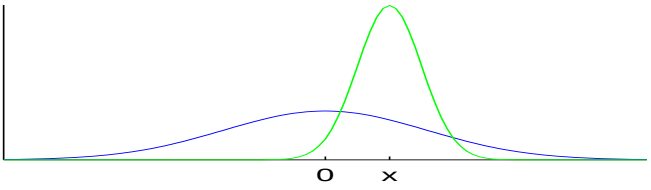
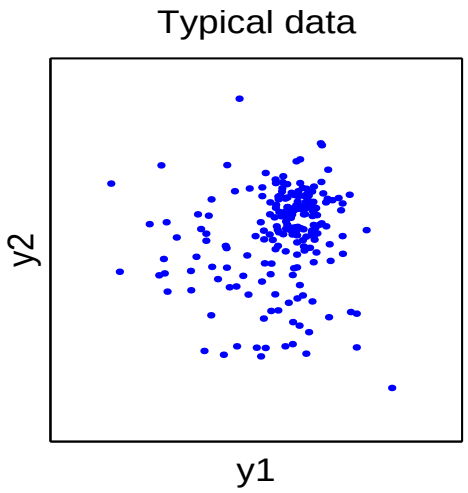
Example

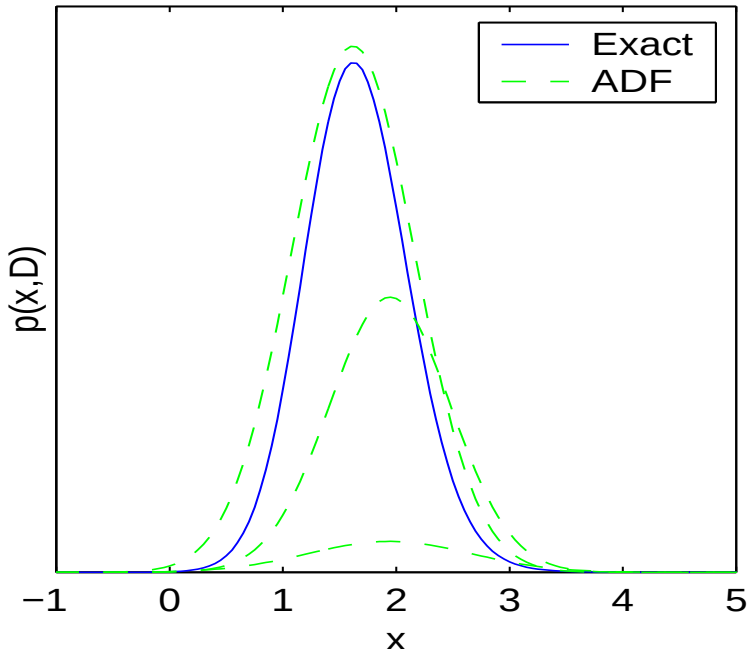
Data model

$$p(y|x) = \frac{1}{2}\mathcal{N}(y; x, 1) + \frac{1}{2}\mathcal{N}(y; 0, 10)$$





ADF posterior for three orderings of same data:



True x = 2

20 data points

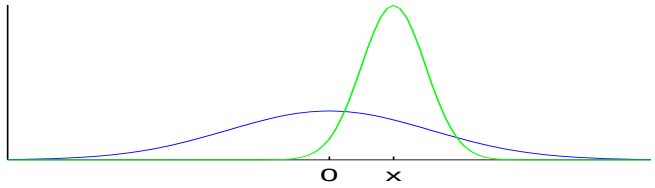
ADF is sensitive to ordering

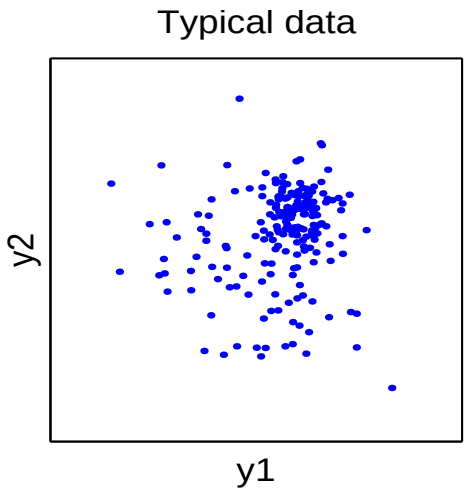
Can we make ADF independent of ordering?

