CS540 Machine learning
Bayesian statistics

Last time

- Number game
- Beta-Bernoulli
- Dirichlet-multinomial

Outline

- Bayesian estimation of
  - Gaussians
  - Generative classifiers
  - MVN
  - Linear regression
  - Logistic regression

Gaussians

- $P(\mu|\sigma^2,D)$
- $P(\sigma^2|\mu,D)$
- $P(\mu,\sigma^2|D)$

Unknown mean

- Conjugate prior is Gaussian

$$
\begin{aligned}
p(\mu|D) &\propto p(D|\mu,\sigma)p(\mu|m_0,\tau_0^2) \\
&\propto \exp\left[-\frac{1}{2\sigma^2}\sum_i (x_i-\mu)^2\right] \times \exp\left[-\frac{1}{2\tau_0^2}(\mu-m_0)^2\right] \\
&= \exp\left[\frac{-1}{2\sigma^2}\sum_i (x_i^2+\mu^2-2x_i\mu) + \frac{-1}{2\tau_0^2}(\mu^2+m_0^2-2m_0\mu)\right] \\
&= \exp\left[-\frac{\mu^2}{2}\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right) + \mu\left(\frac{m_0}{\tau_0^2}+\frac{\sum_i x_i}{\sigma^2}\right) - \left(\frac{m_0^2}{2\tau_0^2}+\frac{\sum_i x_i^2}{2\sigma^2}\right)\right] \\
&\stackrel{\text{def}}{=} \exp\left[-\frac{1}{2\tau_n^2}(\mu^2-2\mu m_n+m_n^2)\right] = \exp\left[-\frac{1}{2\tau_n^2}(\mu-m_n)^2\right] \\
&= \mathcal{N}(\mu|m_n,\tau_n^2)
\end{aligned}
$$

Completing the square

- Match powers in μ^2

$$
\begin{aligned}
p(\mu|\mathcal{D}) &= \exp\left[-\frac{\mu^2}{2}\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)+\mu\left(\frac{m_0}{\tau_0^2}+\frac{\sum_i x_i}{\sigma^2}\right)-\left(\frac{m_0^2}{2\tau_0^2}+\frac{\sum_i x_i^2}{2\sigma^2}\right)\right] \\
&\stackrel{\text{def}}{=} \exp\left[-\frac{1}{2\tau_n^2}(\mu^2-2\mu m_n+m_n^2)\right] \\
\frac{1}{\tau_n^2} &= \frac{1}{\tau_0^2}+\frac{n}{\sigma^2} \\
\tau_n^2 &= \frac{\sigma^2\tau_0^2}{n\tau_0^2+\sigma^2}=\frac{1}{\frac{n}{\sigma^2}+\frac{1}{\tau_0^2}}
\end{aligned}
$$

- Define

$$\lambda = 1/\sigma^2, \;\; \ell_0 = 1/\tau_0^2, \;\; \ell_n = 1/\tau_n^2$$

$$\ell_n = \ell_0 + n\lambda$$

Posterior precision = prior precision + n * measurement precision

Completing the square

- Match powers in μ

$$
\begin{aligned}
p(\mu|\mathcal{D}) &= \exp\left[-\frac{\mu^2}{2}\left(\frac{1}{\tau_0^2}+\frac{n}{\sigma^2}\right)+\mu\left(\frac{m_0}{\tau_0^2}+\frac{\sum_i x_i}{\sigma^2}\right)-\left(\frac{m_0^2}{2\tau_0^2}+\frac{\sum_i x_i^2}{2\sigma^2}\right)\right] \\
&\stackrel{\text{def}}{=} \exp\left[-\frac{1}{2\tau_n^2}(\mu^2-2\mu m_n+m_n^2)\right] \\
\frac{m_n}{\tau_n^2} &= \frac{\sum_{i=1}^n x_i}{\sigma^2}+\frac{m_0}{\tau_0^2}=\frac{\tau_0^2 n\overline{x}+\sigma^2 m_0}{\sigma^2\tau_0^2} \\
m_n &= \frac{\sigma^2}{n\tau_0^2+\sigma^2}m_0+\frac{n\tau_0^2}{n\tau_0^2+\sigma^2}\overline{x}=\tau_n^2\left(\frac{m_0}{\tau_0^2}+\frac{n\overline{x}}{\sigma^2}\right) \\
m_n &= \frac{\overline{x}n\lambda+m_0\ell_0}{\ell_n}=\alpha\overline{x}+(1-\alpha)m_0 \\
\alpha &= \frac{n\lambda}{\ell_n}
\end{aligned}
$$

Posterior mean = convex comb of prior mean and MLE

Posterior mean

- Consider N=1.

$$\mu_1 = \frac{\sigma^2}{\sigma^2+\sigma_0^2}\mu_0 + \frac{\sigma_0^2}{\sigma^2+\sigma_0^2}x$$

Convex comb of prior and MLE

$$= \mu_0 + (x-\mu_0)\frac{\sigma_0^2}{\sigma^2+\sigma_0^2}$$

Prior plus data correction term

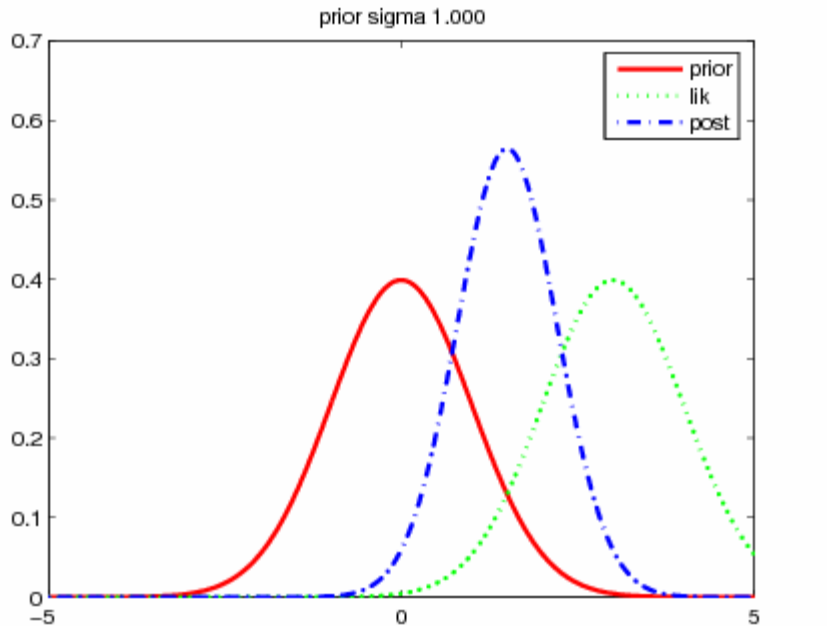
$$= x - (x-\mu_0)\frac{\sigma^2}{\sigma^2+\sigma_0^2}$$

