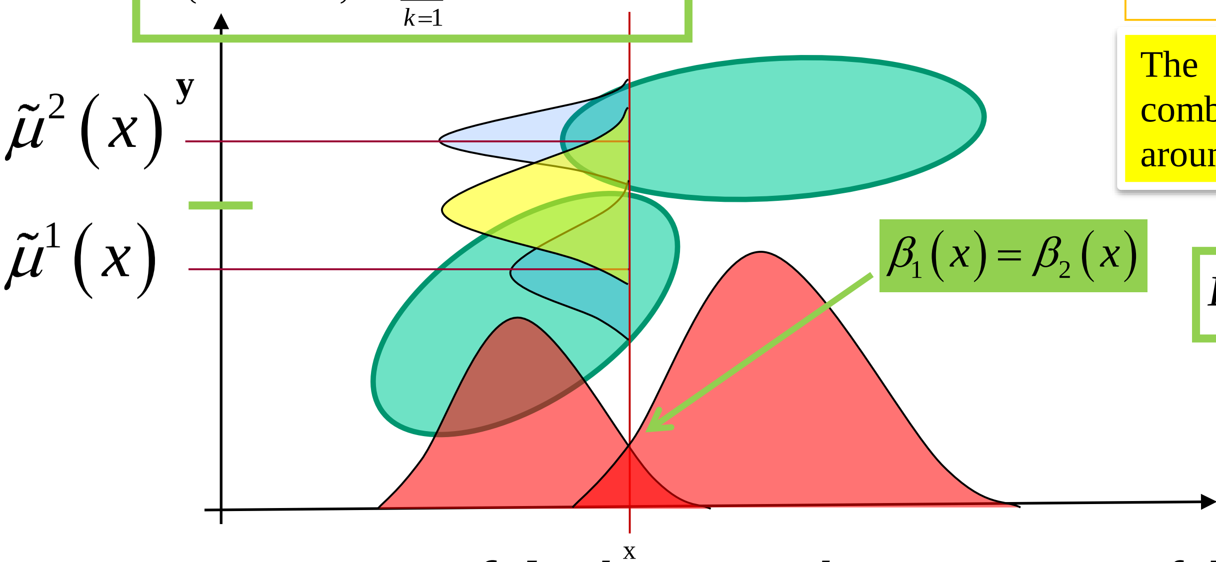
Computing the variance *var{p(y|x)}* provides information on the variability of the prediction.

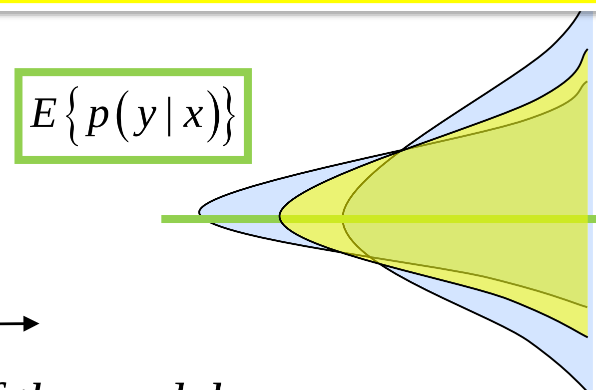
$$\mathrm{var}\left\{p(y\mid x)\right\} = \sum_{k=1}^{K} \beta_k(x)\cdot\left(\left(\tilde{\mu}^k_{y|x}(x)\right)^2 + \tilde{\Sigma}^k_{y|x}\right) - \left(\sum_{k=1}^{K}\left(\beta_k(x)\cdot\tilde{\mu}^k_{y|x}(x)\right)\right)^2$$

$$E\left\{p(y\mid x)\right\} = \sum_{k=1}^{K} \beta_k(x)\cdot\tilde{\mu}^k(x)$$

with $\tilde{\Sigma}^k_{y|x} = \Sigma^k_{yy} - \Sigma^k_{yx}\left(\Sigma^k_{xx}\right)^{-1}\Sigma^k_{xy}$

The variance of the model is a weighted combination of the variances of the models around the weighted mean.





Careful: *This is not the uncertainty of the model.*
Use the likelihood to compute the uncertainty of the model!

