Introduction to Machine Learning

Brown University CSCI 1950-F, Spring 2012
Prof. Erik Sudderth

Lecture 7:
Multivariate Gaussian Distributions
Preview of Gaussian Classification

Many figures courtesy Kevin Murphy's textbook,
Machine Learning: A Probabilistic Perspective

A Learning Toolbox

Tuning Model Parameters to Data

- Maximum likelihood (ML) estimation: Frequency matching
- Bayesian MAP estimation: Regularized ML
- Bayesian posterior predictive distribution, posterior means, …

Model Selection

- Empirical performance on held-out data: Cross-validation
- Model complexity penalization: Bayesian marginal likelihood
- Approximations coming later: Laplace, AIC/BIC, …

Probability Distributions

- Dirichlet-multinoulli, beta-Bernoulli: Discrete data
- Multivariate normal, gamma: Continuous data

Useful Model Families: Remainder of course…

Covariance and Correlation

Covariance:

$$\mathrm{cov}\left[X,Y\right] \triangleq \mathbb{E}\left[(X-\mathbb{E}\left[X\right])(Y-\mathbb{E}\left[Y\right])\right] = \mathbb{E}\left[XY\right] - \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right]$$

$$\mathrm{cov}\left[\mathbf{x}\right] \triangleq \mathbb{E}\left[(\mathbf{x}-\mathbb{E}\left[\mathbf{x}\right])(\mathbf{x}-\mathbb{E}\left[\mathbf{x}\right])^T\right] = \begin{pmatrix} \mathrm{var}\left[X_1\right] & \mathrm{cov}\left[X_1,X_2\right] & \cdots & \mathrm{cov}\left[X_1,X_d\right] \\ \mathrm{cov}\left[X_2,X_1\right] & \mathrm{var}\left[X_2\right] & \cdots & \mathrm{cov}\left[X_2,X_d\right] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}\left[X_d,X_1\right] & \mathrm{cov}\left[X_d,X_2\right] & \cdots & \mathrm{var}\left[X_d\right] \end{pmatrix}$$

$$\Sigma \in \mathbb{R}^{d\times d}$$

Always *positive semidefinite:* $u^T\Sigma u \geq 0 \text{ for any } u \in \mathbb{R}^{d\times 1}, u \neq 0$

Often *positive definite:* $u^T\Sigma u > 0 \text{ for any } u \in \mathbb{R}^{d\times 1}, u \neq 0$

Correlation:

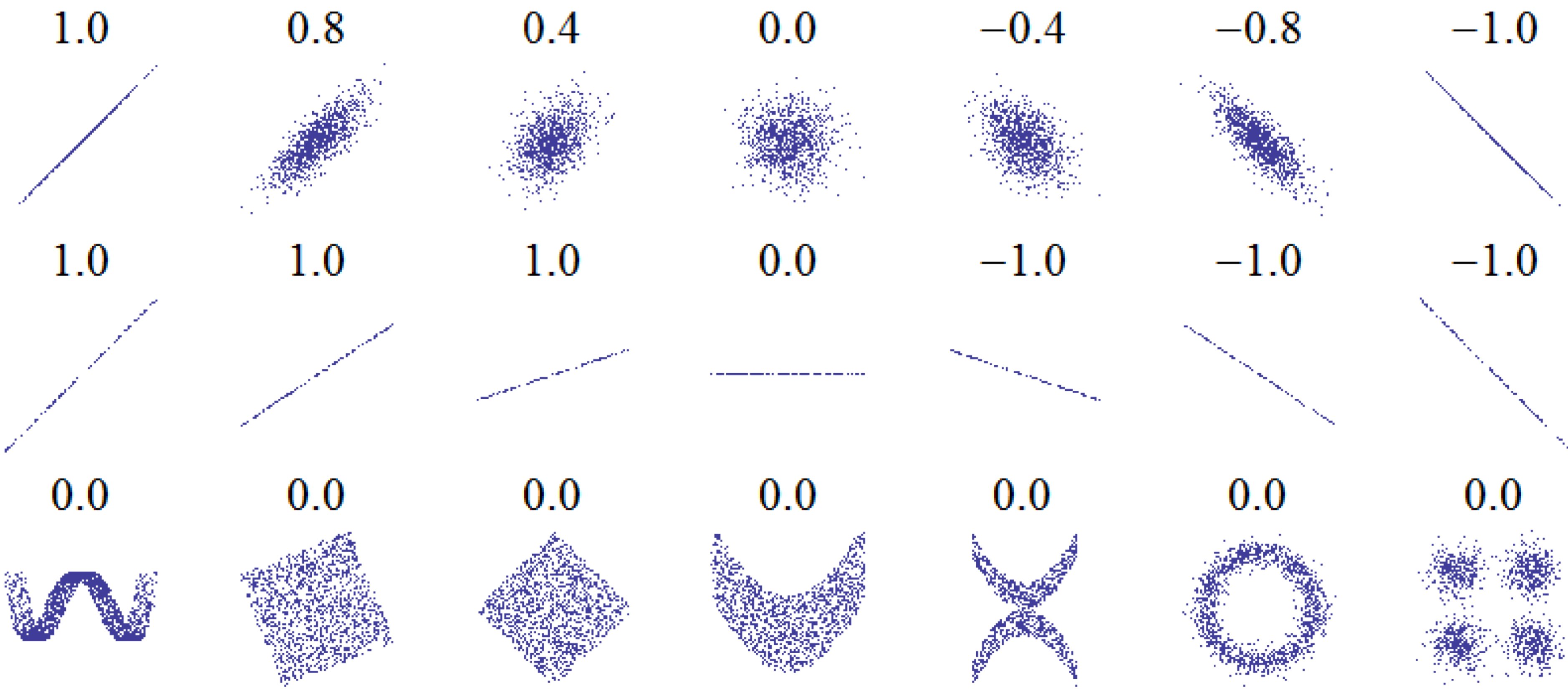
$$\mathrm{corr}\left[X,Y\right] \triangleq \frac{\mathrm{cov}\left[X,Y\right]}{\sqrt{\mathrm{var}\left[X\right]\mathrm{var}\left[Y\right]}} \qquad -1 \leq \mathrm{corr}\left[X,Y\right] \leq 1$$