Example continued

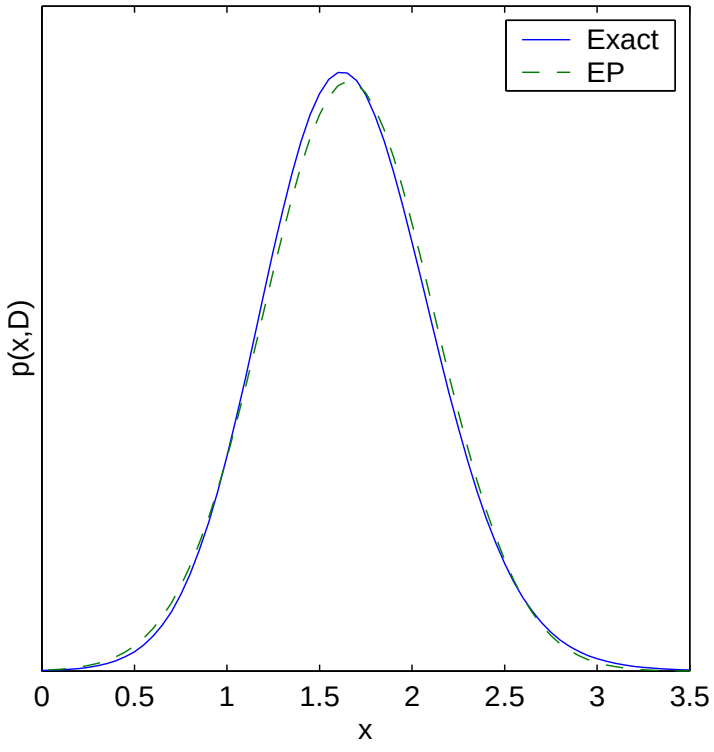
Data model

$$p(y|x) = \frac{1}{2}\mathcal{N}(y; x, 1) + \frac{1}{2}\mathcal{N}(y; 0, 10)$$

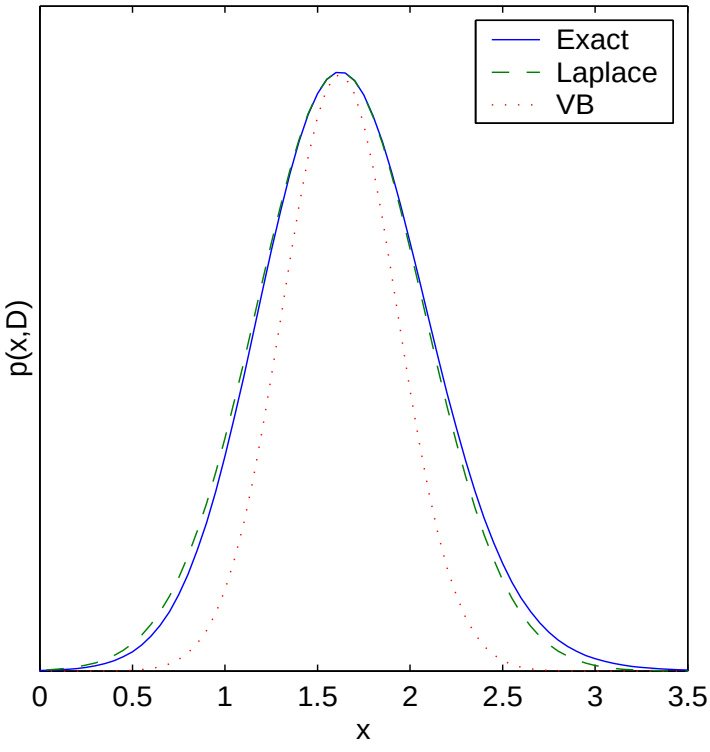




EP posterior at convergence

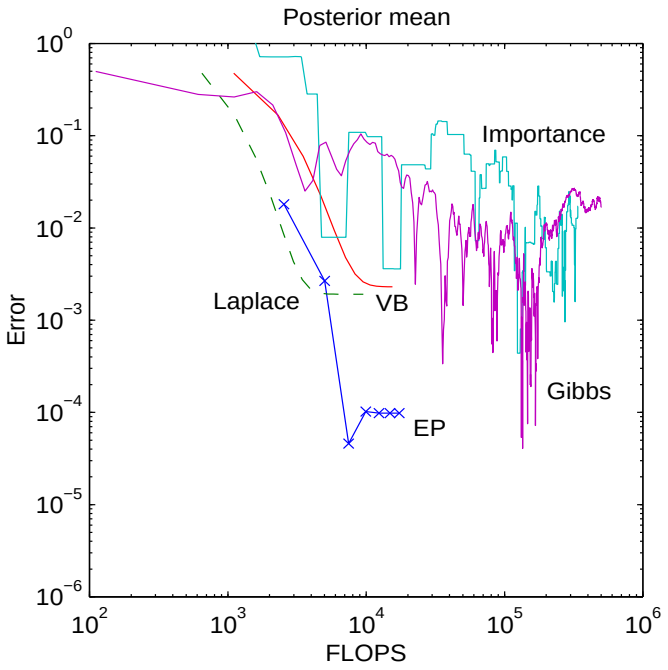


Other methods

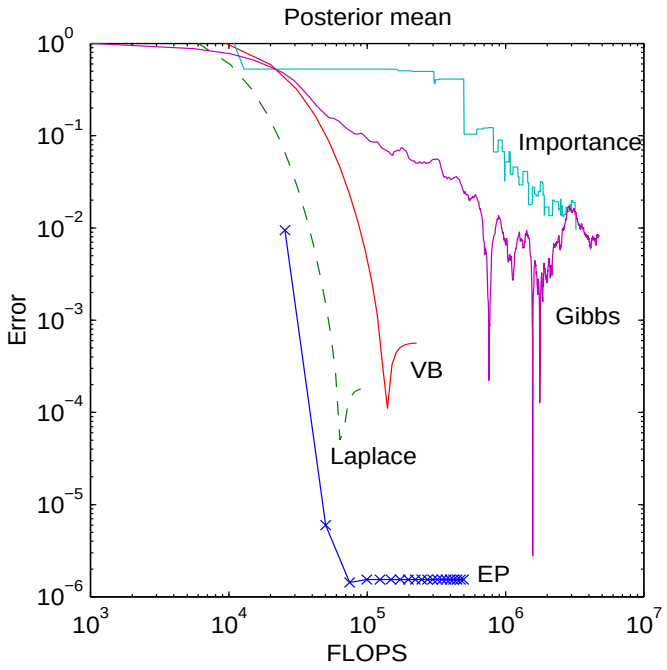


All independent of data ordering

Performance



Data size n=20



n=200

ADF = first ‘x’ of EP

VB = variational bound

Deterministic methods improve with more data

(posterior is more Gaussian)

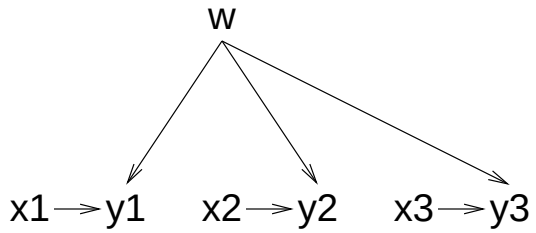
Sampling methods do not care

Bayes point machine

Bayesian approach to linear classification

Use w to classify x:

$$\begin{aligned} \mathbf{w}^{\mathrm{T}}\mathbf{x}_i &> 0 \qquad (\text{class } 1) \\ \mathbf{w}^{\mathrm{T}}\mathbf{x}_i &< 0 \qquad (\text{class } 2) \end{aligned}$$