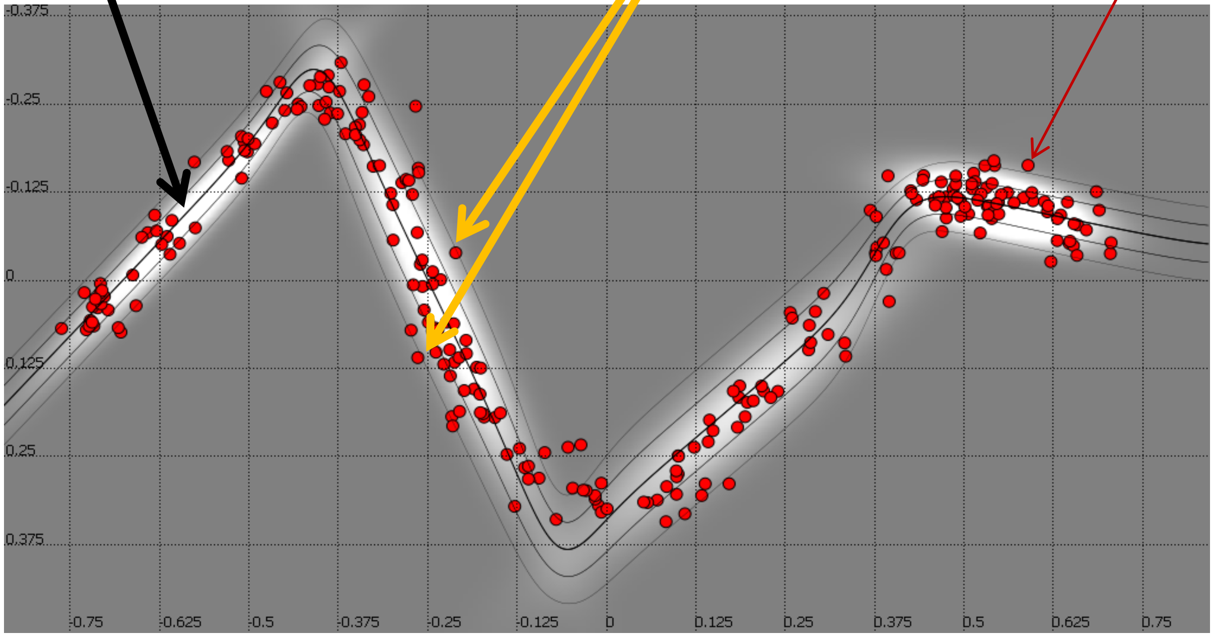
Interpreting the variance

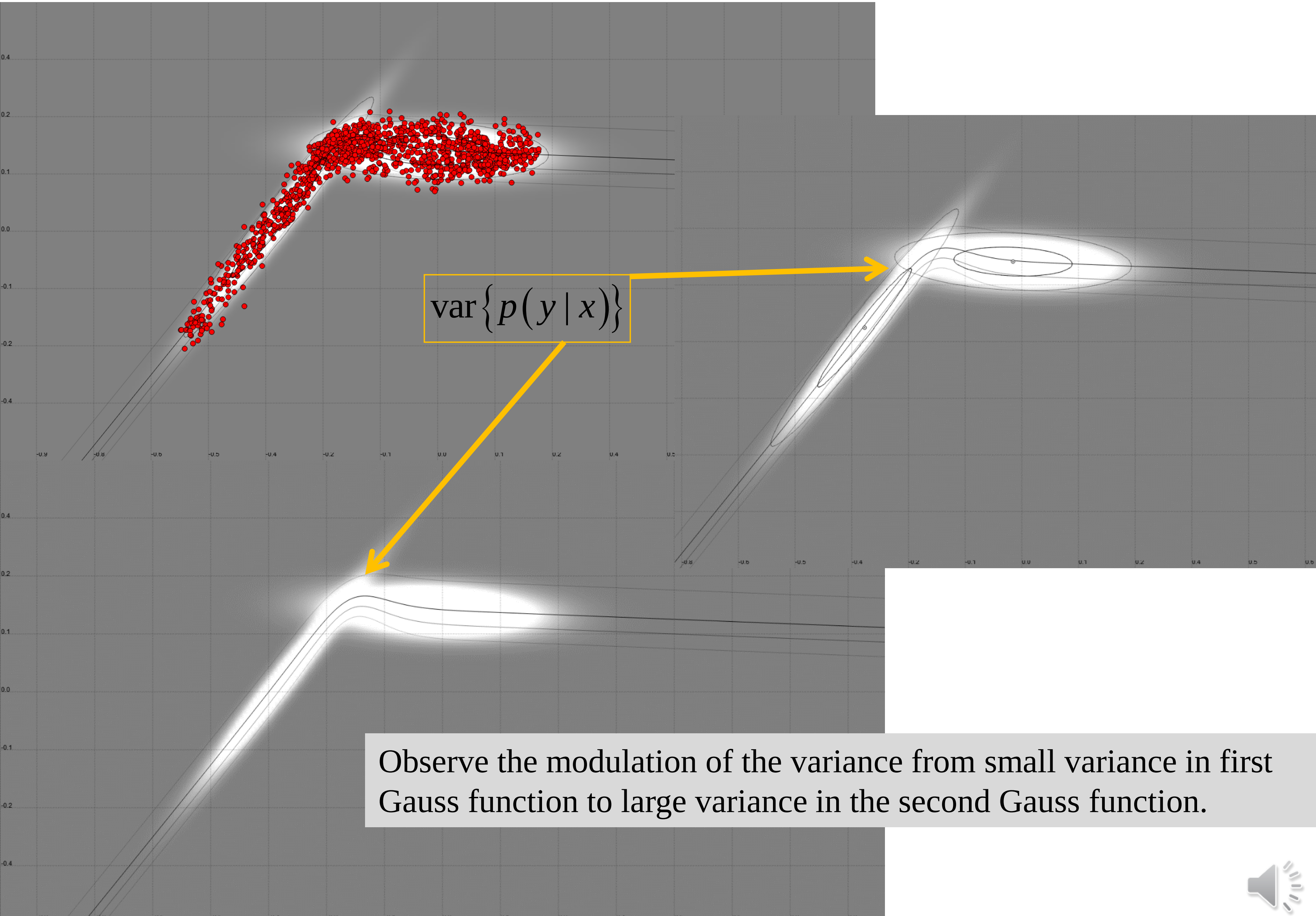
The variance *var{p(x,y)}* can be visualized all around the regressive line. It provides information on the evolution of the noise in state space.