$$\mathbf{R} = \begin{pmatrix} \mathrm{corr}\left[X_1,X_1\right] & \mathrm{corr}\left[X_1,X_2\right] & \cdots & \mathrm{corr}\left[X_1,X_d\right] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{corr}\left[X_d,X_1\right] & \mathrm{corr}\left[X_d,X_2\right] & \cdots & \mathrm{corr}\left[X_d,X_d\right] \end{pmatrix}$$

Independence:

$$p(X,Y) = p(X)p(Y) \Rightarrow \mathrm{cov}\left[X,Y\right] = 0 \Leftrightarrow \mathrm{corr}\left[X,Y\right] = 0$$

Empirical Correlations



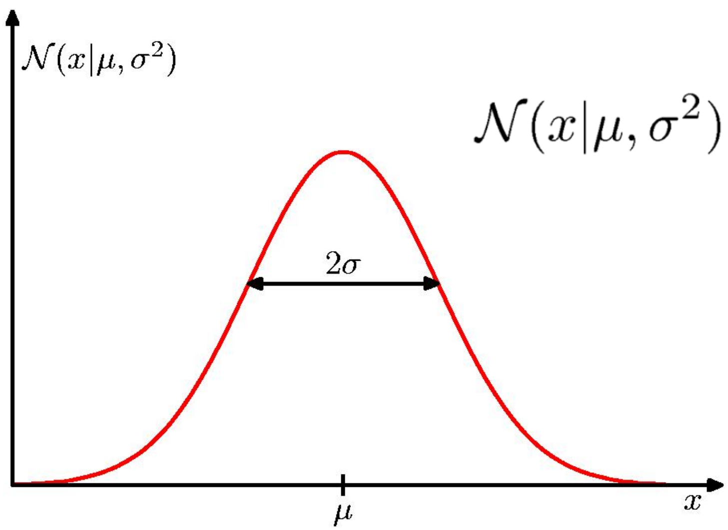
Sample mean:

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

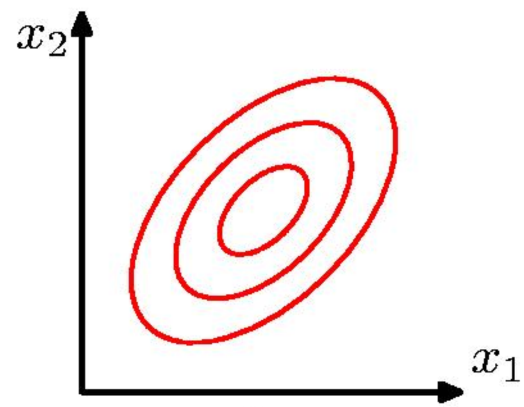
Sample covariance:

$$\hat{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \hat{\mu})(x_i - \hat{\mu})^T$$

Gaussian Distributions



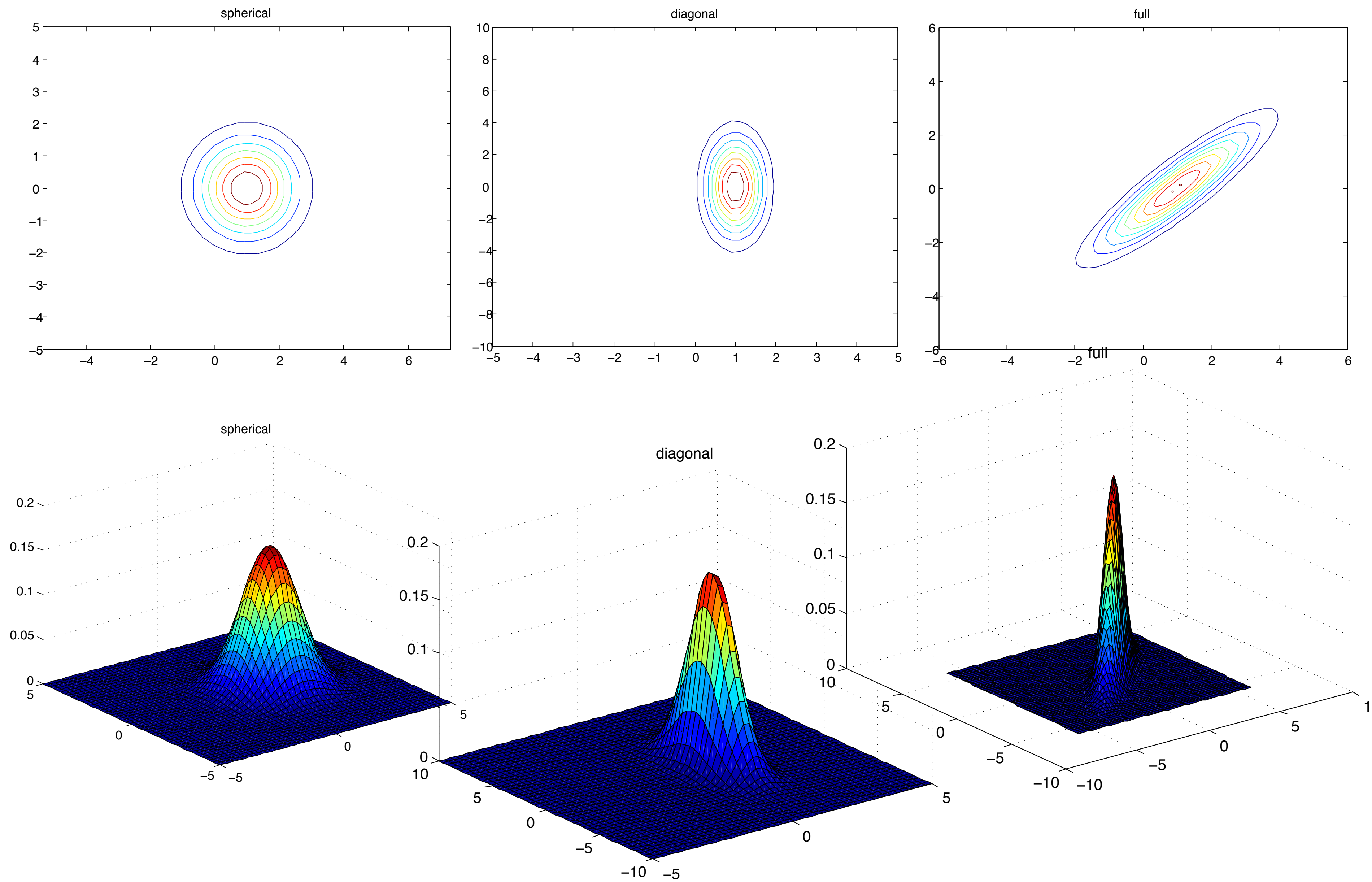
$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$



