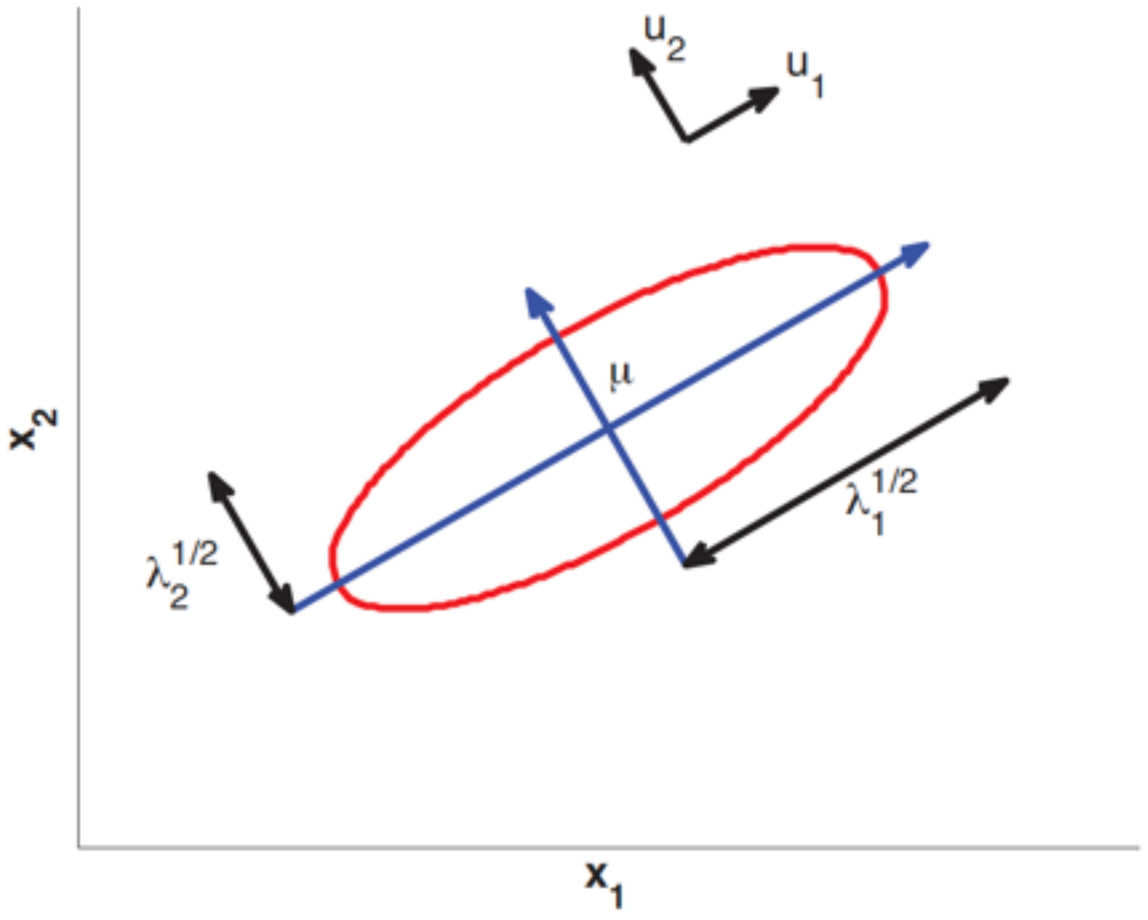
Gaussian models

Sep 2022

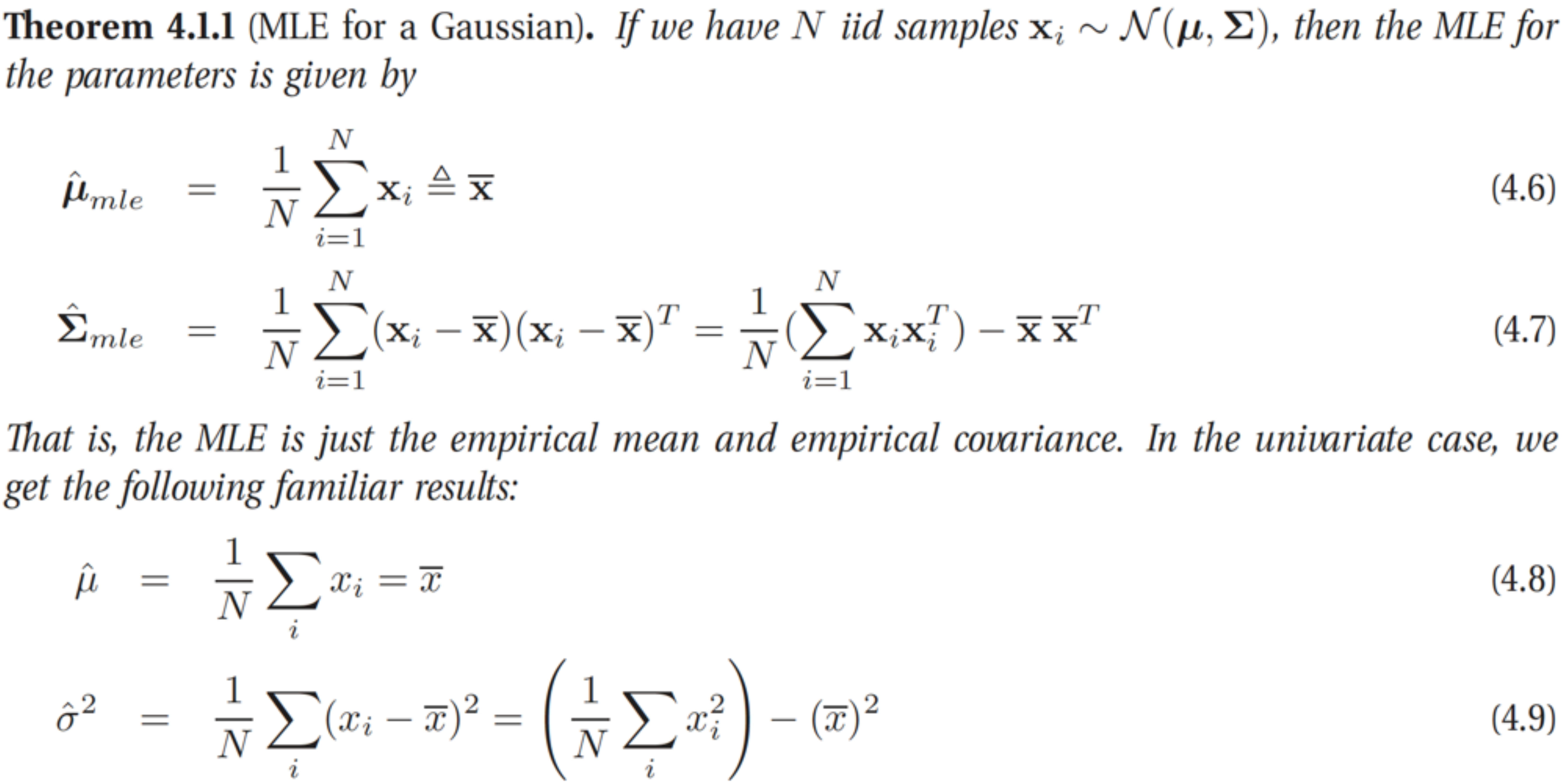
Murphy chap4

Multivariate normal distribution

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



Maximum likelihood estimate



Theorem 4.1.1 (MLE for a Gaussian)**.** *If we have N iid samples* $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*, then the MLE for the parameters is given by*

$$\hat{\boldsymbol{\mu}}_{mle} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i \triangleq \overline{\mathbf{x}} \tag{4.6}$$

$$\hat{\boldsymbol{\Sigma}}_{mle} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T = \frac{1}{N}(\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^T) - \overline{\mathbf{x}}\,\overline{\mathbf{x}}^T \tag{4.7}$$

That is, the MLE is just the empirical mean and empirical covariance. In the univariate case, we get the following familiar results:

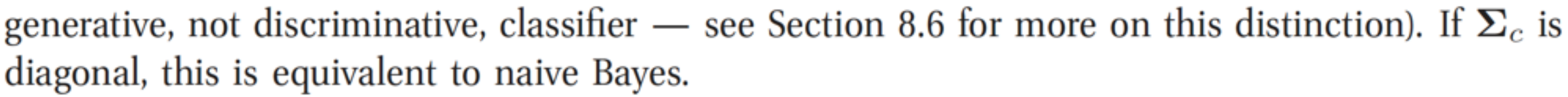