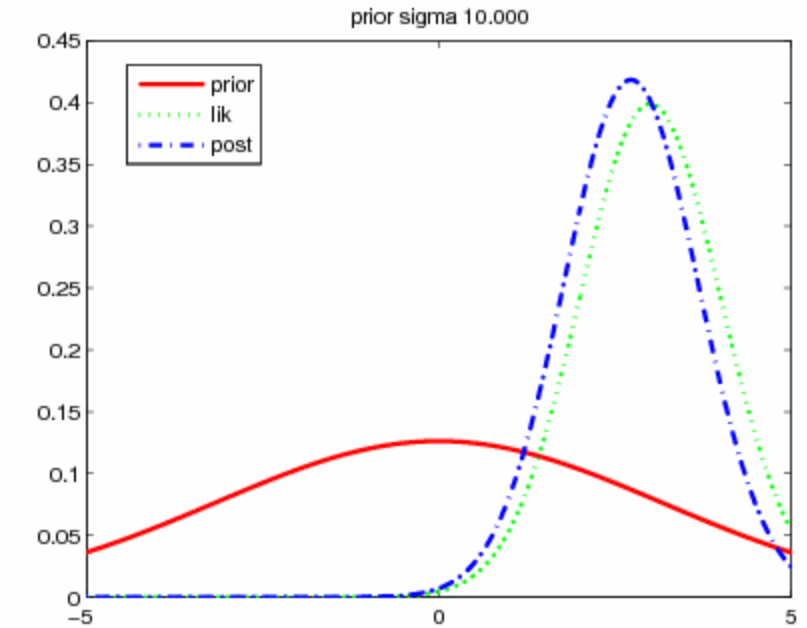
Data shrunk towards prior

Posterior

- Precision = 1/variance, $\lambda = 1/\sigma^2$.
- Precisions add, means are averaged.

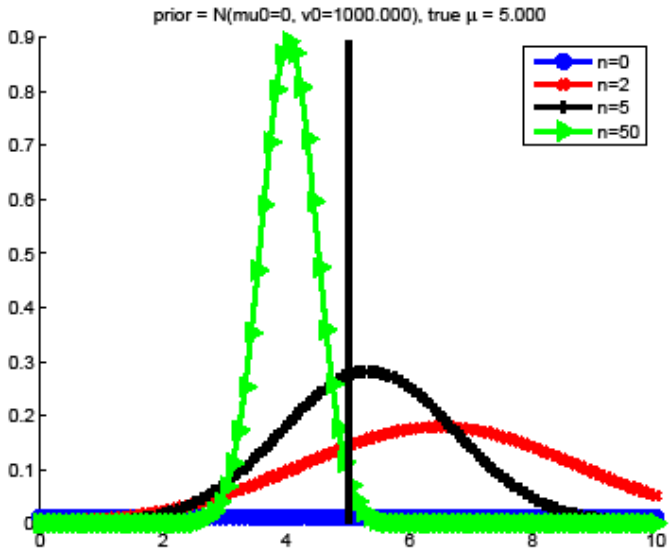
$$
\begin{aligned}
p(\mu|D,\lambda) &= \mathcal{N}(\mu|\mu_N,\lambda_N) \\
\lambda_N &= \lambda_0 + N\lambda \\
\mu_N &= \frac{\overline{x}N\lambda + \mu_0\lambda_0}{\lambda_N} = w\overline{x} + (1-w)\mu_0
\end{aligned}
$$





Sequential updating

- p(μ|D) rapidly approaches a delta function centered on the true mean.



Posterior predictive distribution

- The predictive variance is the observation noise σ^2 plus the uncertainty about μ, σ^2_N

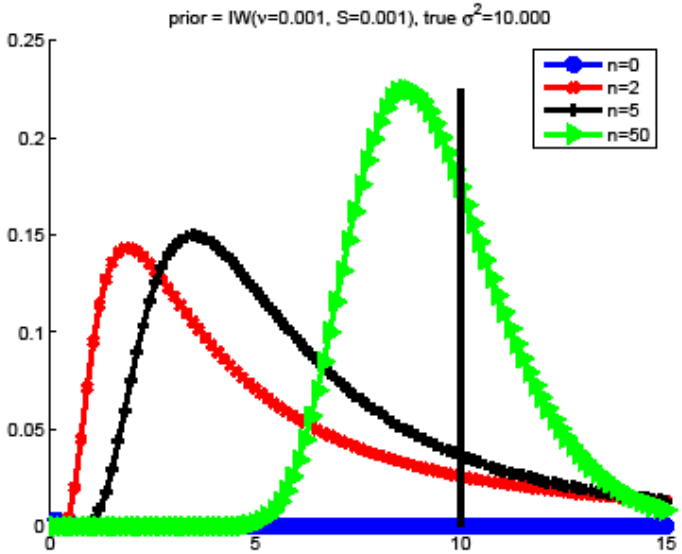
$$
\begin{aligned}
p(x|D) &= \int p(x|\mu)p(\mu|D)d\mu \\
&= \int \mathcal{N}(x|\mu,\sigma^2)\mathcal{N}(\mu|\mu_N,\sigma_N^2)d\mu \\
&= \mathcal{N}(x|\mu_N,\sigma_N^2+\sigma^2)
\end{aligned}
$$

Unknown variance

$$
\begin{aligned}
p(\mathcal{D}|\sigma^2) &\propto (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right) \\
p(\sigma^2) &= \mathrm{IW}(\sigma^2|a_0,b_0) \propto (\sigma^2)^{-\frac{a_0+2}{2}} \exp(-\frac{b_0}{2\sigma^2}) \\
p(\sigma^2|\mathcal{D}) &= \mathrm{IW}(\sigma^2|a_n,b_n) \\
a_n &= a_0+n \\
b_n &= b_0+\sum_{i=1}^{n}(x_i-\mu)^2
\end{aligned}
$$

Inverse wishart = inverse Gamma

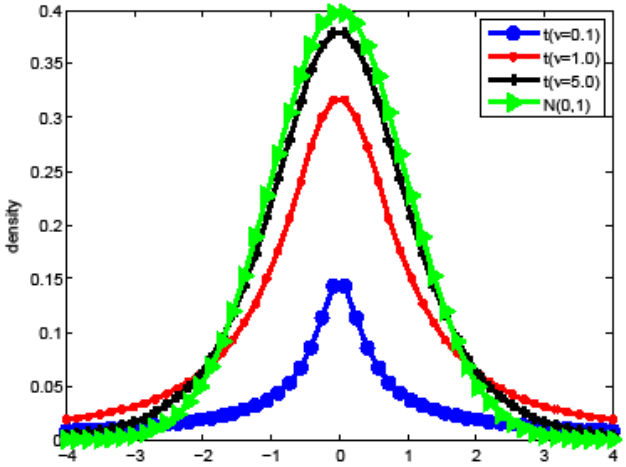
$$\mathrm{IW}(\sigma^2|a,b) = IG(\sigma^2|\frac{a}{2},\frac{b}{2})$$