$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

- Simplest joint distribution that can capture arbitrary mean & covariance
- Justifications from *central limit theorem* and *maximum entropy* criterion
- Probability density above assumes covariance is *positive definite*
- ML parameter estimates are *sample mean* & *sample covariance*

Two-Dimensional Gaussians



Gaussian Geometry

- Eigenvalues and eigenvectors:

$$\Sigma u_i = \lambda_i u_i, i = 1, \ldots, d$$

- For a *symmetric* matrix:

$$\lambda_i \in \mathbb{R} \qquad u_i^T u_i = 1 \qquad u_i^T u_j = 0$$

$$\Sigma = U \Lambda U^T = \sum_{i=1}^{d} \lambda_i u_i u_i^T$$

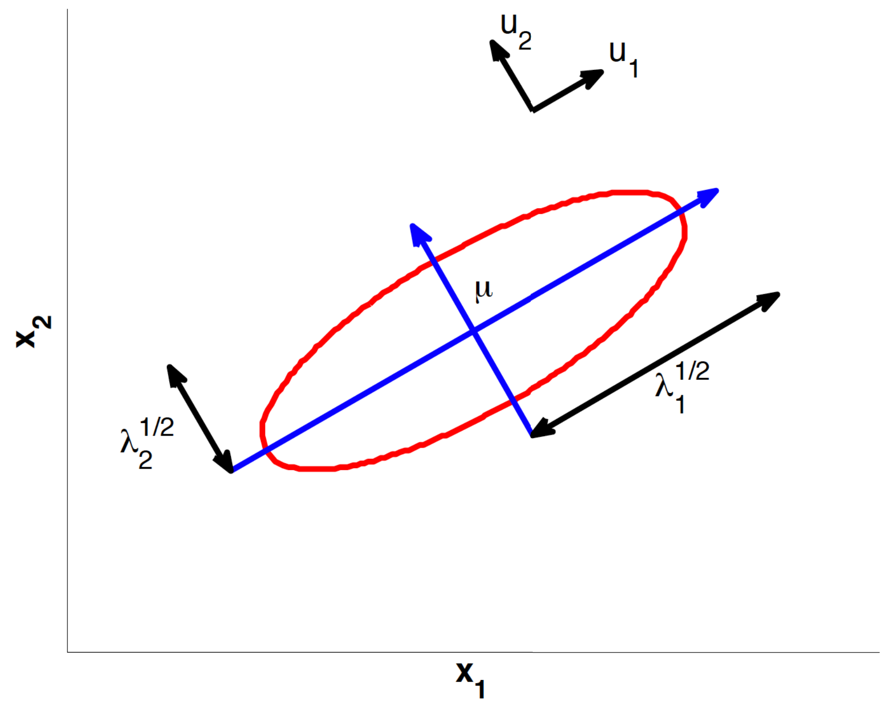
- For a *positive semidefinite* matrix:

$$\lambda_i \geq 0$$

- For a *positive definite* matrix:

$$\lambda_i > 0$$

$$\Sigma^{-1} = U \Lambda^{-1} U^T = \sum_{i=1}^{d} \frac{1}{\lambda_i} u_i u_i^T$$