$$\hat{\mu} = \frac{1}{N}\sum_i x_i = \overline{x} \tag{4.8}$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_i (x_i - \overline{x})^2 = \left(\frac{1}{N}\sum_i x_i^2\right) - (\overline{x})^2 \tag{4.9}$$

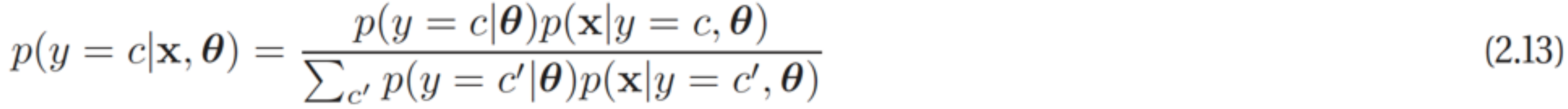
Gaussian discriminant analysis

class conditional density:

$$p(\mathbf{x}|y=c,\boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) \tag{4.30}$$

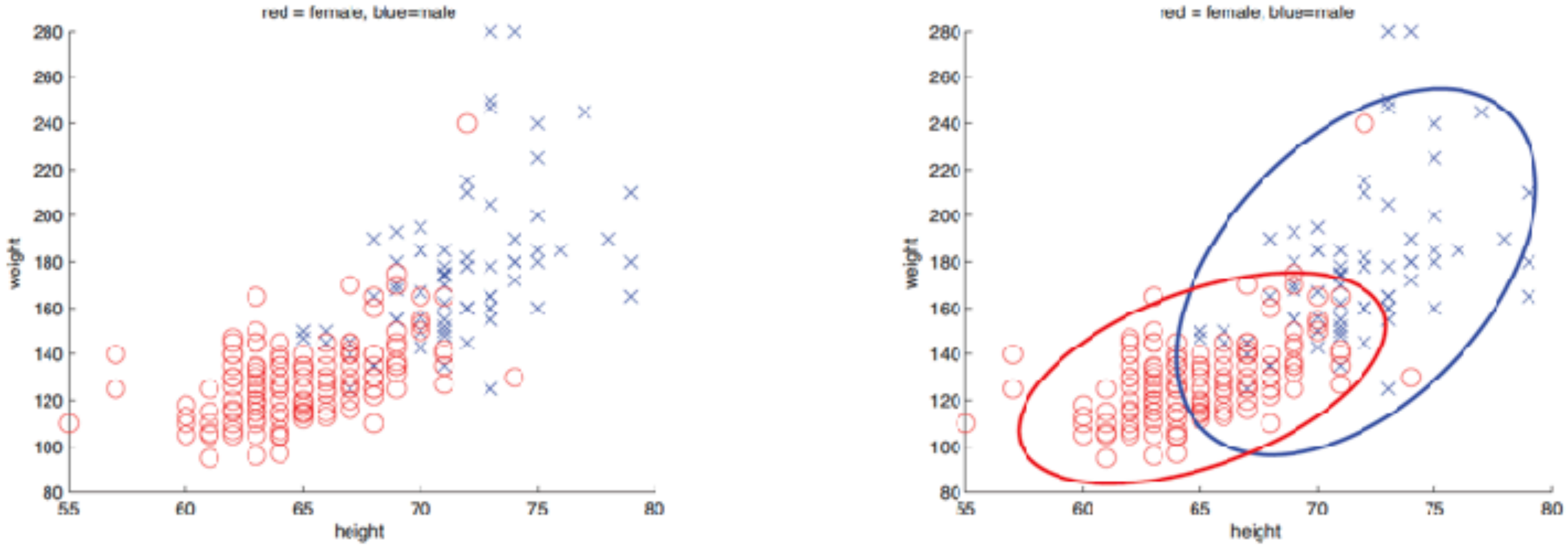


generative, not discriminative, classifier — see Section 8.6 for more on this distinction). If $\boldsymbol{\Sigma}_c$ is diagonal, this is equivalent to naive Bayes.



$$p(y=c|\mathbf{x},\boldsymbol{\theta}) = \frac{p(y=c|\boldsymbol{\theta})p(\mathbf{x}|y=c,\boldsymbol{\theta})}{\sum_{c'} p(y=c'|\boldsymbol{\theta})p(\mathbf{x}|y=c',\boldsymbol{\theta})} \tag{2.13}$$

Gaussian discriminant analysis



(a)

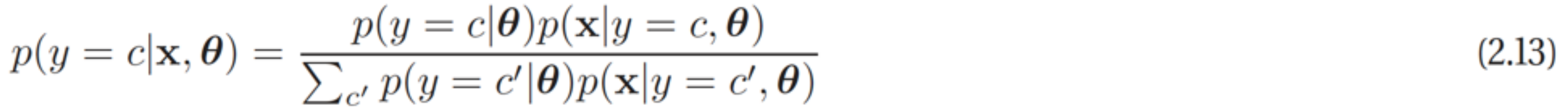
(b)

Figure 4.2 (a) Height/weight data. (b) Visualization of 2d Gaussians fit to each class. 95% of the probability mass is inside the ellipse. Figure generated by `gaussHeightWeight`.