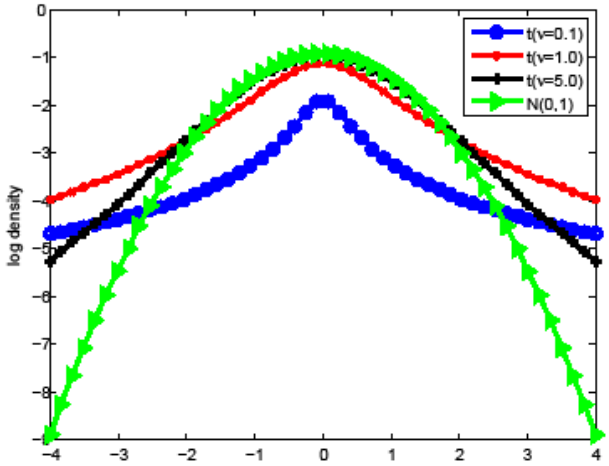


Posterior predictive

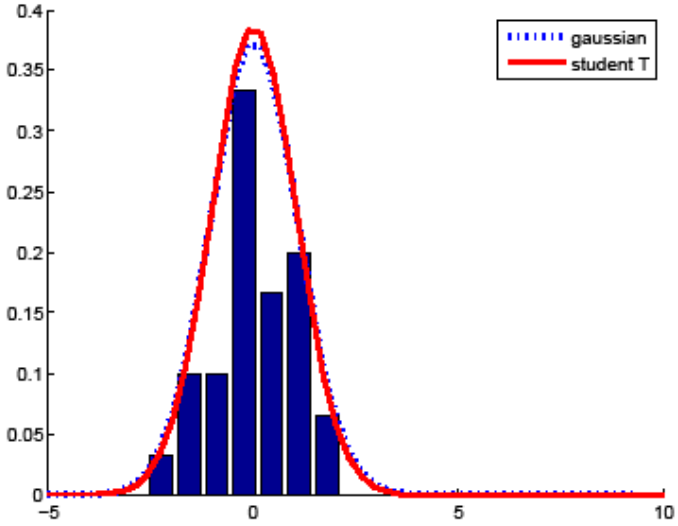
- Student T (infinite mixture of Gaussians)

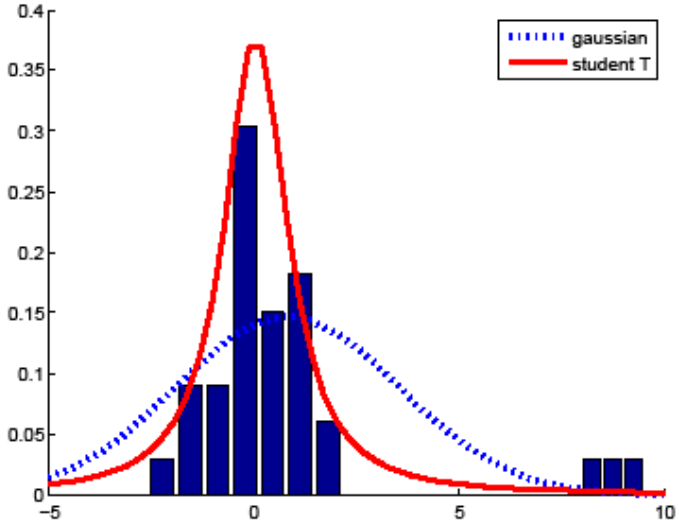
$$p(x|\mu, \mathcal{D}) = \int \mathcal{N}(x|\mu, \sigma^2)\mathrm{IW}(\sigma^2|a_n, b_n)d\sigma^2 = \mathcal{T}(x|a_n, \mu, \frac{b_n}{a_n})$$





Robustness of Student T





Unknown mean and variance

- Factored prior is not conjugate

$$p(\mu, \sigma^2) = p(\mu)p(\sigma^2) = \mathcal{N}(\mu|m_0, \frac{1}{\kappa_0})\mathsf{IW}(\sigma^2|a_0, b_0)$$

- Use NIW instead

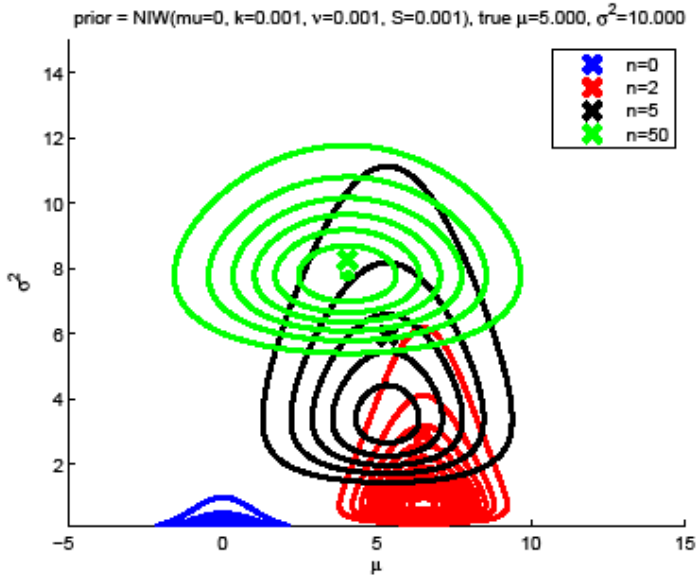
$$NIW(\mu, \sigma^2|m_0, \kappa_0, a_0, b_0) \stackrel{\text{def}}{=} \mathcal{N}(\mu|m_0, \frac{\sigma^2}{\kappa_0})\mathsf{IW}(\sigma^2|a_0, b_0)$$

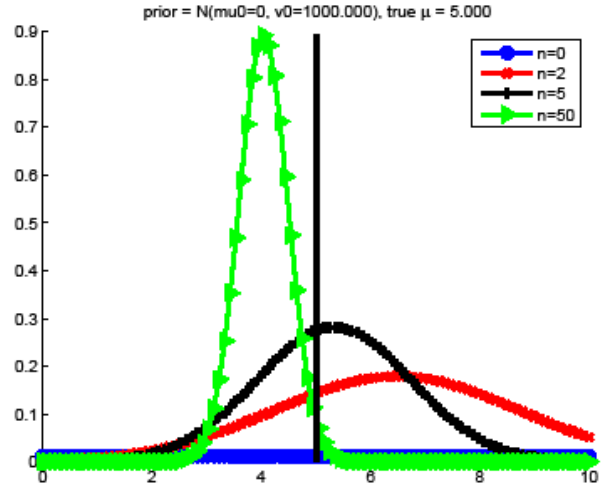
$$\propto \sigma^{-1}(\sigma^2)^{-(a_0/2+1)} \exp\left(-\frac{1}{2\sigma^2}[b_0 + \kappa_0(m_0 - \mu)^2]\right)$$

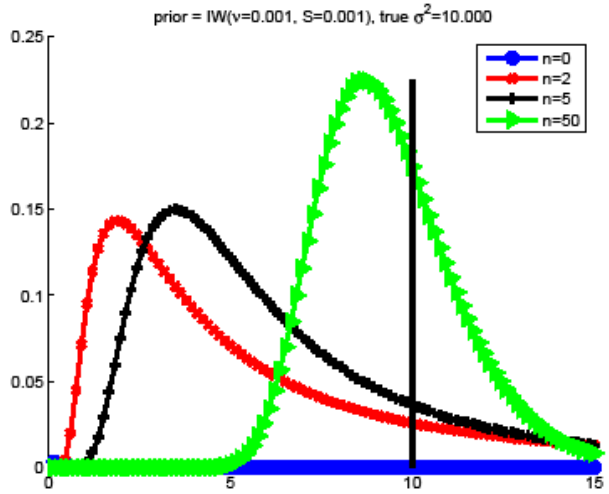
- Posterior is also NIW

$$\begin{aligned} p(\mu, \sigma^2|\mathcal{D}) &= NIW(\mu, \sigma^2|m_n, \kappa_n, a_n, b_n) \\ m_n &= \frac{\kappa_0 m_0 + n\overline{x}}{\kappa_n} \\ \kappa_n &= \kappa_0 + n \\ a_n &= a_0 + n \\ b_n &= b_0 + \sum_{i=1}^{n}(x_i - \overline{x})^2 + \frac{n\kappa_0}{\kappa_0 + n}(m_0 - \overline{x})^2 \end{aligned}$$