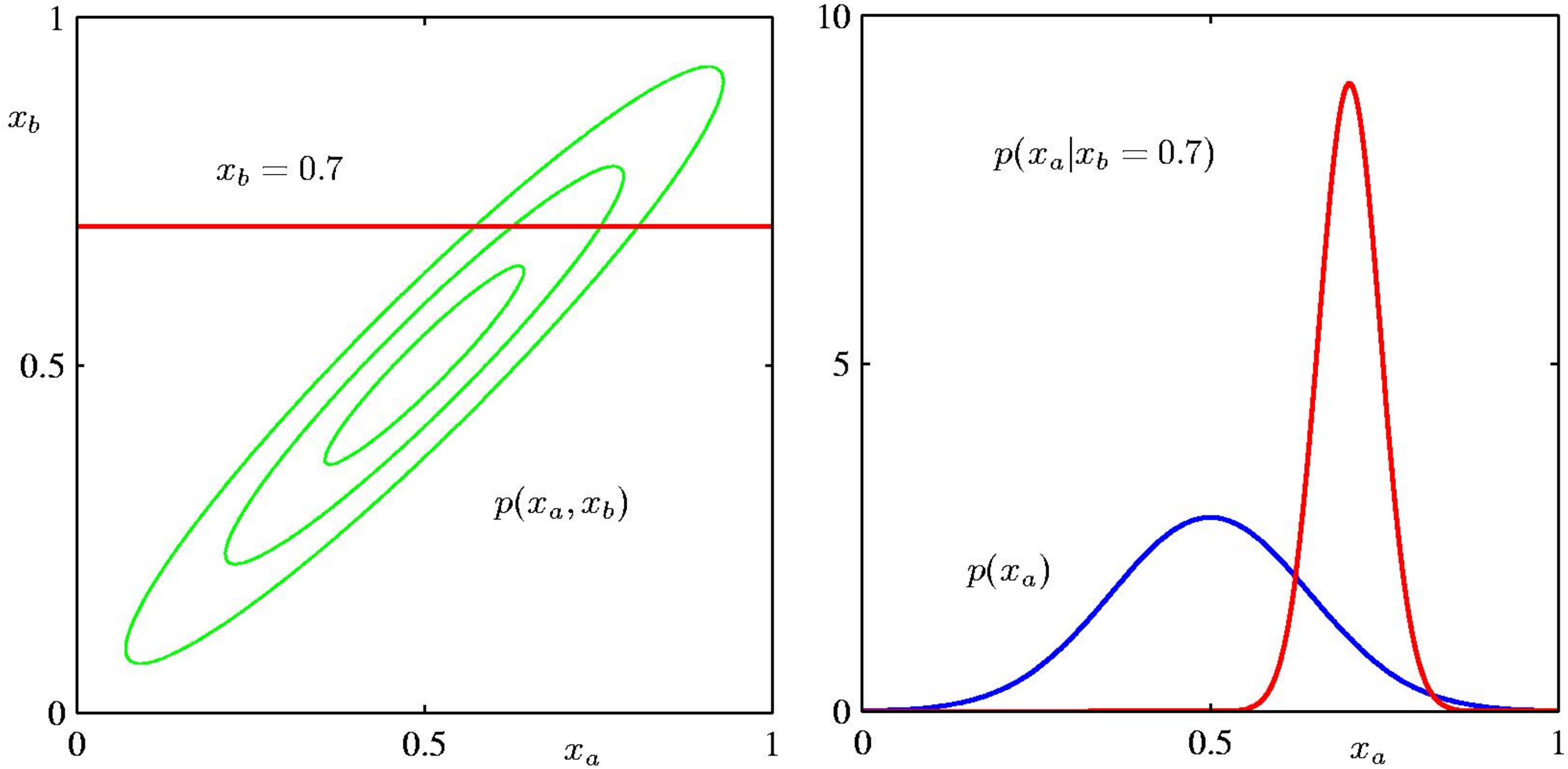


$$\Delta^2 = (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

$$\Delta^2 = \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i}$$

$$y_i = u_i^T (x - \mu)$$

Gaussian Conditionals & Marginals



For any joint multivariate Gaussian distribution,
all marginal distributions are Gaussians,
and all conditional distributions are Gaussians

Partitioned Gaussian Distributions

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}, \quad \boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \boldsymbol{\Lambda}_{11} & \boldsymbol{\Lambda}_{12} \\ \boldsymbol{\Lambda}_{21} & \boldsymbol{\Lambda}_{22} \end{pmatrix}$$

Marginals:

$$\begin{aligned} p(\mathbf{x}_1) &= \mathcal{N}(\mathbf{x}_1|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) \\ p(\mathbf{x}_2) &= \mathcal{N}(\mathbf{x}_2|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22}) \end{aligned}$$

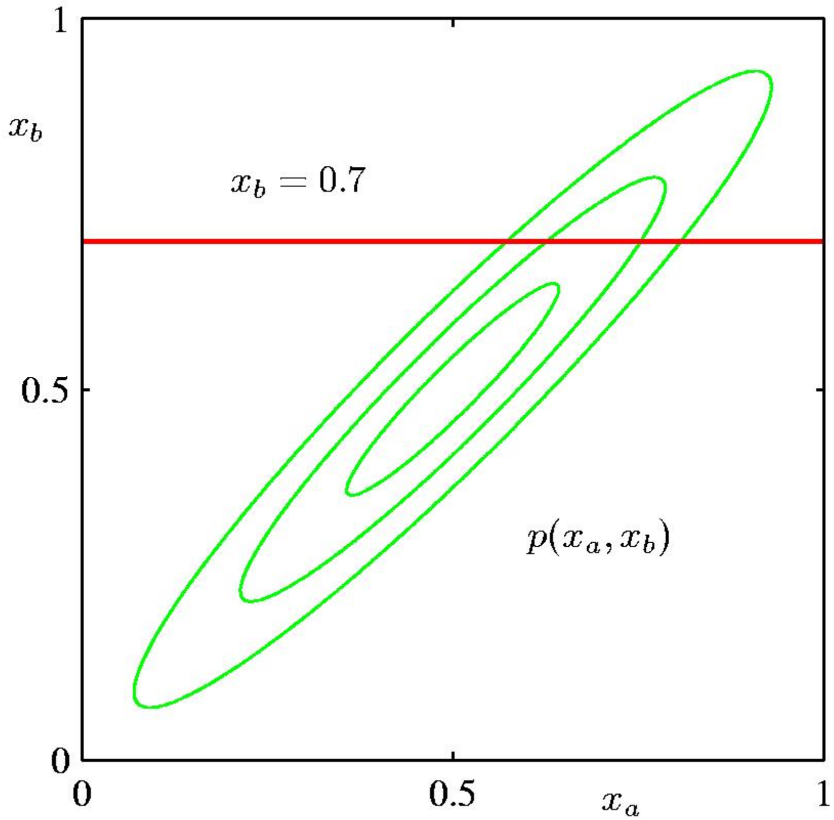
Conditionals:

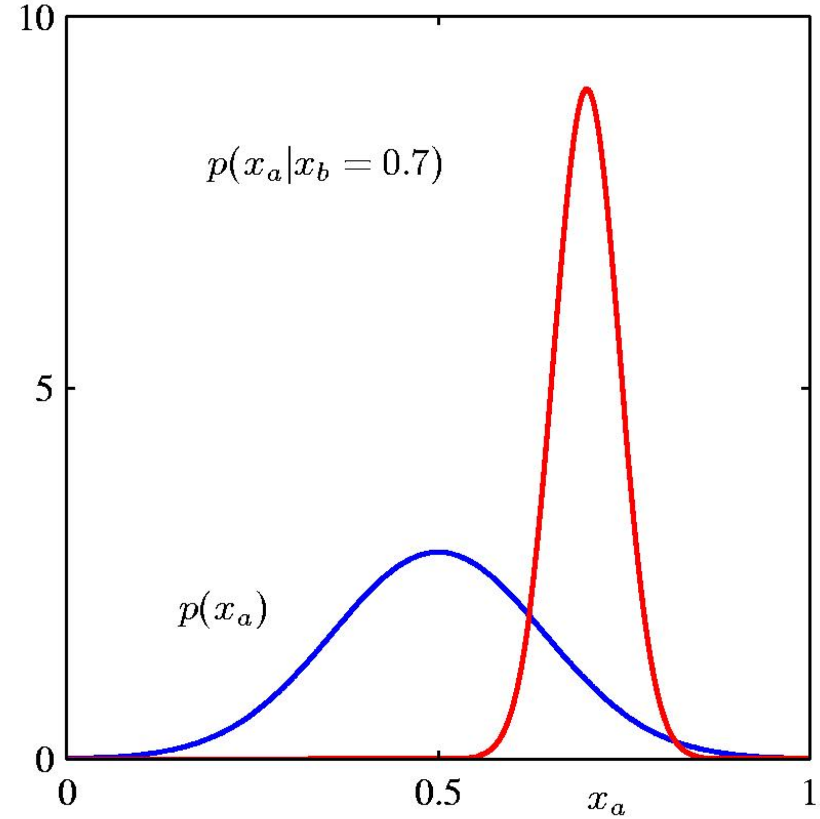
$$\begin{aligned} p(\mathbf{x}_1|\mathbf{x}_2) &= \mathcal{N}(\mathbf{x}_1|\boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{1|2}) \\ \boldsymbol{\mu}_{1|2} &= \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \\ &= \boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{11}^{-1}\boldsymbol{\Lambda}_{12}(\mathbf{x}_2 - \boldsymbol{\mu}_2) \\ &= \boldsymbol{\Sigma}_{1|2}\left(\boldsymbol{\Lambda}_{11}\boldsymbol{\mu}_1 - \boldsymbol{\Lambda}_{12}(\mathbf{x}_2 - \boldsymbol{\mu}_2)\right) \\ \boldsymbol{\Sigma}_{1|2} &= \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} = \boldsymbol{\Lambda}_{11}^{-1} \end{aligned}$$

Intuition:

$$p(x_1 \mid x_2) = \frac{p(x_1, x_2)}{p(x_2)} \propto \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) + \frac{1}{2}(x_2-\mu_2)^T\Sigma_{22}^{-1}(x_2-\mu_2)\right\}$$

Gaussian Conditionals & Marginals





$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

$$p(x_1) = \mathcal{N}(x_1|\mu_1, \sigma_1^2)$$

$$p(x_1|x_2) \quad = \quad \mathcal{N}\left(x_1|\mu_1 + \frac{\rho\sigma_1\sigma_2}{\sigma_2^2}(x_2 - \mu_2),\ \sigma_1^2 - \frac{(\rho\sigma_1\sigma_2)^2}{\sigma_2^2}\right)$$

Linear Gaussian Systems

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x) \qquad p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \boldsymbol{\Sigma}_y)$$

Marginal Likelihood:

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu}_x + \mathbf{b}, \boldsymbol{\Sigma}_y + \mathbf{A}\boldsymbol{\Sigma}_x\mathbf{A}^T)$$

Posterior Distribution:

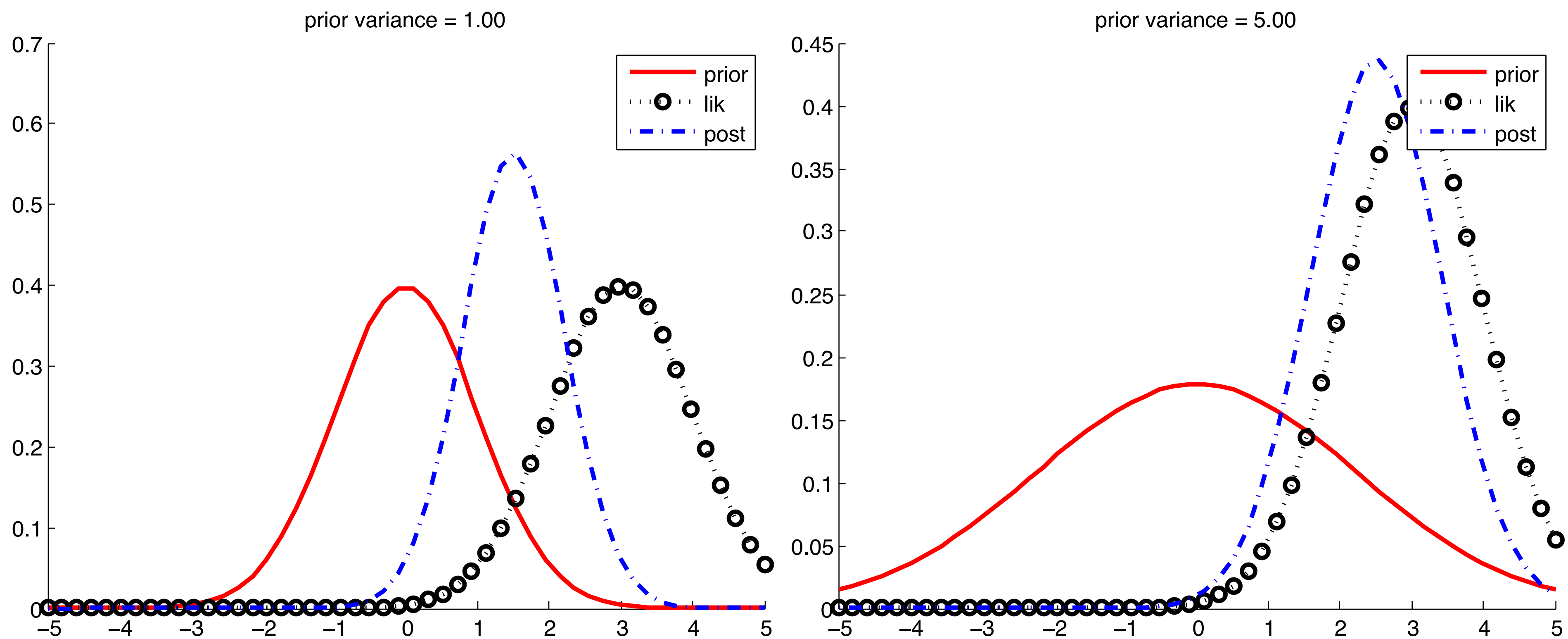
$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{x|y}, \boldsymbol{\Sigma}_{x|y})$$

$$\boldsymbol{\Sigma}_{x|y}^{-1} = \boldsymbol{\Sigma}_x^{-1} + \mathbf{A}^T\boldsymbol{\Sigma}_y^{-1}\mathbf{A}$$

$$\boldsymbol{\mu}_{x|y} = \boldsymbol{\Sigma}_{x|y}[\mathbf{A}^T\boldsymbol{\Sigma}_y^{-1}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Sigma}_x^{-1}\boldsymbol{\mu}_x]$$

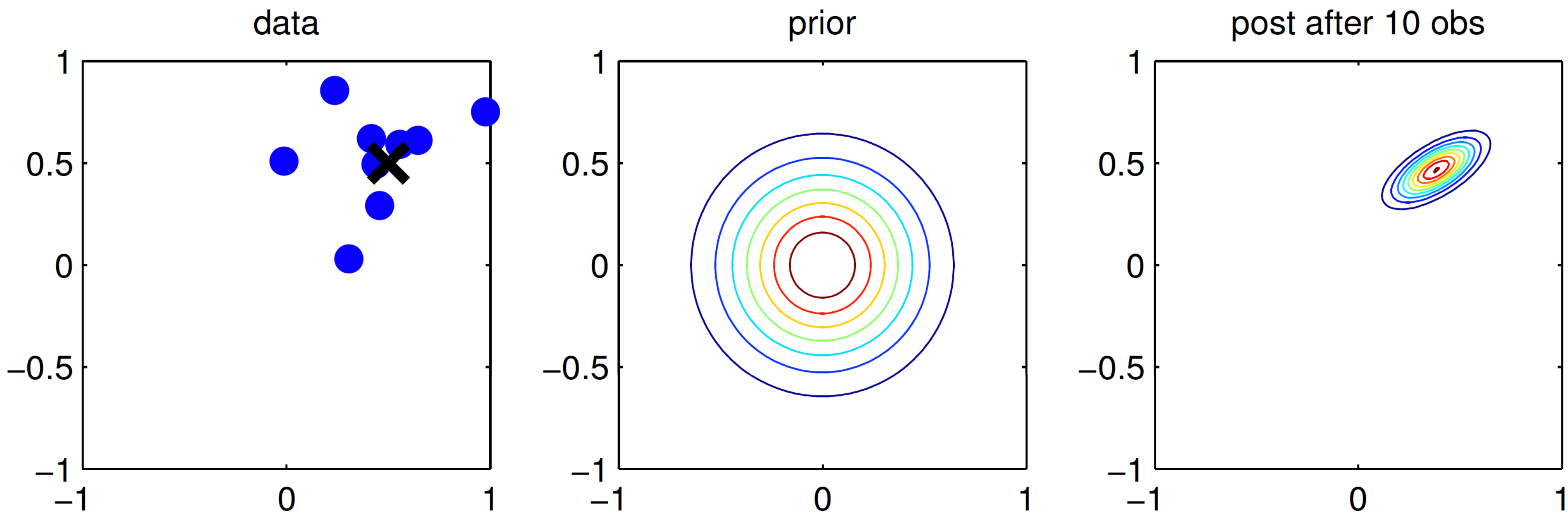
Impact of Prior Covariance

$$p(x) = \mathcal{N}(x|\mu_0, \lambda_0^{-1}) \qquad p(y_i|x) = \mathcal{N}(y_i|x, \lambda_y^{-1})$$



$$
\begin{aligned}
p(x|\mathbf{y}) &= \mathcal{N}(x|\mu_N, \lambda_N^{-1}) \\
\lambda_N &= \lambda_0 + N\lambda_y \\
\mu_N &= \frac{N\lambda_y \overline{y} + \lambda_0 \mu_0}{\lambda_N} = \frac{N\lambda_y}{N\lambda_y + \lambda_0}\overline{y} + \frac{\lambda_0}{N\lambda_y + \lambda_0}\mu_0
\end{aligned}
$$