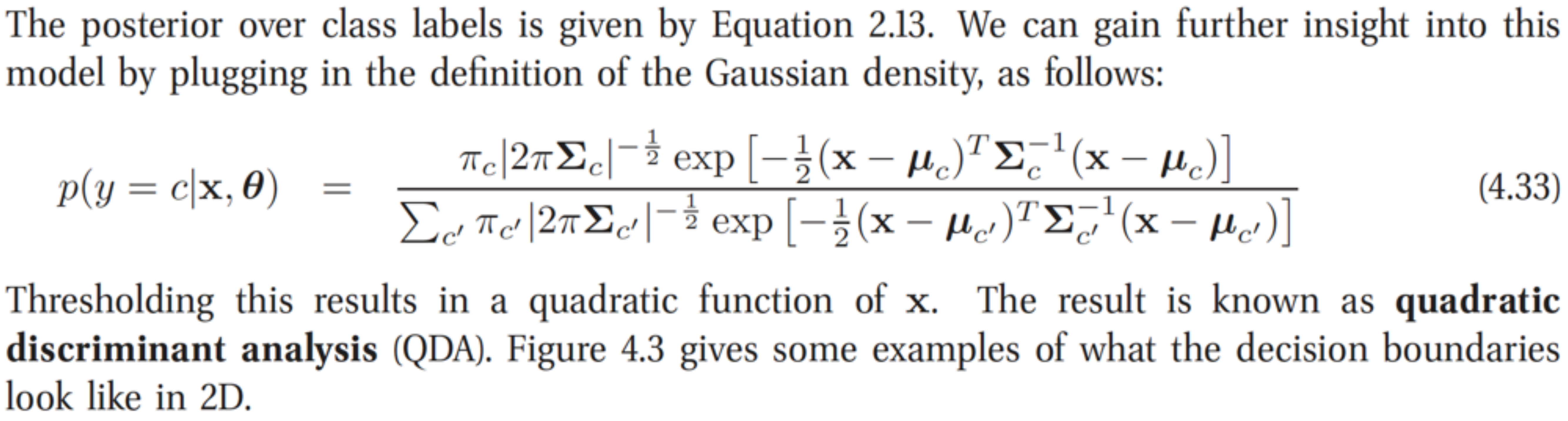
As an example, Figure 4.2 shows two Gaussian class-conditional densities in 2d, representing the height and weight of men and women. We can see that the features are correlated, as is to be expected (tall people tend to weigh more). The ellipses for each class contain 95% of the probability mass. If we have a uniform prior over classes, we can classify a new test vector as follows:

$$\hat{y}(\mathbf{x}) = \underset{c}{\operatorname{argmin}} (\mathbf{x} - \boldsymbol{\mu}_c)^T \boldsymbol{\Sigma}_c^{-1} (\mathbf{x} - \boldsymbol{\mu}_c) \tag{4.32}$$

Quadratic discriminant analysis (QDA)



$$p(y = c|\mathbf{x}, \boldsymbol{\theta}) = \frac{p(y = c|\boldsymbol{\theta})p(\mathbf{x}|y = c, \boldsymbol{\theta})}{\sum_{c'} p(y = c'|\boldsymbol{\theta})p(\mathbf{x}|y = c', \boldsymbol{\theta})} \tag{2.13}$$



The posterior over class labels is given by Equation 2.13. We can gain further insight into this model by plugging in the definition of the Gaussian density, as follows:

$$p(y = c|\mathbf{x}, \boldsymbol{\theta}) \quad = \quad \frac{\pi_c|2\pi\boldsymbol{\Sigma}_c|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_c)^T\boldsymbol{\Sigma}_c^{-1}(\mathbf{x} - \boldsymbol{\mu}_c)\right]}{\sum_{c'} \pi_{c'}|2\pi\boldsymbol{\Sigma}_{c'}|^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_{c'})^T\boldsymbol{\Sigma}_{c'}^{-1}(\mathbf{x} - \boldsymbol{\mu}_{c'})\right]} \tag{4.33}$$

Thresholding this results in a quadratic function of $\mathbf{x}$. The result is known as **quadratic discriminant analysis** (QDA). Figure 4.3 gives some examples of what the decision boundaries look like in 2D.

Quadratic discriminant analysis (QDA)