$$p(\mathbf{w}, D) = p(\mathbf{w}) \prod_i p(y_i|\mathbf{x}_i, \mathbf{w})$$

p(w) is uniform

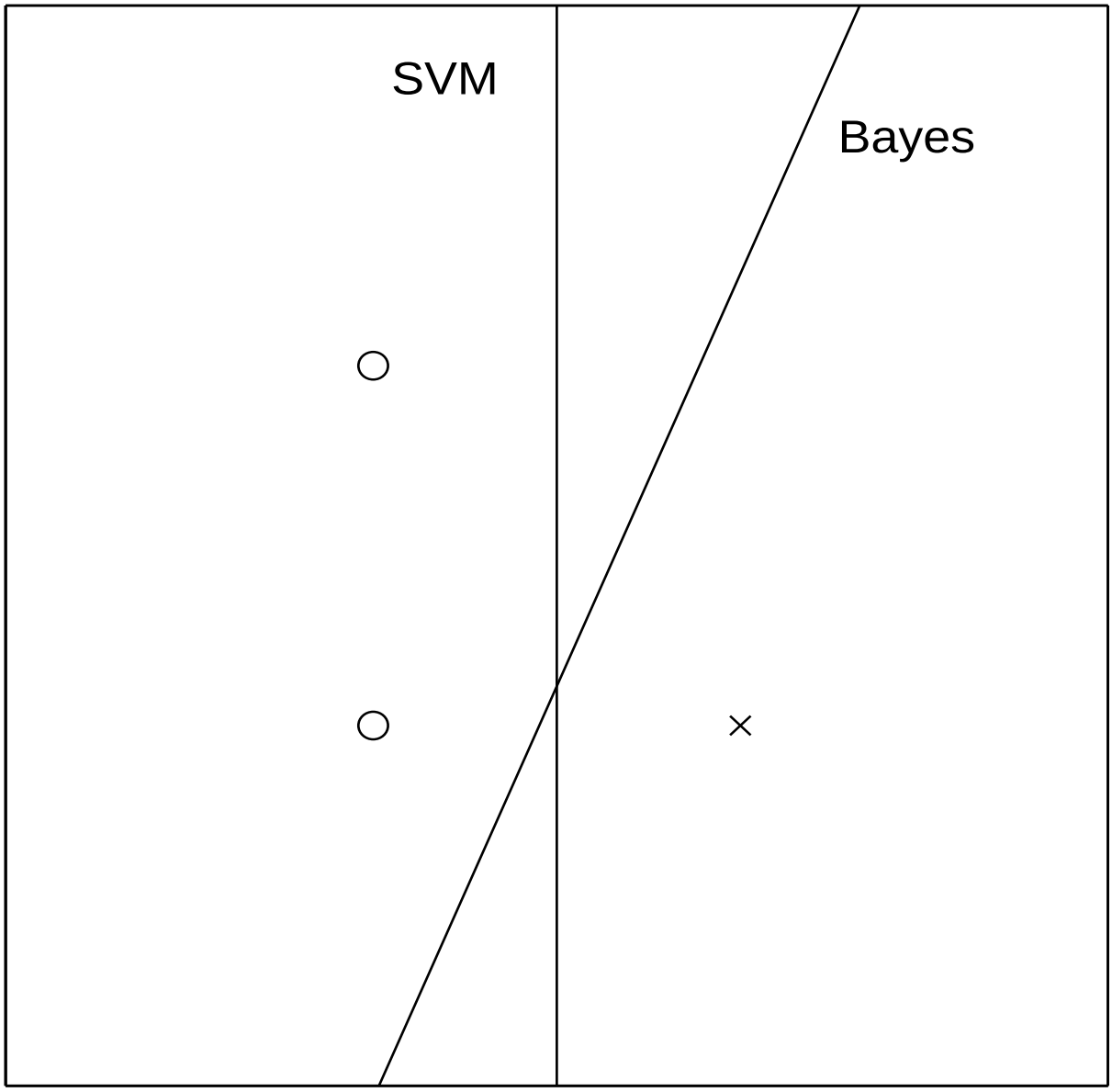
$$\begin{aligned} p(y|\mathbf{x}, \mathbf{w}) &= \Theta(y\mathbf{w}^{\mathrm{T}}\mathbf{x}) \\ &= \begin{cases} 1 & \text{if } \mathbf{w} \text{ is a perfect separator} \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Classify a new data point by voting:

$$\begin{aligned} p(y|\mathbf{x}, D) &= \int_{\mathbf{w}} p(y|\mathbf{x}, \mathbf{w}) p(\mathbf{w}|D) d\mathbf{w} \\ y &= E[\mathrm{sign}(\mathbf{w}^{\mathrm{T}}\mathbf{x})|D] \\ &\approx \mathrm{sign}(E[\mathbf{w}|D]^{\mathrm{T}}\mathbf{x}) \end{aligned}$$