$E\{p(y|x)\}$

$\operatorname{var}\{p(y|x)\}$

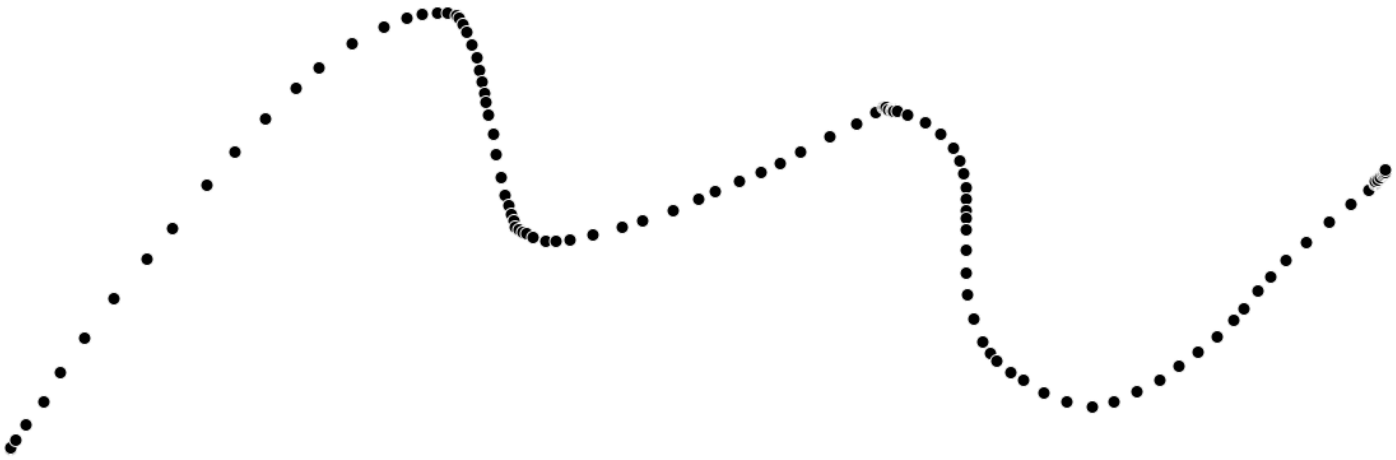
Color shading gives the likelihood of the model (uncertainty).