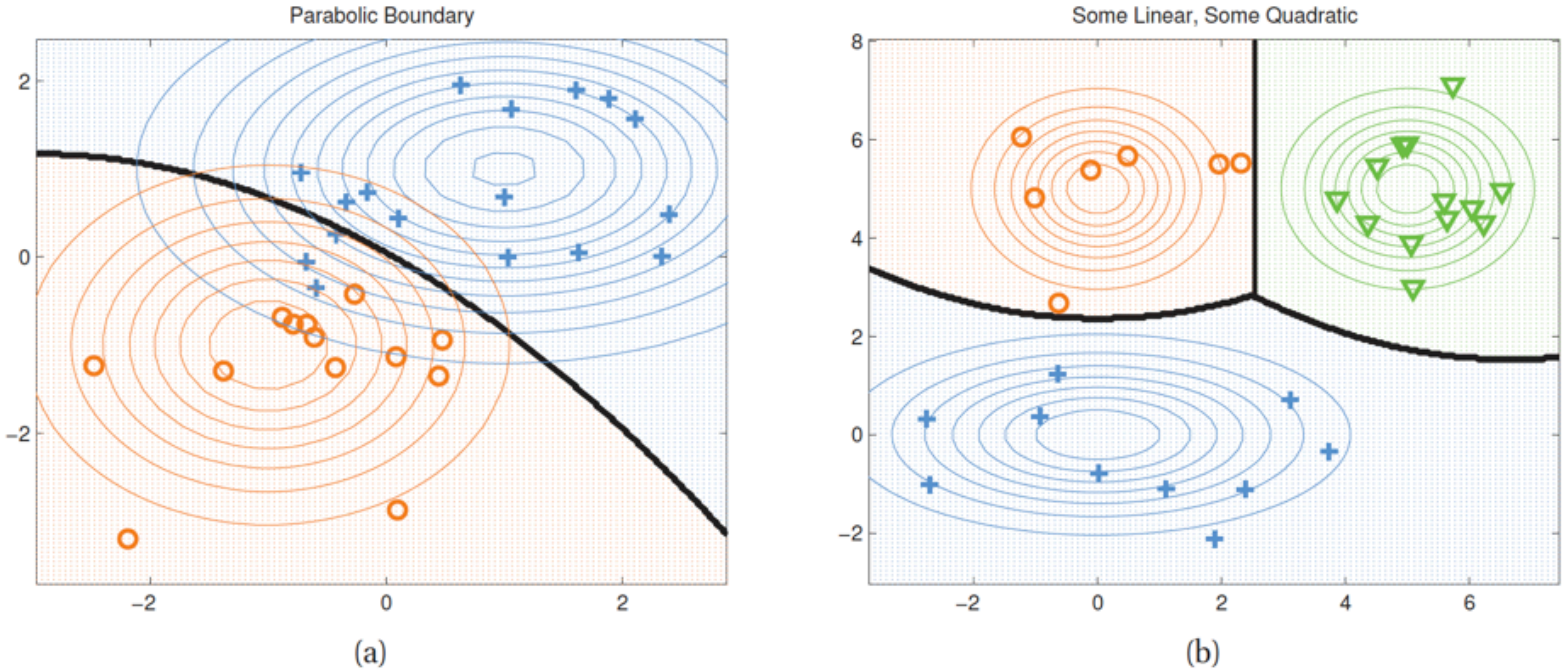


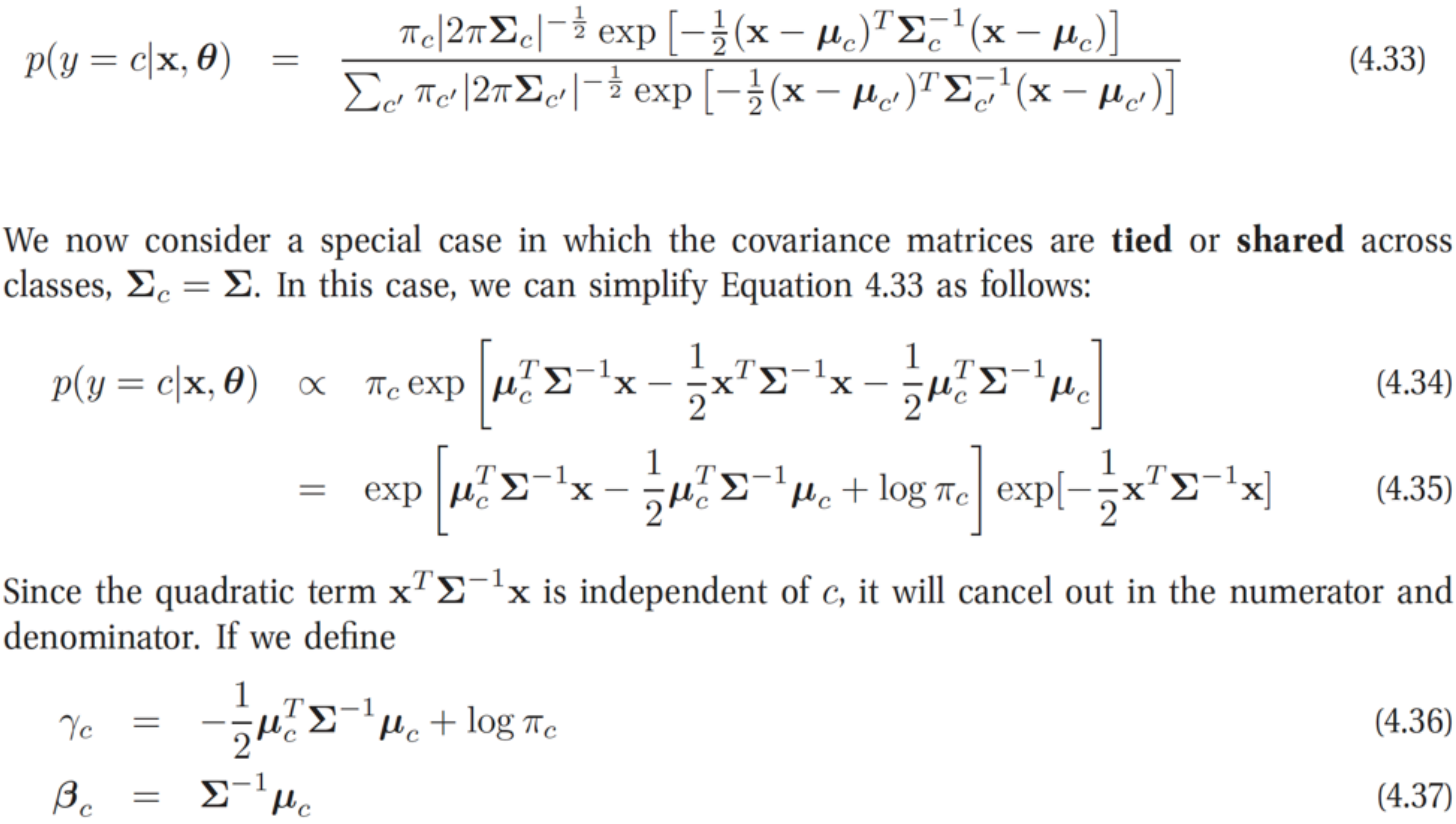
Figure 4.3 Quadratic decision boundaries in 2D for the 2 and 3 class case. Figure generated by `discrimAnalysisDboundariesDemo`.