E[w|D] is the Bayes Point

Bayes point machine example



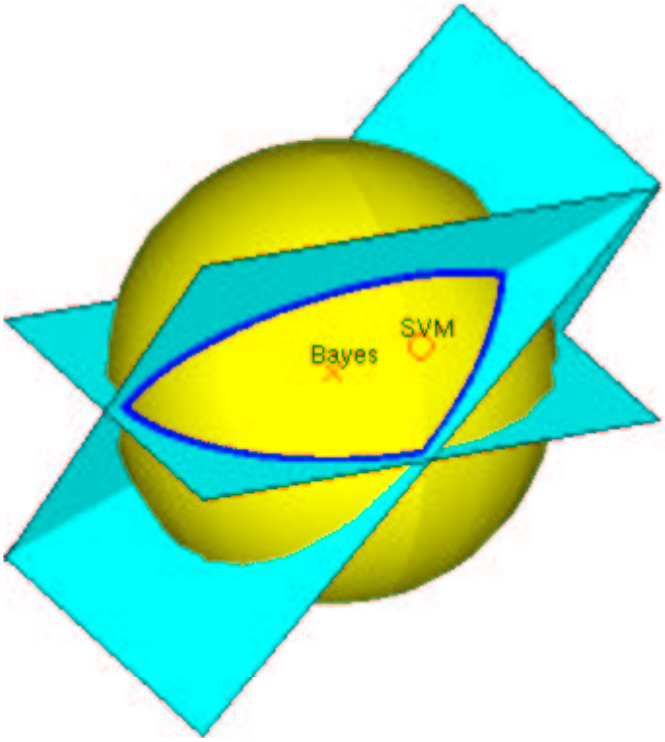
SVM → Maximize margin

(distance to closest data point)

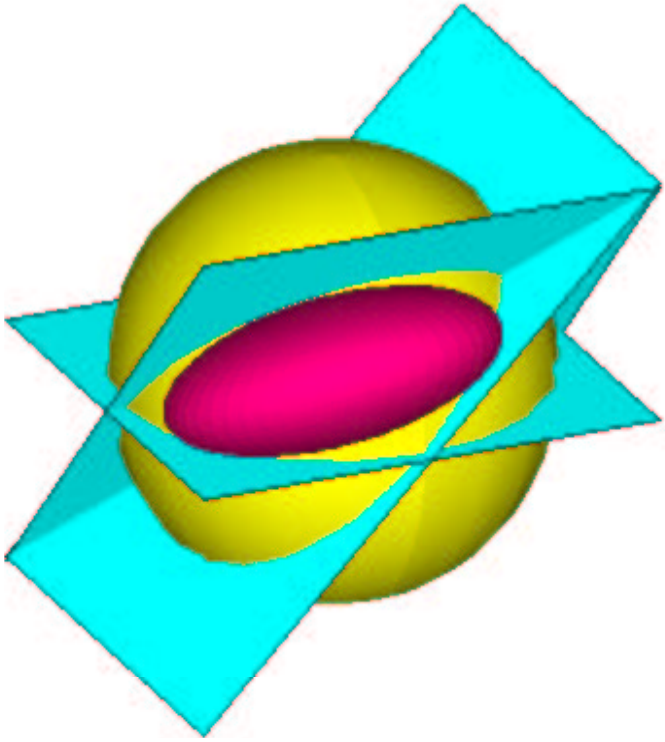
Bayes → Vote all perfect separators

Performance of EP

Version space



EP Gaussian posterior

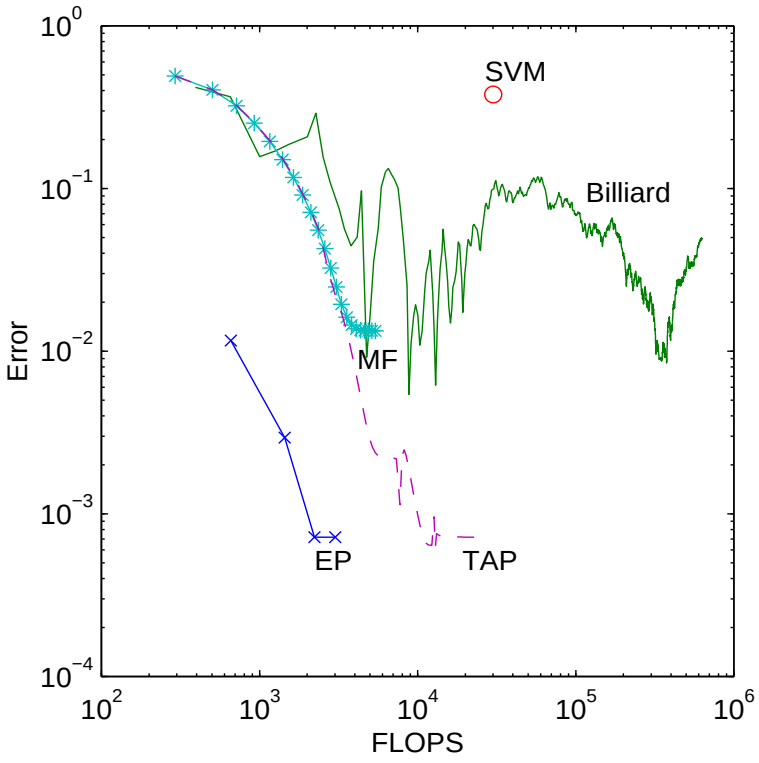


Billiard = Monte Carlo

Opper&Winther's algs:

MF = mean−field theory

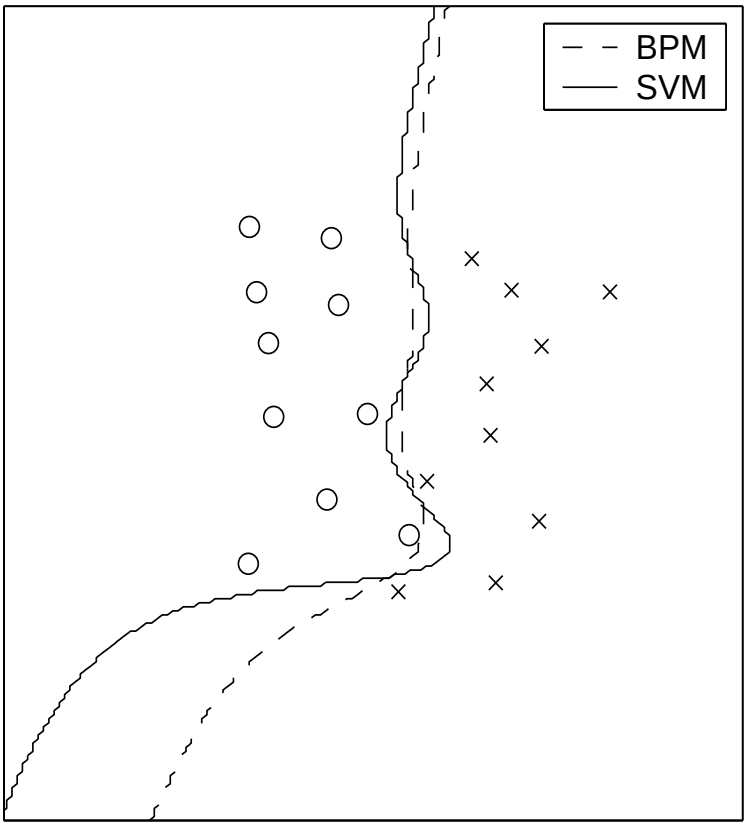
TAP = cavity method
(equiv to Gaussian EP)