NIW


prior = NIW(mu=0, k=0.001, ν=0.001, S=0.001), true μ=5.000, σ²=10.000
n=0
n=2
n=5
n=50
σ²
μ


prior = N(mu0=0, v0=1000.000), true μ = 5.000
n=0
n=2
n=5
n=50


prior = IW(ν=0.001, S=0.001), true σ²=10.000
n=0
n=2
n=5
n=50

Marginals and predictive

$$
\begin{aligned}
p(\sigma^2|\mathcal{D}) &= \int p(\mu, \sigma^2|\mathcal{D}) d\mu = \int p(\sigma^2|\mathcal{D}) p(\mu|\sigma^2, \mathcal{D}) d\mu = \mathsf{IW}(\sigma^2|a_n, b_n) \\
p(\mu|\mathcal{D}) &= \int p(\mu, \sigma^2|D) d\sigma^2 = \mathcal{T}(\mu|a_n, m_n, \frac{b_n}{a_n \kappa_n}) \\
p(x|\mathcal{D}) &= \int \int \mathcal{N}(x|\mu, \sigma^2) \mathsf{NIW}(\mu, \sigma^2|m_n, \kappa_n, a_n, b_n) d\mu d\sigma^2 \\
&= \mathcal{T}\left(x|a_n, m_n, \frac{b_n(\kappa_n + 1)}{b_n \kappa_n}\right)
\end{aligned}
$$

Outline

- Bayesian estimation of
  - Gaussians
  - Generative classifiers
  - MVN
  - Linear regression
  - Logistic regression

MLE for gen classif

$$
\begin{aligned}
p(\mathcal{D}|\boldsymbol{\theta}) &= \prod_{i=1}^{n} p(y_i|\boldsymbol{\pi})p(\mathbf{x}_i|y_i,\boldsymbol{\phi}) \\
&= \prod_{c=1}^{C} \pi_c^{n_c} \prod_{i:y_i=c} p(\mathbf{x}_i|\boldsymbol{\phi}_c) \\
\ell(\boldsymbol{\theta}) &= \left[\sum_c n_c \log \pi_c\right] + \sum_{c=1}^{C}\left[\sum_{i=1}^{n} I(y_i=c)\mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c)\right] \\
\hat{\pi}_k &= \frac{N_k}{N} \\
\hat{\boldsymbol{\mu}}_c &= \frac{1}{N_c}\sum_{i:y_i=c}\mathbf{x}_i \\
\hat{\Sigma}_c &= \frac{1}{N_c}\sum_{i:y_i=c}(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_c)(\mathbf{x}_i-\hat{\boldsymbol{\mu}}_c)^T
\end{aligned}
$$

Plugin predictive for gen classif

$$
\begin{aligned}
p(y=c|\mathbf{x},\mathcal{D}) &= \int p(y=c|\mathbf{x},\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})d\boldsymbol{\theta} \\
&\approx \int p(y=c|\mathbf{x},\boldsymbol{\theta})\delta_{\hat{(\boldsymbol{\theta})}}(\boldsymbol{\theta})d\boldsymbol{\theta} \\
&= p(y=c|\mathbf{x},\hat{\boldsymbol{\theta}}) \\
&= \frac{\pi_c \mathcal{N}(\mathbf{x}|\hat{\boldsymbol{\mu}}_c,\hat{\boldsymbol{\Sigma}}_c)}{\sum_{k=1}^{C}\pi_k \mathcal{N}(\mathbf{x}|\hat{\boldsymbol{\mu}}_k,\hat{\boldsymbol{\Sigma}}_k)}
\end{aligned}
$$

Bayesian gen classif

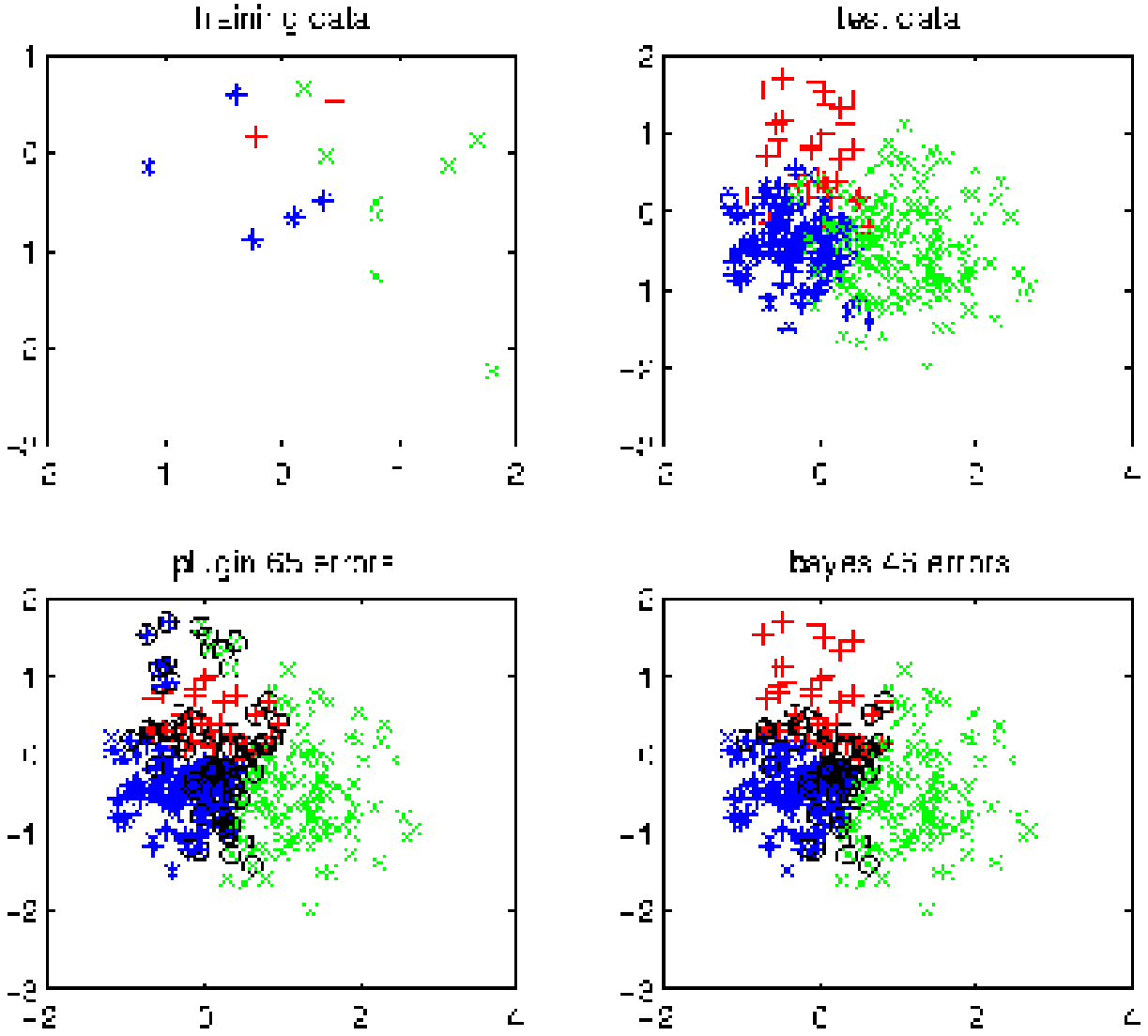
- Just “fit” separate density to every class and feature