Quadratic discriminant analysis (QDA)

Linear or parabolic decision boundaries

$$
\begin{aligned}
\log \frac{p(c=k|x)}{p(c=l|x)} &= \log \frac{\pi_k (2\pi)^{-d/2} |\Sigma|^{-1/2} \exp\Big(-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1} (x-\mu_k)\Big)}{\pi_l (2\pi)^{-d/2} |\Sigma|^{-1/2} \exp\Big(-\frac{1}{2}(x-\mu_l)^T \Sigma^{-1} (x-\mu_l)\Big)} \\
&= \log \frac{\pi_k}{\pi_l} - \frac{1}{2}(x-\mu_k)^T \Sigma^{-1} (x-\mu_k) + \frac{1}{2}(x-\mu_l)^T \Sigma^{-1} (x-\mu_l) \\
&= \log \frac{\pi_k}{\pi_l} - x^T \Sigma^{-1} (\mu_k - \mu_l) - \frac{1}{2}(\mu_k + \mu_l)^T \Sigma^{-1} (\mu_k - \mu_l),
\end{aligned}
$$

Linear discriminant analysis (LDA)



$$p(y=c|\mathbf{x},\boldsymbol{\theta}) = \frac{\pi_c|2\pi\boldsymbol{\Sigma}_c|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_c)^T\boldsymbol{\Sigma}_c^{-1}(\mathbf{x}-\boldsymbol{\mu}_c)\right]}{\sum_{c'}\pi_{c'}|2\pi\boldsymbol{\Sigma}_{c'}|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_{c'})^T\boldsymbol{\Sigma}_{c'}^{-1}(\mathbf{x}-\boldsymbol{\mu}_{c'})\right]} \tag{4.33}$$

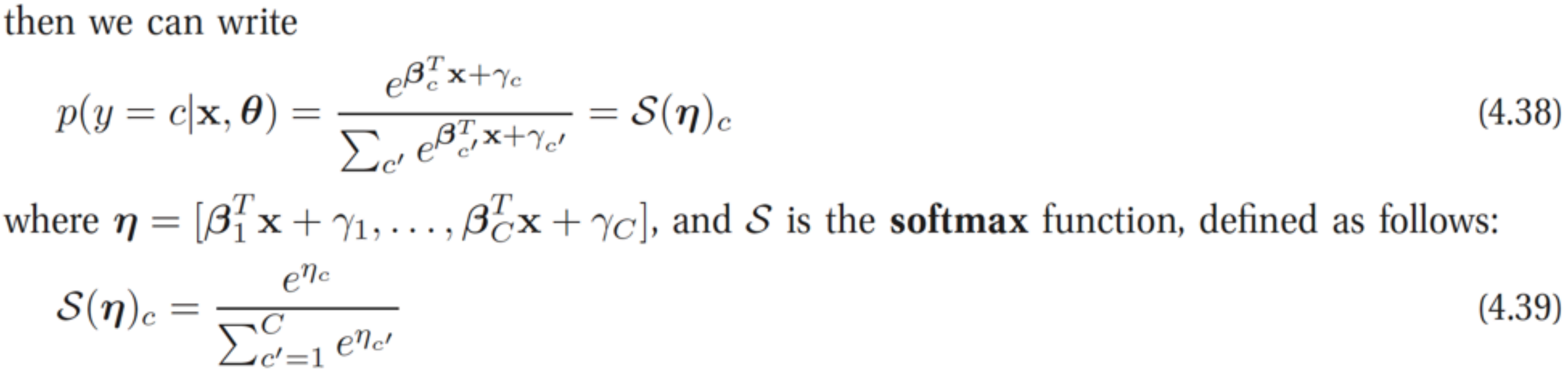
We now consider a special case in which the covariance matrices are **tied** or **shared** across classes, $\boldsymbol{\Sigma}_c = \boldsymbol{\Sigma}$. In this case, we can simplify Equation 4.33 as follows:

$$\begin{aligned} p(y=c|\mathbf{x},\boldsymbol{\theta}) &\propto \pi_c \exp\left[\boldsymbol{\mu}_c^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_c^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_c\right] && (4.34) \\ &= \exp\left[\boldsymbol{\mu}_c^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_c^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_c + \log\pi_c\right]\exp[-\frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}] && (4.35) \end{aligned}$$

Since the quadratic term $\mathbf{x}^T\boldsymbol{\Sigma}^{-1}\mathbf{x}$ is independent of c, it will cancel out in the numerator and denominator. If we define

$$\begin{aligned} \gamma_c &= -\frac{1}{2}\boldsymbol{\mu}_c^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_c + \log\pi_c && (4.36) \\ \boldsymbol{\beta}_c &= \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_c && (4.37) \end{aligned}$$

Linear discriminant analysis (LDA)



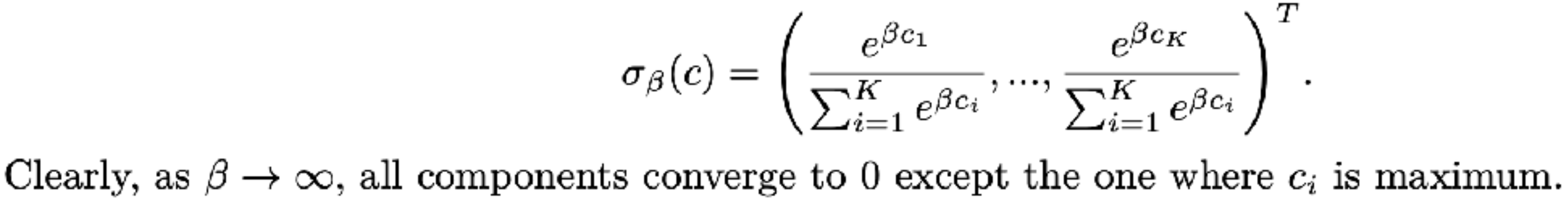
then we can write

$$p(y=c|\mathbf{x},\boldsymbol{\theta}) = \frac{e^{\boldsymbol{\beta}_c^T\mathbf{x}+\gamma_c}}{\sum_{c'} e^{\boldsymbol{\beta}_{c'}^T\mathbf{x}+\gamma_{c'}}} = \mathcal{S}(\boldsymbol{\eta})_c \tag{4.38}$$

where $\boldsymbol{\eta} = [\boldsymbol{\beta}_1^T\mathbf{x}+\gamma_1,\ldots,\boldsymbol{\beta}_C^T\mathbf{x}+\gamma_C]$, and $\mathcal{S}$ is the **softmax** function, defined as follows:

$$\mathcal{S}(\boldsymbol{\eta})_c = \frac{e^{\eta_c}}{\sum_{c'=1}^{C} e^{\eta_{c'}}} \tag{4.39}$$