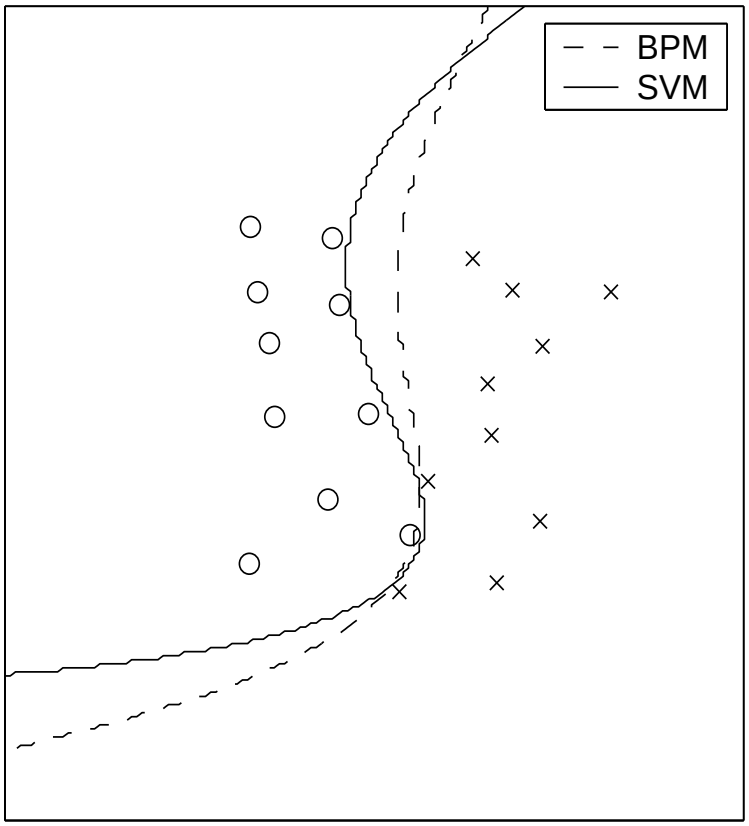
Gaussian kernels

Map data into high dimensional space so that

$$\phi(\mathbf{x}_i)^{\mathrm{T}}\phi(\mathbf{x}_j) = \exp(-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{2\sigma^2})$$



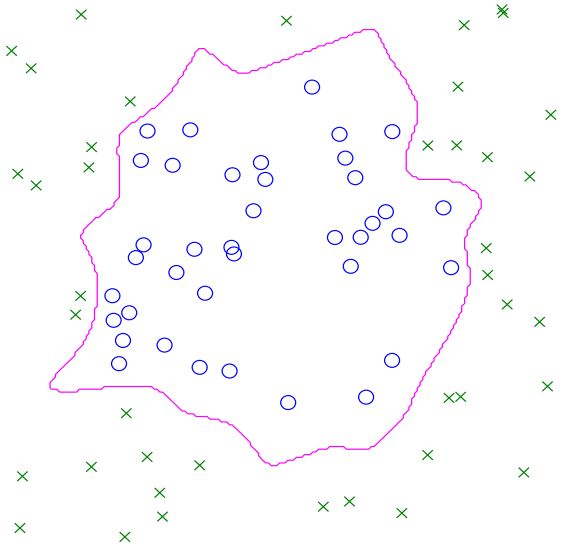
narrow width 0.2



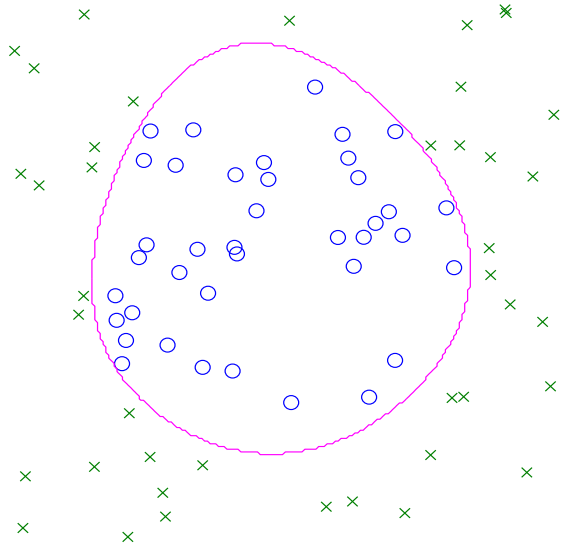
wide width 0.5

SVM boundaries are more contrived, sensitive to kernel

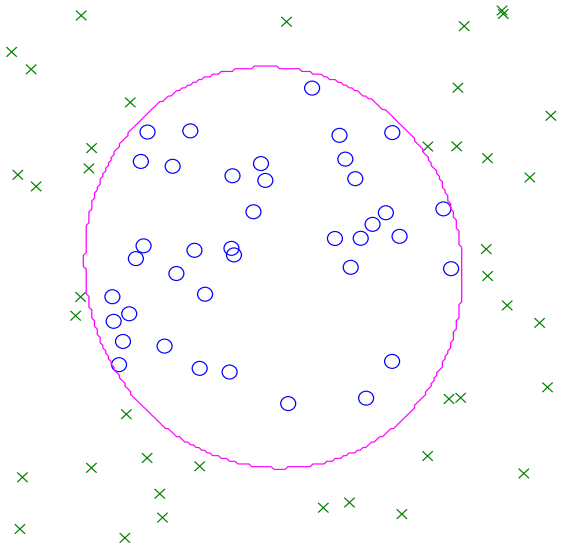
Kernel selection



Gaussian kernel, width 0.08
(SVM choice)



Gaussian kernel, width 0.6
(Bayes choice among Gaussians)



Quadratic kernel
(Bayes choice)

Kernel	R^2/ρ^2	$\log(p(D))$
$\sigma = 0.08$	18	-39
$\sigma = 0.6$	108	-19
quadratic	656	-16

SVM and EP have similar boundaries, but prefer different kernels