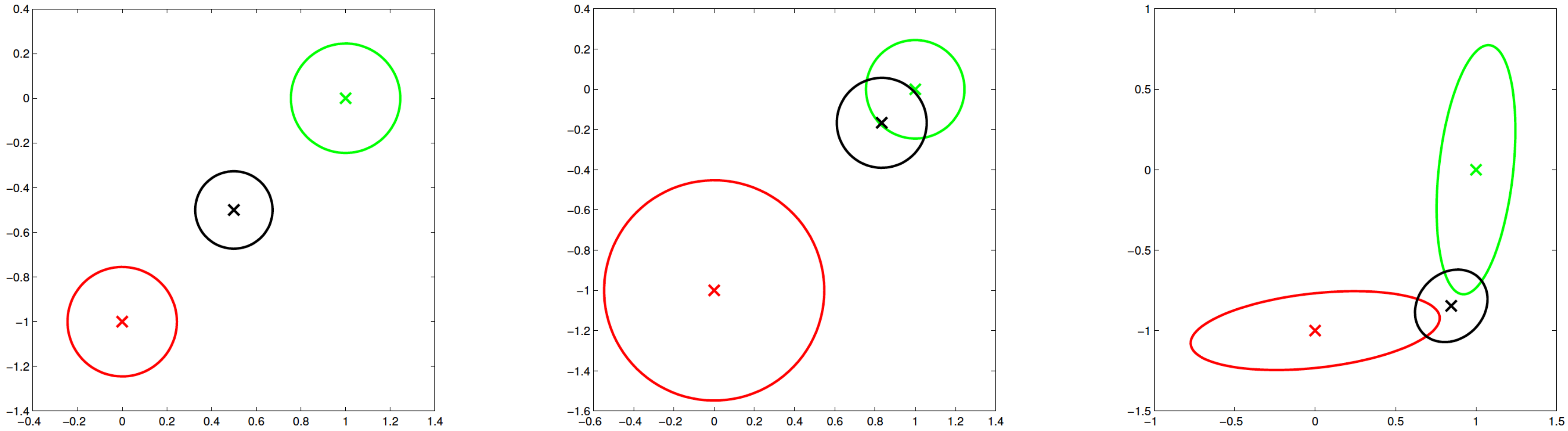
Vectors of Unknown Parameters



$\mathbf{y}_i \sim \mathcal{N}(\mathbf{x}, \boldsymbol{\Sigma}_y)$ $\mathcal{N}(\mathbf{x}|\mathbf{0}, 0.1\mathbf{I}_2)$

$$
\begin{aligned}
p(\mathbf{x}|\mathbf{y}_1, \ldots, \mathbf{y}_N) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N) \\
\boldsymbol{\Sigma}_N^{-1} &= \boldsymbol{\Sigma}_0^{-1} + N\boldsymbol{\Sigma}_y^{-1} \\
\boldsymbol{\mu}_N &= \boldsymbol{\Sigma}_N(\boldsymbol{\Sigma}_y^{-1}(N\overline{\mathbf{y}}) + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0)
\end{aligned}
$$