Smoothed argmax function

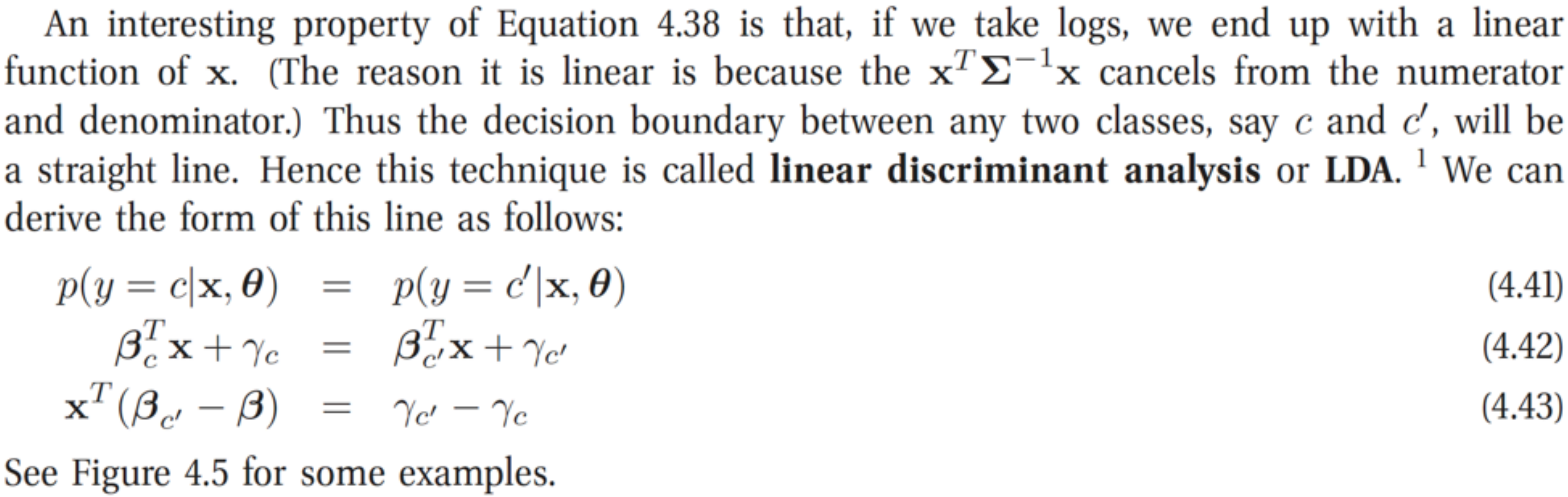


$$\sigma_\beta(c) = \left(\frac{e^{\beta c_1}}{\sum_{i=1}^{K} e^{\beta c_i}},\ldots,\frac{e^{\beta c_K}}{\sum_{i=1}^{K} e^{\beta c_i}}\right)^T.$$

Clearly, as $\beta\to\infty$, all components converge to 0 except the one where c_i is maximum.

Linear discriminant analysis (LDA)

$$p(y = c|\mathbf{x}, \boldsymbol{\theta}) = \frac{e^{\boldsymbol{\beta}_c^T \mathbf{x} + \gamma_c}}{\sum_{c'} e^{\boldsymbol{\beta}_{c'}^T \mathbf{x} + \gamma_{c'}}} = \mathcal{S}(\boldsymbol{\eta})_c \tag{4.38}$$



An interesting property of Equation 4.38 is that, if we take logs, we end up with a linear function of $\mathbf{x}$. (The reason it is linear is because the $\mathbf{x}^T \boldsymbol{\Sigma}^{-1} \mathbf{x}$ cancels from the numerator and denominator.) Thus the decision boundary between any two classes, say c and c', will be a straight line. Hence this technique is called **linear discriminant analysis** or **LDA**. [1] We can derive the form of this line as follows:

$$\begin{aligned} p(y = c|\mathbf{x}, \boldsymbol{\theta}) &= p(y = c'|\mathbf{x}, \boldsymbol{\theta}) && (4.41) \\ \boldsymbol{\beta}_c^T \mathbf{x} + \gamma_c &= \boldsymbol{\beta}_{c'}^T \mathbf{x} + \gamma_{c'} && (4.42) \\ \mathbf{x}^T (\boldsymbol{\beta}_{c'} - \boldsymbol{\beta}) &= \gamma_{c'} - \gamma_c && (4.43) \end{aligned}$$

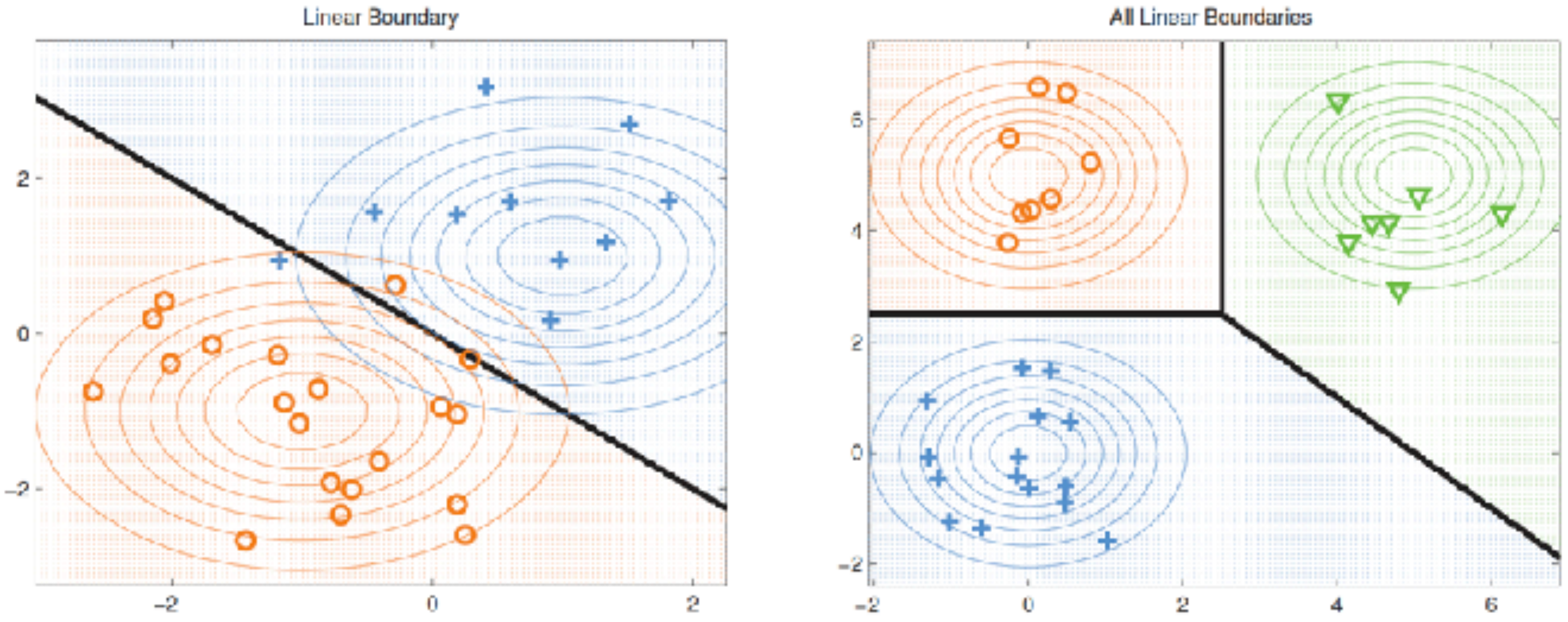
See Figure 4.5 for some examples.