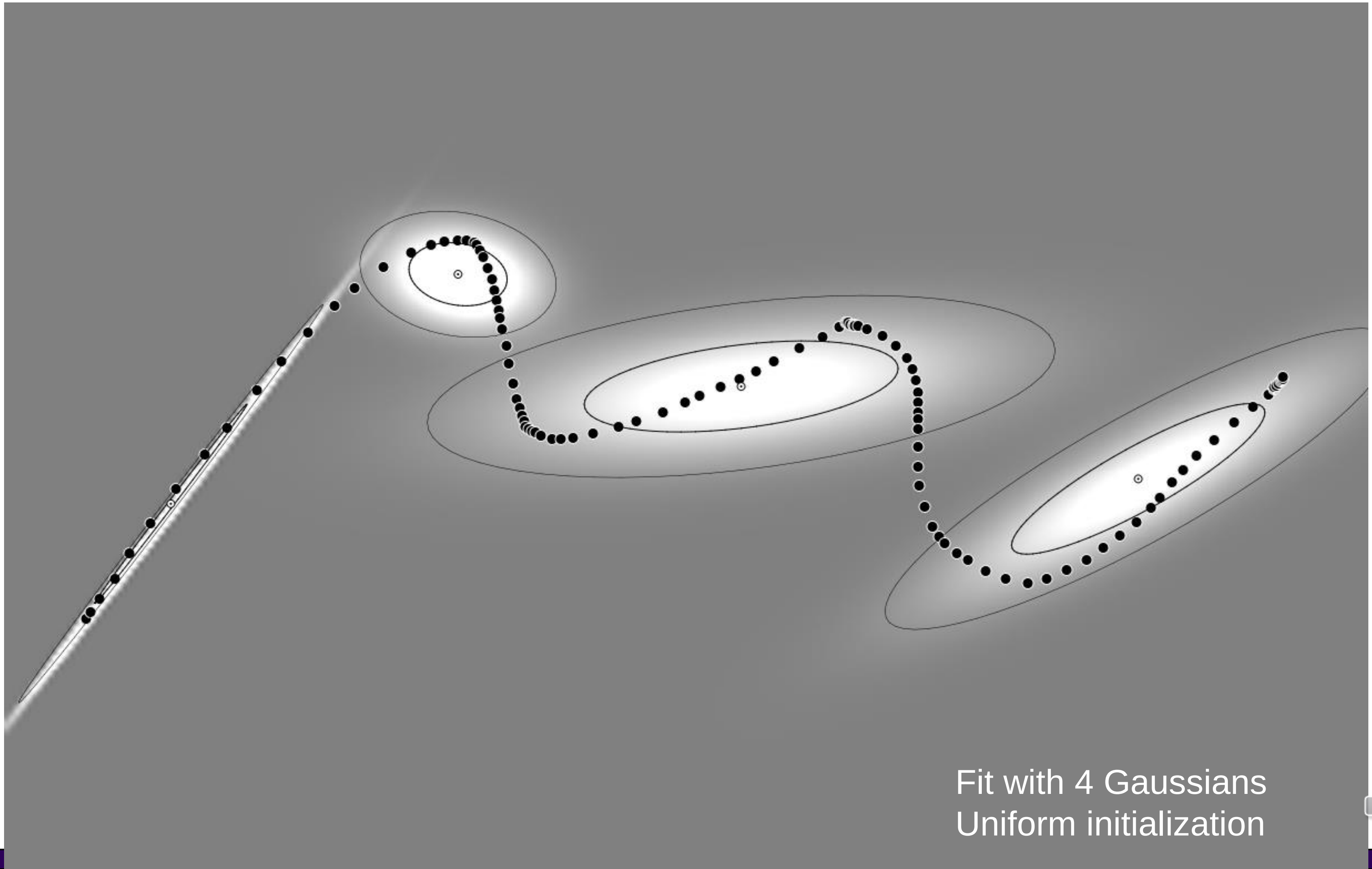


Observe the modulation of the variance from small variance in first Gauss function to large variance in the second Gauss function.