Observation Reliability

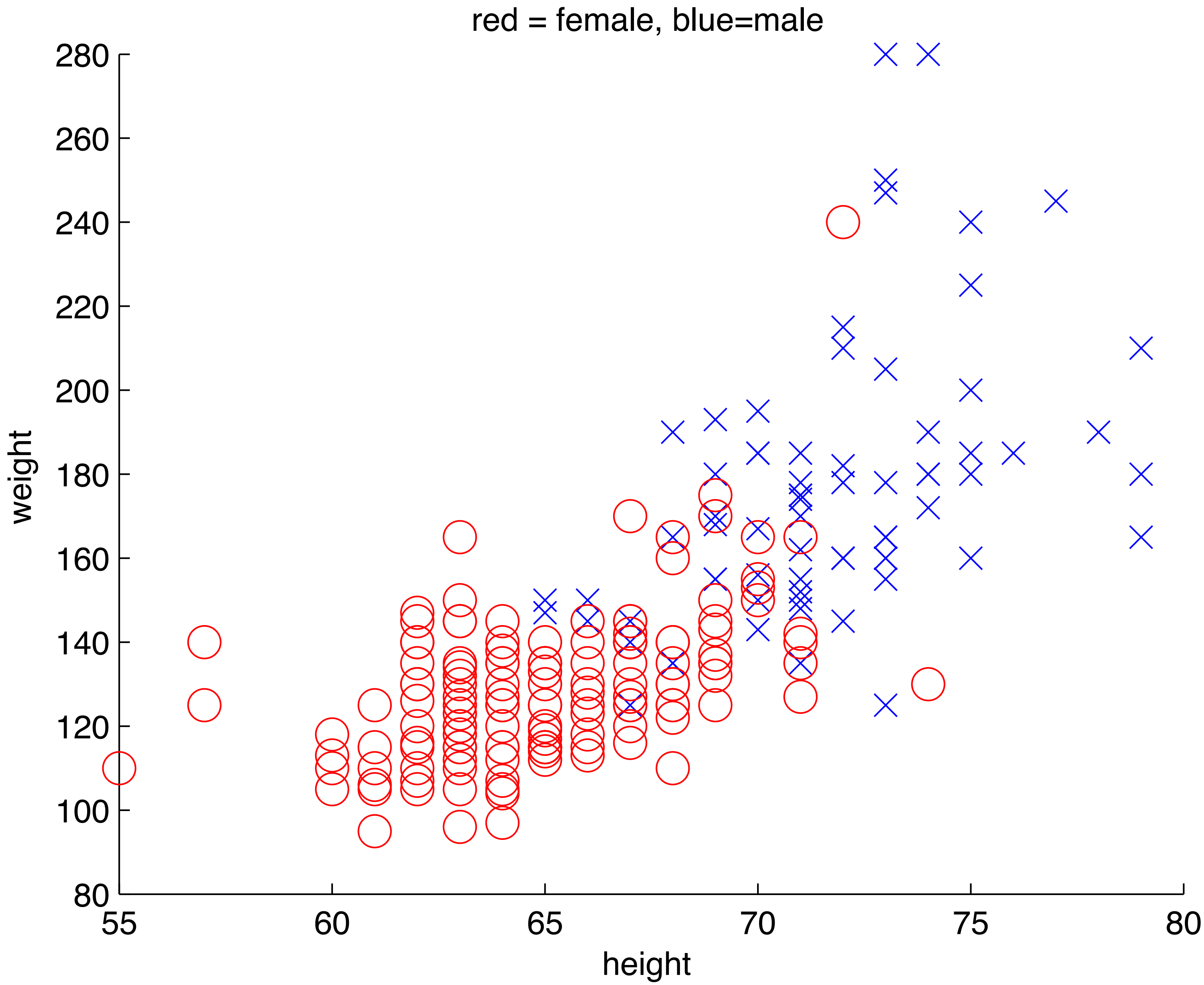


Red: Likelihood of observation A

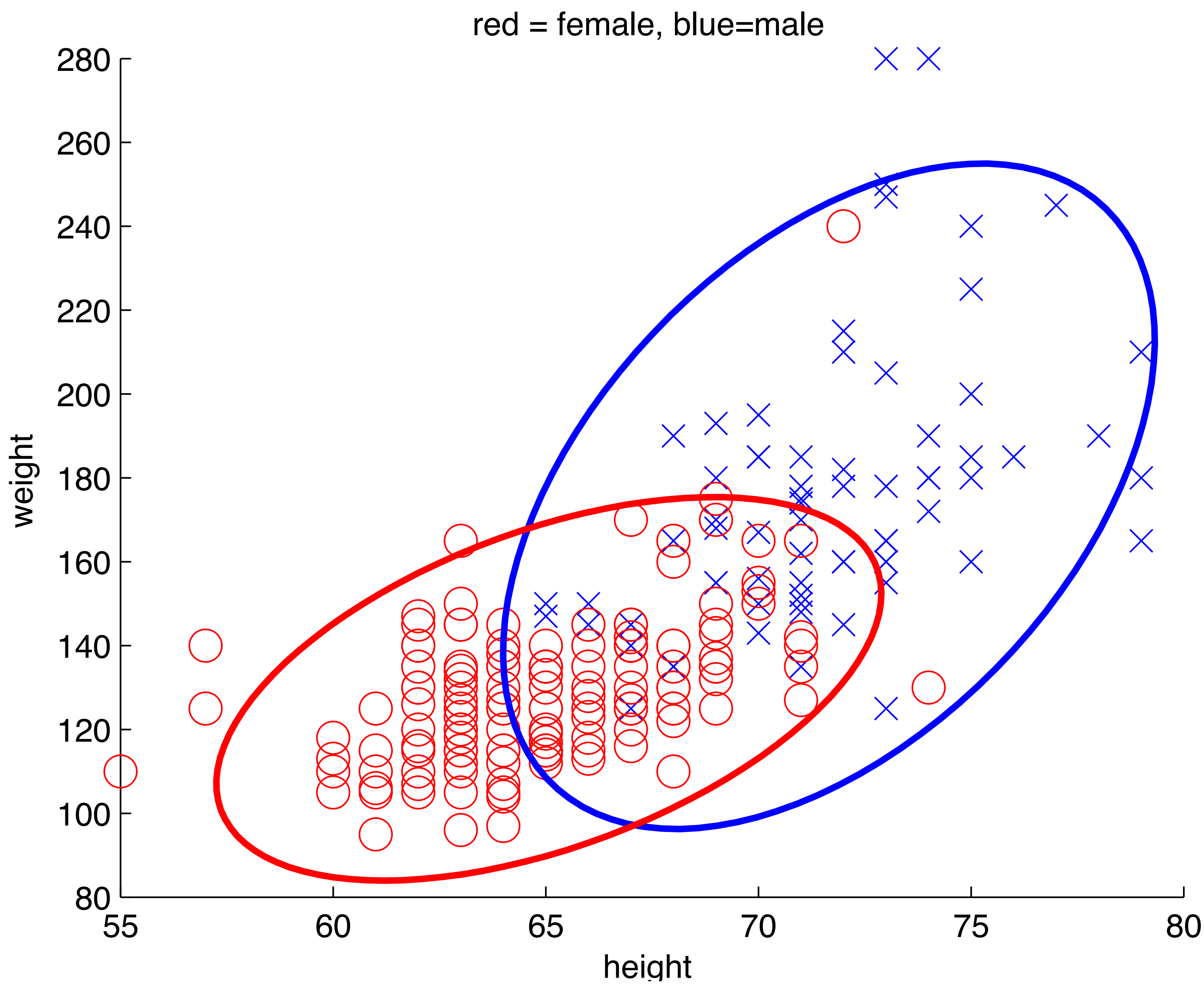
Green: Likelihood of observation B

Black: Gaussian posterior distribution

Class-Specific Gaussians



Class-Specific Gaussians



Discriminant Analysis