Two-class LDA

To gain further insight into the meaning of these equations, let us consider the binary case. In this case, the posterior is given by

$$p(y=1|\mathbf{x},\boldsymbol{\theta}) = \frac{e^{\boldsymbol{\beta}_1^T\mathbf{x}+\gamma_1}}{e^{\boldsymbol{\beta}_1^T\mathbf{x}+\gamma_1}+e^{\boldsymbol{\beta}_0^T\mathbf{x}+\gamma_0}} \tag{4.44}$$

$$= \frac{1}{1+e^{(\boldsymbol{\beta}_0-\boldsymbol{\beta}_1)^T\mathbf{x}+(\gamma_0-\gamma_1)}} = \text{sigm}\left((\boldsymbol{\beta}_1-\boldsymbol{\beta}_0)^T\mathbf{x}+(\gamma_1-\gamma_0)\right) \tag{4.45}$$

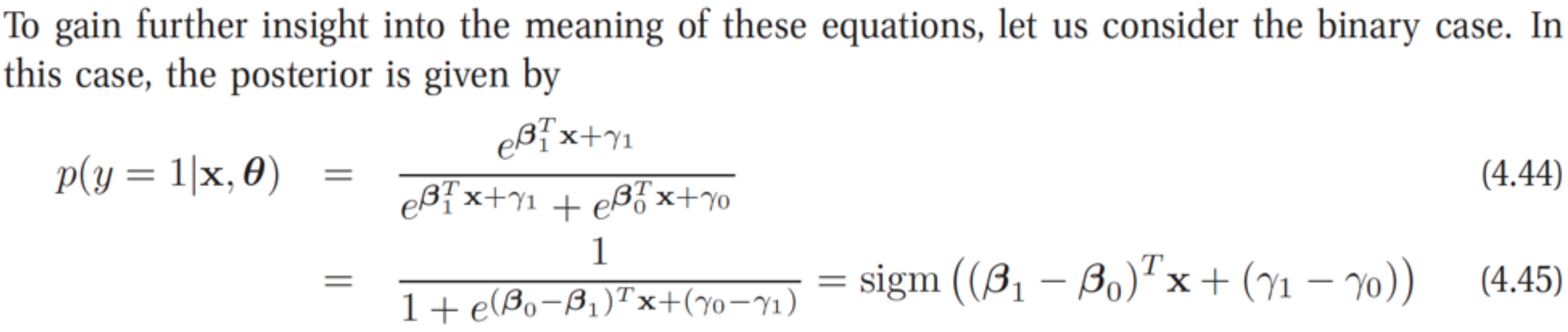


(a)

(b)

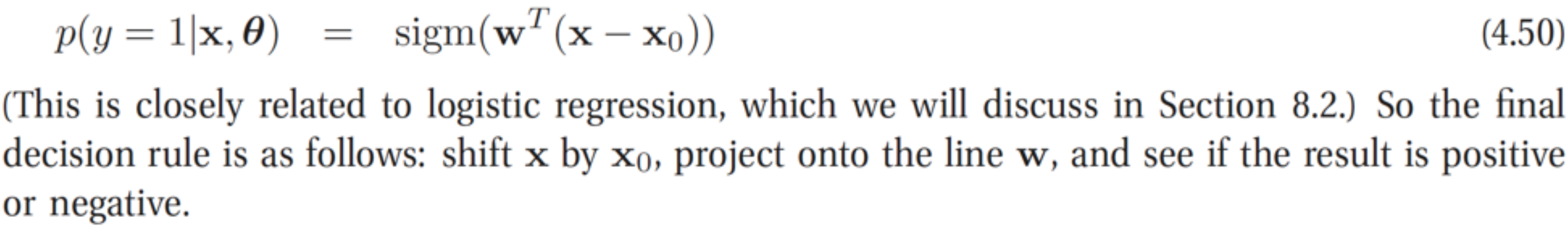
Figure 4.5 Linear decision boundaries in 2D for the 2 and 3 class case. Figure generated by `discrimAnalysisDboundariesDemo`.

Two-class LDA



To gain further insight into the meaning of these equations, let us consider the binary case. In this case, the posterior is given by

$$\begin{aligned} p(y=1|\mathbf{x},\boldsymbol{\theta}) &= \frac{e^{\boldsymbol{\beta}_1^T\mathbf{x}+\gamma_1}}{e^{\boldsymbol{\beta}_1^T\mathbf{x}+\gamma_1}+e^{\boldsymbol{\beta}_0^T\mathbf{x}+\gamma_0}} && (4.44)\\ &= \frac{1}{1+e^{(\boldsymbol{\beta}_0-\boldsymbol{\beta}_1)^T\mathbf{x}+(\gamma_0-\gamma_1)}} = \mathrm{sigm}\left((\boldsymbol{\beta}_1-\boldsymbol{\beta}_0)^T\mathbf{x}+(\gamma_1-\gamma_0)\right) && (4.45) \end{aligned}$$



$$p(y=1|\mathbf{x},\boldsymbol{\theta}) = \mathrm{sigm}(\mathbf{w}^T(\mathbf{x}-\mathbf{x}_0)) \qquad (4.50)$$

(This is closely related to logistic regression, which we will discuss in Section 8.2.) So the final decision rule is as follows: shift $\mathbf{x}$ by $\mathbf{x}_0$, project onto the line $\mathbf{w}$, and see if the result is positive or negative.