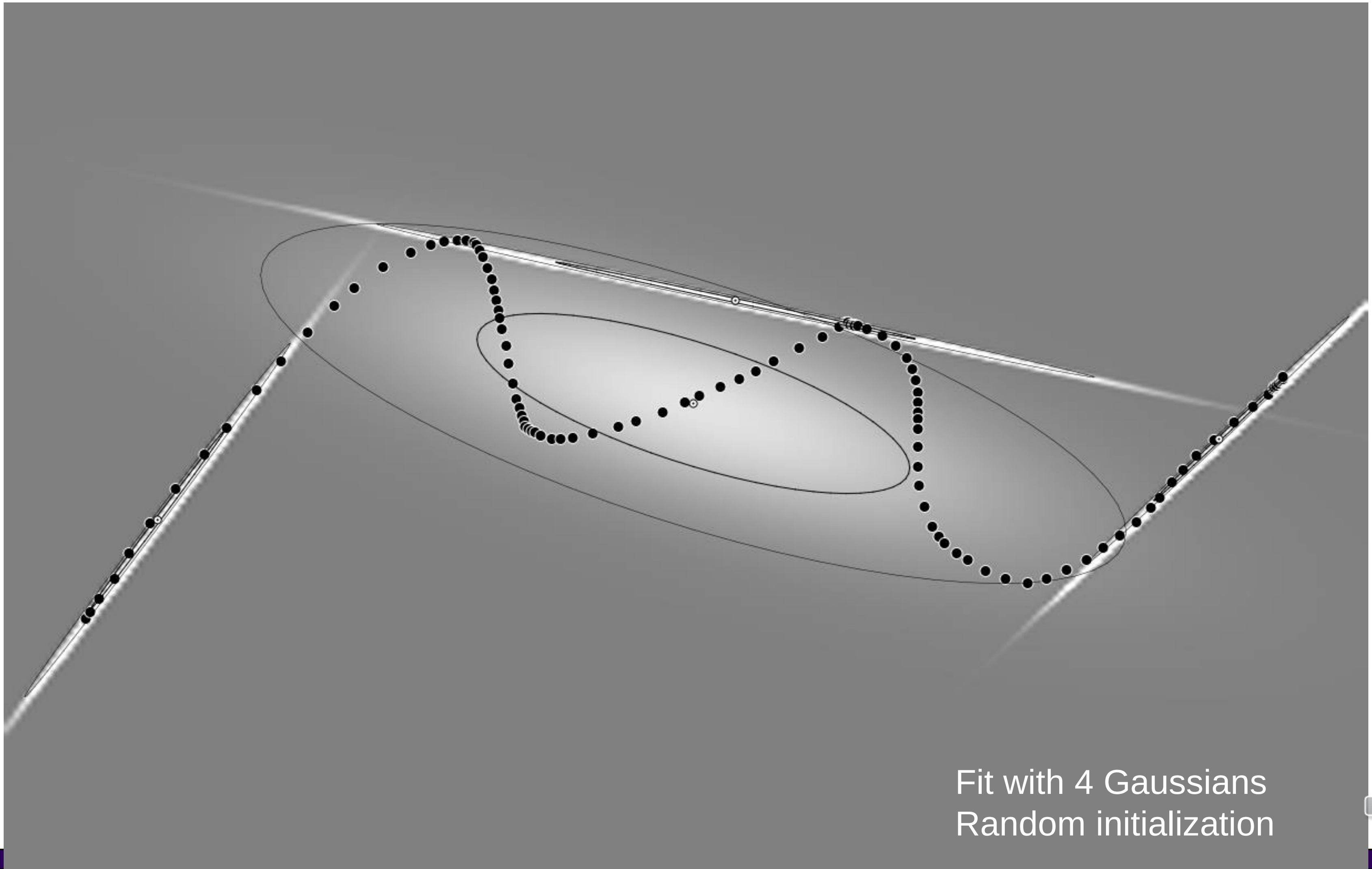
GMR: Sensitivity to Choice of K and Initialization



GMR: Sensitivity to Choice of *K* and Initialization



GMR: Sensitivity to Choice of *K* and Initialization



Fit with 4 Gaussians
Random initialization

GMR: Take home message

GMR proceeds in two steps:

- ❑ Parametrize the density p(x,y) and then estimate solely the parameters. The density is constructed from a mixture of K Gaussians
- ❑ Compute the regression from the expectation over the conditional.

Such *generative model* provides more information than models that directly computing *p(y|x).*

→ It allows to learn to predict a multi-dimensional output *y.*
→ It allows to embed correlations across the output dimensions *y.*
→ It allows to query *x* given y, i.e. to compute *p(x|y).*

It comes at the cost of computing the full distribution, while often we care only for the conditional.

The estimate depends on E-M which is very sensitive to initialization.