$$
\begin{aligned}
p(y=c|\mathbf{x},\mathcal{D}) &\propto p(y=c|\mathcal{D})p(\mathbf{x}|y=c,\mathcal{D}) \\
p(y=c|\mathcal{D}) &= \int \mathsf{Mu}(y=c|\boldsymbol{\pi},1)\mathsf{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}_n)d\boldsymbol{\pi} = \frac{N_c+\alpha_c}{\sum_k N_k+\alpha_k} \\
p(\mathbf{x}|y=c,\mathcal{D}) &= \int \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c)\mathsf{MVNIW}(\boldsymbol{\mu}_c,\boldsymbol{\Sigma}_c|\boldsymbol{\beta}_n)d\boldsymbol{\mu}_c d\boldsymbol{\Sigma}_c \\
&= \mathcal{T}(\mathbf{x}|\nu_{nc}-d+1,\mathbf{m}_{nc},\frac{\mathbf{S}_{nc}(\kappa_{nc}+1)}{\kappa_{nc}(\nu_{nc}-d+1)})
\end{aligned}
$$

Outline

- Bayesian estimation of
  - Gaussians
  - Generative classifiers
  - MVN
  - Linear regression
  - Logistic regression

Bayesian vs plugin

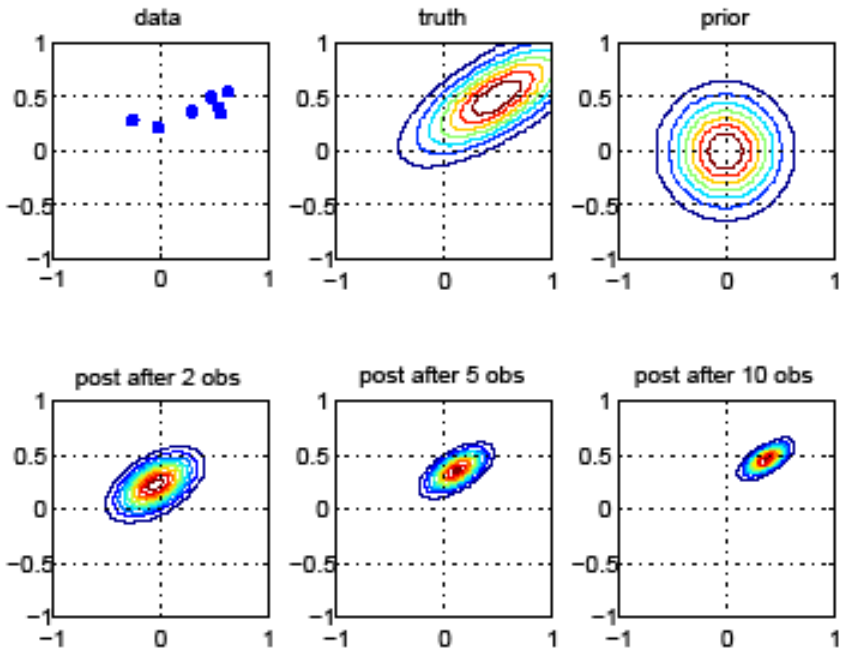


Outline

- Bayesian estimation of
  - Gaussians
  - Generative classifiers
  - MVN
  - Linear regression
  - Logistic regression

Unknown mean

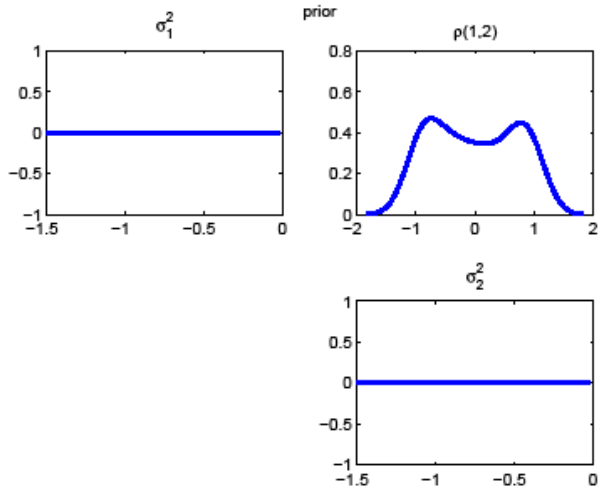
$$
\begin{aligned}
p(\boldsymbol{\mu}|\mathcal{D}, \boldsymbol{\Sigma}) &\propto \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \prod_{i=1}^{n} \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&= \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n) \\
\boldsymbol{\Sigma}_n^{-1} &= \boldsymbol{\Sigma}_0^{-1} + n\boldsymbol{\Sigma}^{-1} \\
\boldsymbol{\mu}_n &= \boldsymbol{\Sigma}_n(\boldsymbol{\Sigma}^{-1}(n\overline{\mathbf{x}}) + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0)
\end{aligned}
$$

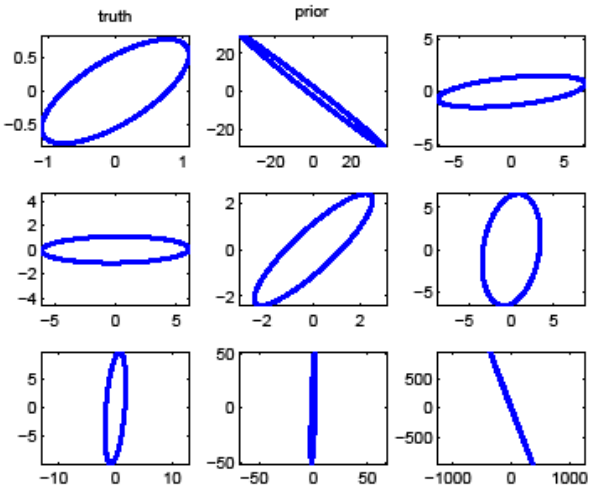


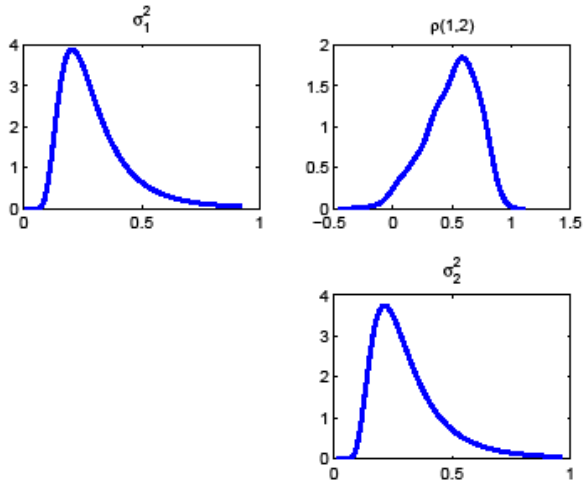
Unknown cov

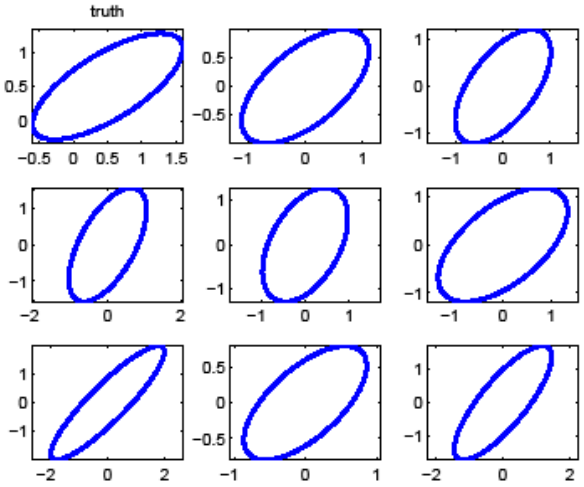
$$
\begin{aligned}
p(\boldsymbol{\Sigma}|\boldsymbol{\mu}, \mathcal{D}) &= \mathsf{IW}(\boldsymbol{\Sigma}|\nu_n, \mathbf{S}_n) \\
\nu_n &= \nu_0 + n \\
\mathbf{S}_n &= \mathbf{S}_0 + \mathbf{M} \\
\mathbf{M} &= \sum_{i=1}^{n} (x_i - \boldsymbol{\mu})(x_i - \boldsymbol{\mu})^T
\end{aligned}
$$

Unknown cov


prior
σ_1^2
$\rho(1,2)$
σ_2^2


truth
prior


σ_1^2
$\rho(1,2)$
σ_2^2


truth

Unknown mean and cov

$$
\begin{aligned}
\mathrm{MVNIW}(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \mathbf{m}_0, \kappa_0, \mathbf{S}_0, \nu_0) &\stackrel{\text{def}}{=} \mathcal{N}(\boldsymbol{\mu} | \mathbf{m}_0, \frac{1}{\kappa_0}\boldsymbol{\Sigma}) \times \mathrm{IW}(\boldsymbol{\Sigma} | \nu_0, \mathbf{S}_0) \\
&\propto |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{\kappa_0}{2}(\boldsymbol{\mu} - \mathbf{m}_0)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu} - \mathbf{m}_0)\right) \\
&\times |\boldsymbol{\Sigma}|^{-\frac{\nu_0 + d + 1}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{S}_0)\right)
\end{aligned}
$$

$$
\begin{aligned}
p(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \mathcal{D}) &= \mathrm{MVNIW}(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \mathbf{m}_n, \kappa_n, \mathbf{S}_n, \nu_n) \\
\mathbf{m}_n &= \frac{\kappa_0 \mathbf{m}_0 + n\overline{\mathbf{x}}}{\kappa_n} = \frac{\kappa_0}{\kappa_0 + n}\mathbf{m}_0 + \frac{n}{\kappa_0 + n}\overline{\mathbf{x}} \\
\kappa_n &= \kappa_0 + n \\
\nu_n &= \nu_0 + n \\
\mathbf{S}_n &= \mathbf{S}_0 + n\mathbf{S} + \frac{\kappa_0 n}{\kappa_0 + n}(\overline{\mathbf{x}} - \mathbf{m}_0)(\overline{\mathbf{x}} - \mathbf{m}_0)^T
\end{aligned}
$$

Outline

- Bayesian estimation of
  - Gaussians
  - Generative classifiers
  - MVN
  - Linear regression
  - Logistic regression

Bayesian linear regression

- Gaussian prior on weights.
- Assume σ^2 is known.

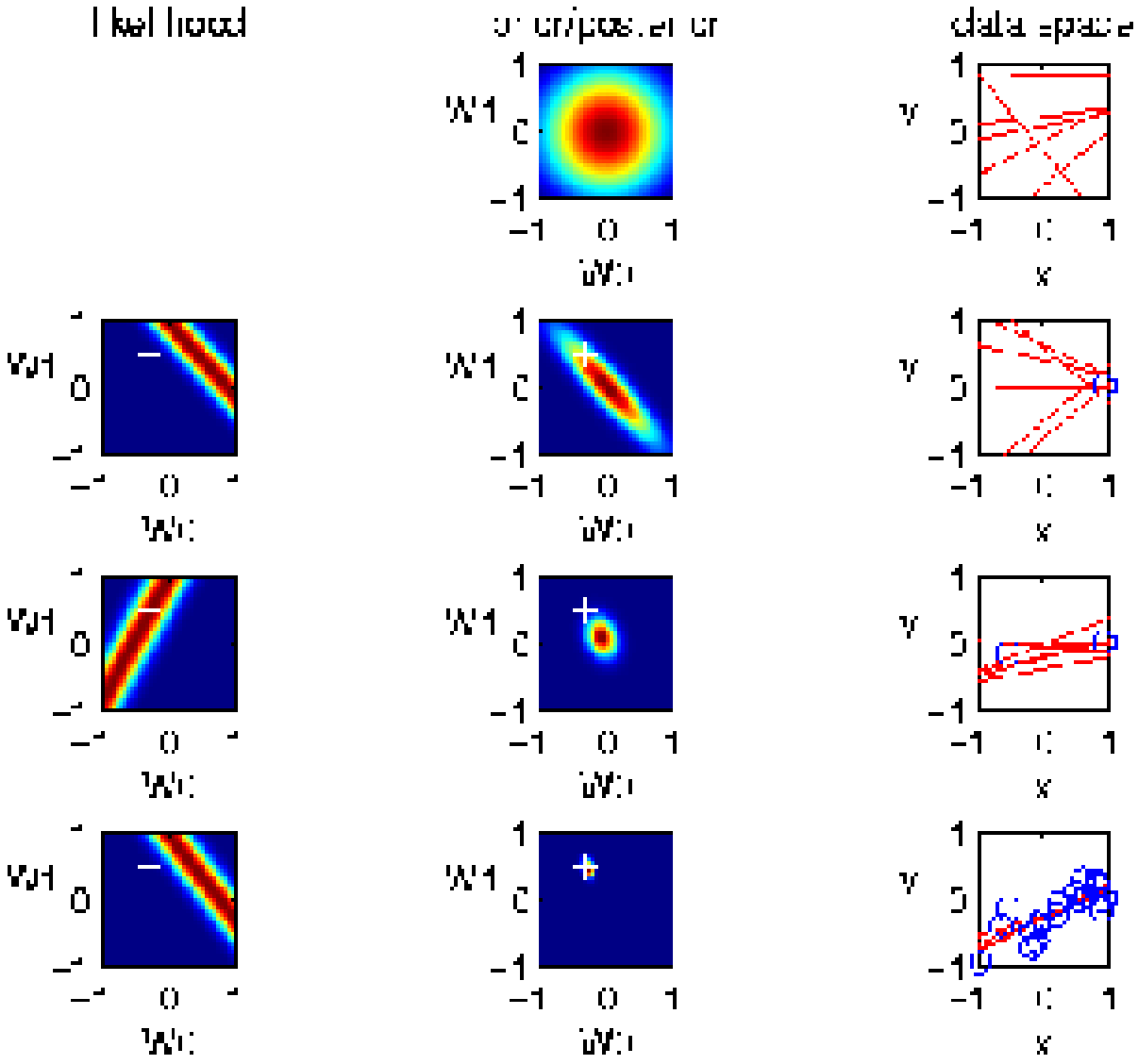
$$
\begin{aligned}
p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \sigma^2) &\propto \mathcal{N}(\mathbf{w}|\mathbf{w}_0, \mathbf{S}_0)\mathcal{N}(\mathbf{y}|\mathbf{X}\mathbf{w}, \sigma^2\mathbf{I}_n) \\
&= \mathcal{N}(\mathbf{w}|\mathbf{w}_n, \mathbf{S}_n) \\
\mathbf{w}_n &= \mathbf{S}_n\mathbf{S}_0^{-1}\mathbf{w}_0 + \frac{1}{\sigma^2}\mathbf{S}_n\mathbf{X}^T\mathbf{y} \\
\mathbf{S}_n^{-1} &= \mathbf{S}_0^{-1} + \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}
\end{aligned}
$$

Connection with Ridge regression

- Let prior be $w_0 = 0$, $S_0 = \tau^2 I$. Let $\lambda = \sigma^2/\tau^2$.

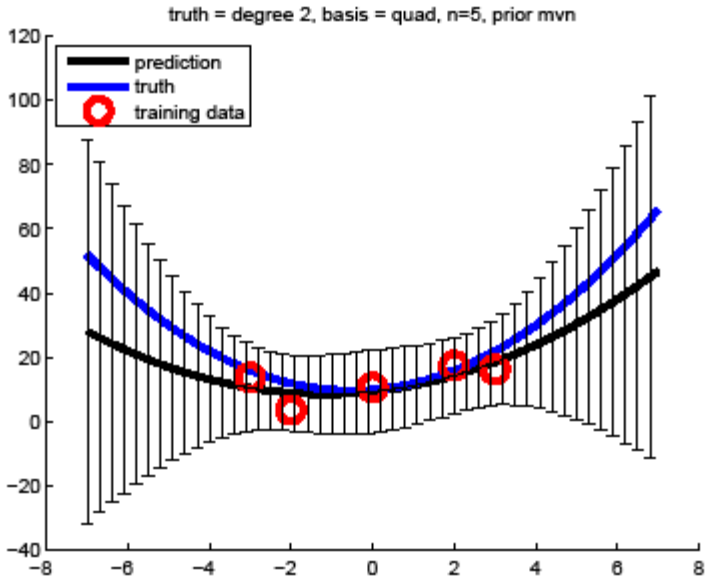
$$
\begin{aligned}
\mathbf{w}_n &= \frac{1}{\sigma^2}\mathbf{S}_n\mathbf{X}^T\mathbf{y} = \frac{1}{\sigma^2}\left(\frac{1}{\tau_0^2}\mathbf{I}_d + \frac{1}{\sigma^2}\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \\
&= \frac{1}{\sigma^2}\left(\frac{1}{\frac{1}{\sigma^2}}(\frac{\frac{1}{\tau_0^2}}{\frac{1}{\sigma^2}}\mathbf{I}_d + \mathbf{X}^T\mathbf{X})\right)^{-1}\mathbf{X}^T\mathbf{y} \\
&= \left(\lambda\mathbf{I}_d + \mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}
\end{aligned}
$$

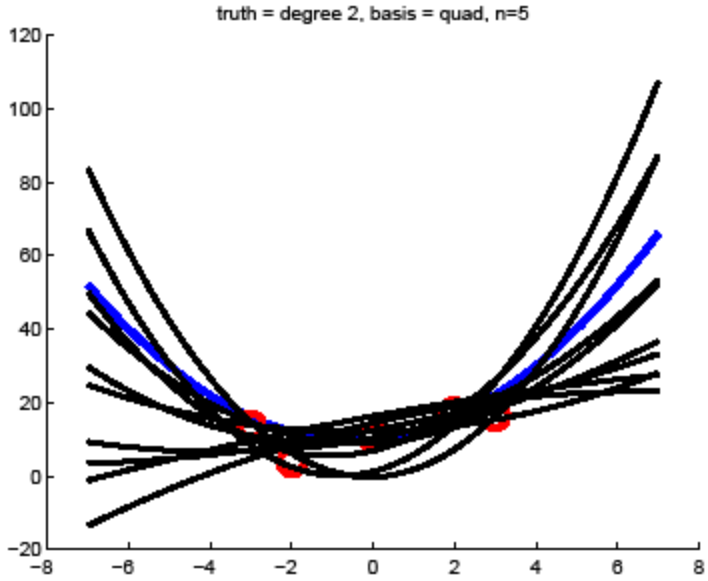
2d example



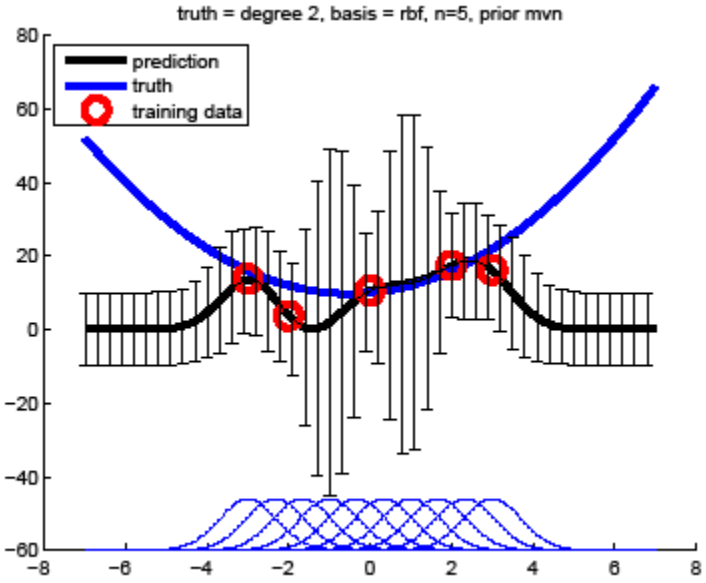
Posterior predictive

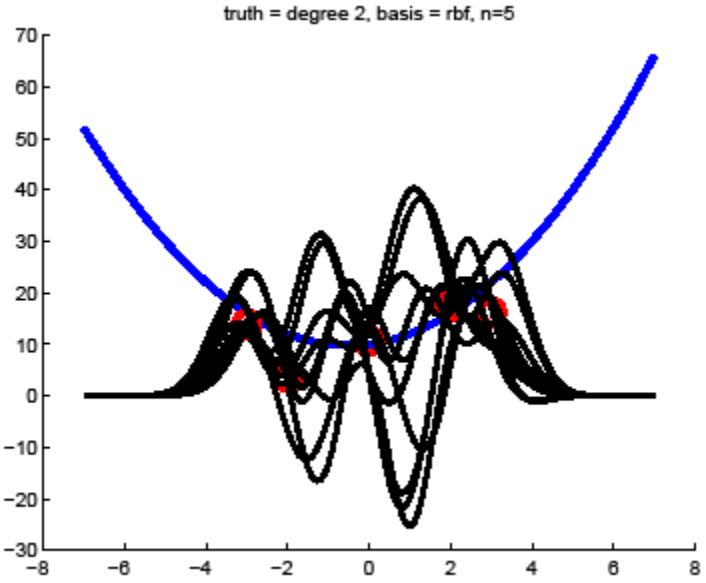
$$
\begin{aligned}
p(y|\mathbf{x}, \mathcal{D}, \sigma^2, \mathbf{w}_0, \tau_0^2) &= \int \mathcal{N}(y|\mathbf{x}^T\mathbf{w}, \sigma^2)\mathcal{N}(\mathbf{w}|\mathbf{w}_n, \mathbf{S}_n)d\mathbf{w} \\
&= \mathcal{N}(y|\mathbf{w}_n^T\mathbf{x}, \sigma_n^2(\mathbf{x})) \\
\sigma_n^2(\mathbf{x}) &= \sigma^2 + \mathbf{x}^T\mathbf{S}_n\mathbf{x}
\end{aligned}
$$





RBF basis

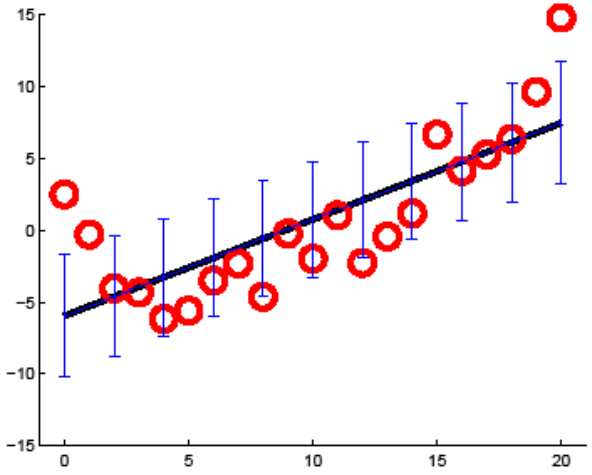


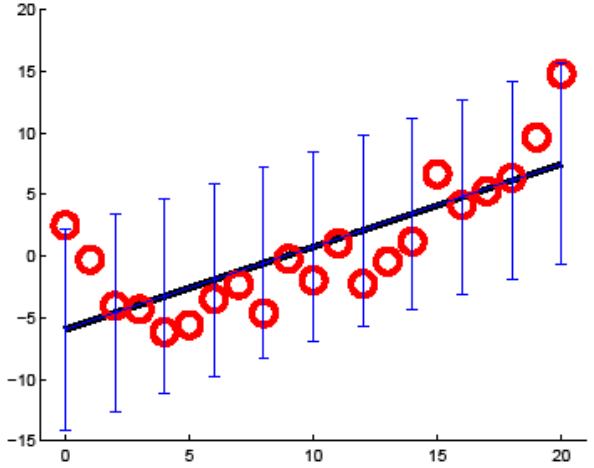


Underestimates uncertainty far from data – GPs will fix this

Handling σ^2

- We can put an IW prior on σ^2.
- When we integrate out σ^2, the posterior predictive becomes a T distribution.





Outline

- Bayesian estimation of
  - Gaussians
  - Generative classifiers
  - MVN
  - Linear regression
  - Logistic regression

Logistic regression

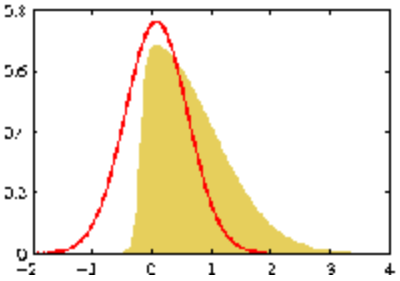
- No conjugate prior. Use Laplace approximation.

$$p(\mathbf{w}|\mathcal{D}) \propto \mathcal{N}(\mathbf{w}|\mathbf{w}_0, \mathbf{C}_0) \prod_{i=1}^{n} \sigma(y_i \mathbf{w}^T \mathbf{x}_i) \approx \mathcal{N}(\mathbf{w}|\hat{\mathbf{w}}_{MAP}, \mathbf{C}_n)$$

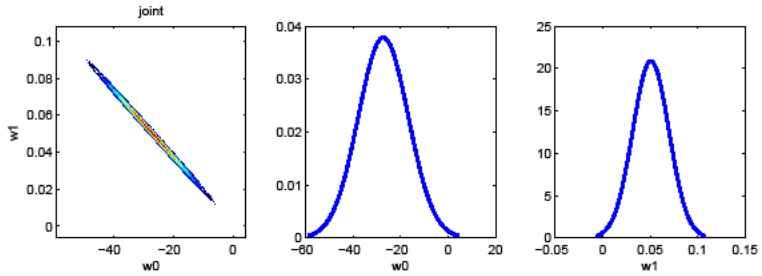
Laplace approximation

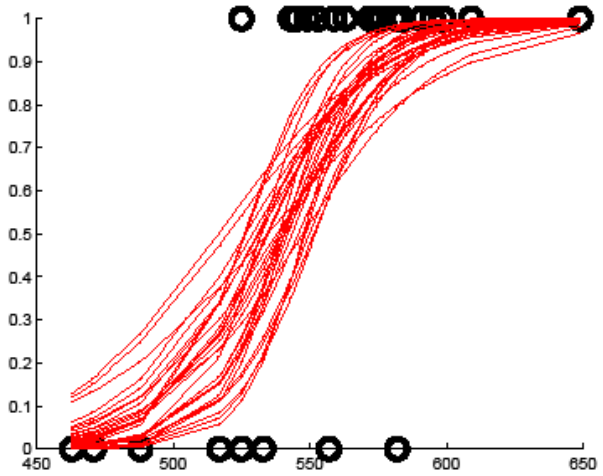
$$
\begin{aligned}
f(\boldsymbol{\theta}) &= p(\boldsymbol{\theta}, D) \\
\ln f(\boldsymbol{\theta}) &\approx \ln f(\boldsymbol{\theta}_0) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \mathbf{H} (\boldsymbol{\theta} - \boldsymbol{\theta}_0) \\
\hat{f}(\boldsymbol{\theta}) &= f(\boldsymbol{\theta}_0) \exp\left[-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \mathbf{C}^{-1} (\boldsymbol{\theta} - \boldsymbol{\theta}_0) \right] \\
\mathbf{C} &= -\mathbf{H}^{-1}
\end{aligned}
$$

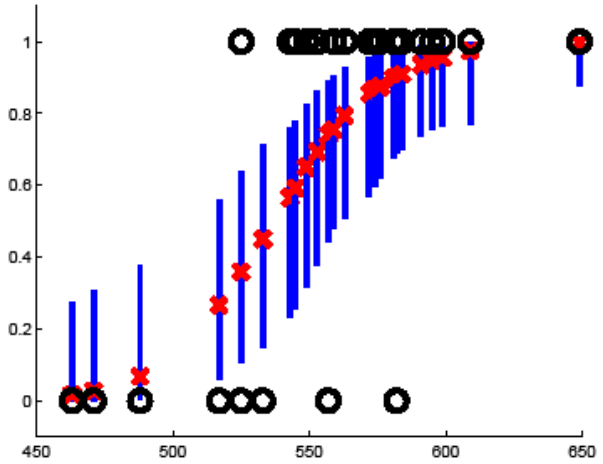
Use ILRS to compute θ_0 and H



SAT example







2d example