Two-class LDA

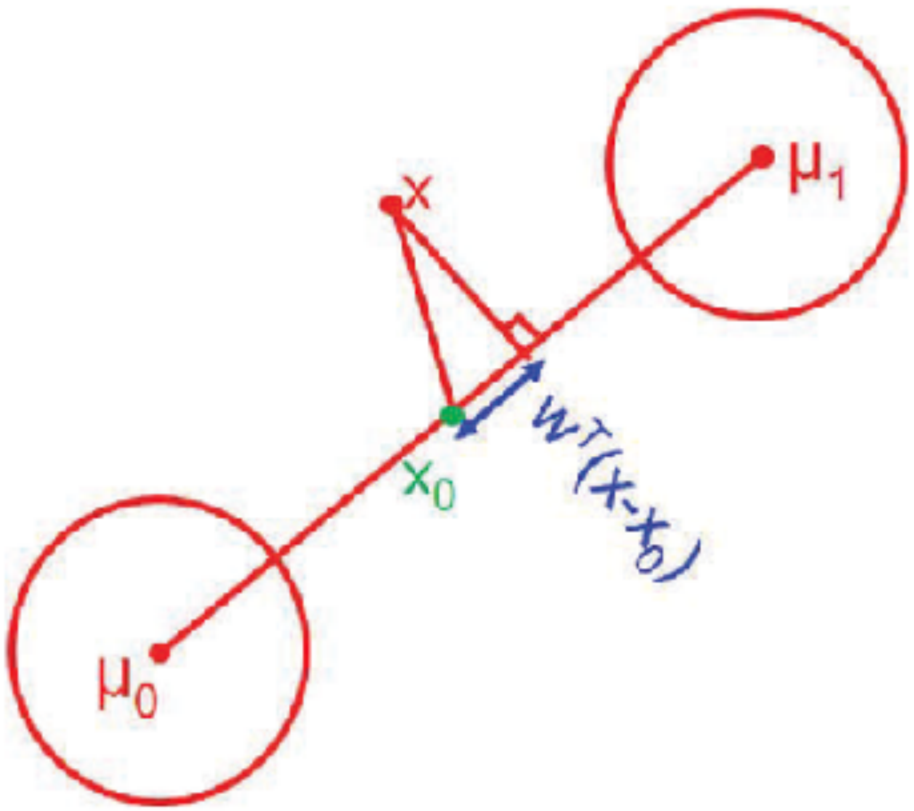
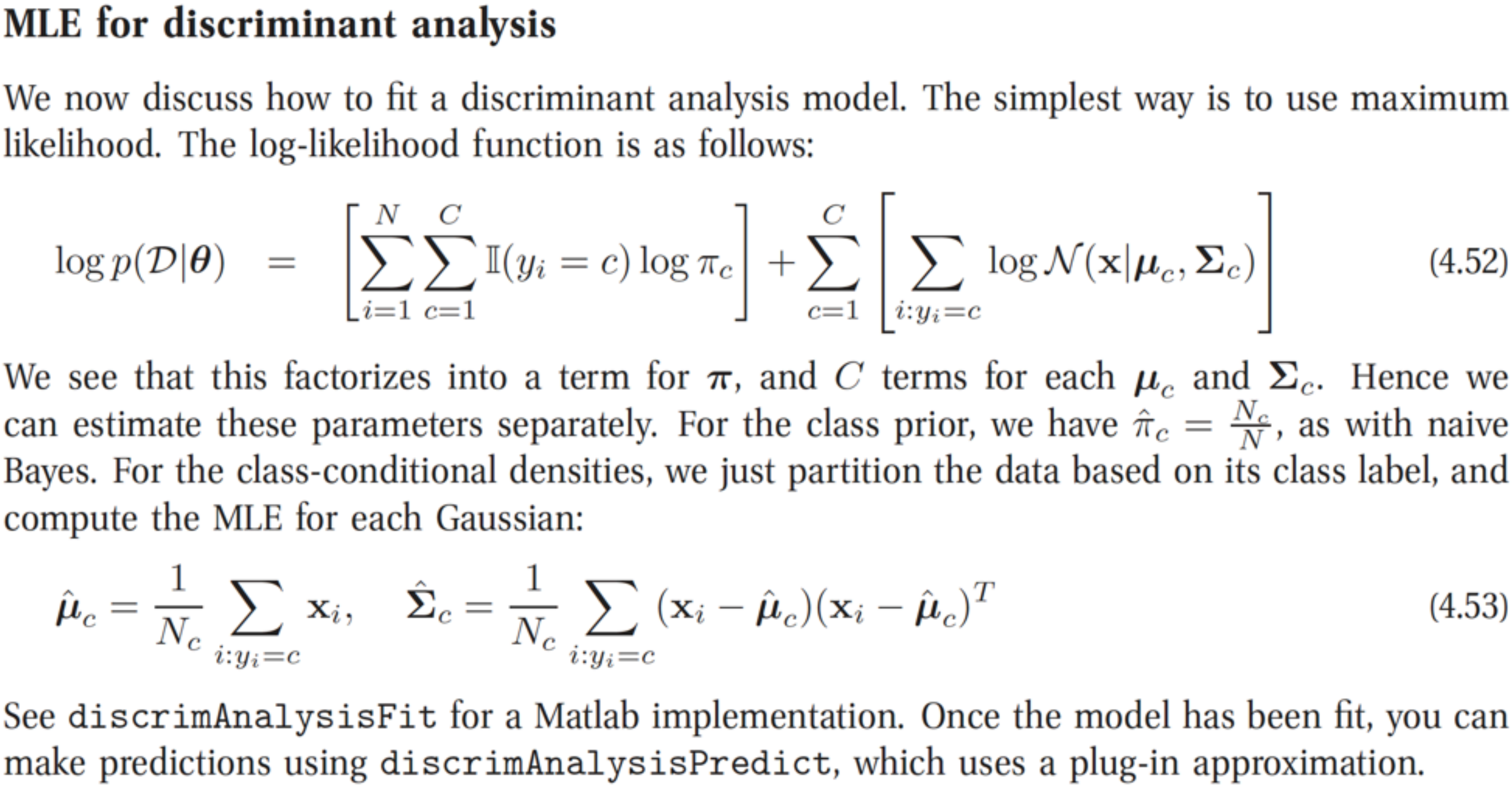


Figure 4.6 Geometry of LDA in the 2 class case where $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{I}$.

$$p(y = 1|\mathbf{x}, \boldsymbol{\theta}) = \text{sigm}(\mathbf{w}^T(\mathbf{x} - \mathbf{x}_0)) \tag{4.50}$$

(This is closely related to logistic regression, which we will discuss in Section 8.2.) So the final decision rule is as follows: shift $\mathbf{x}$ by $\mathbf{x}_0$, project onto the line $\mathbf{w}$, and see if the result is positive or negative.

Fit discriminant models



MLE for discriminant analysis

We now discuss how to fit a discriminant analysis model. The simplest way is to use maximum likelihood. The log-likelihood function is as follows:

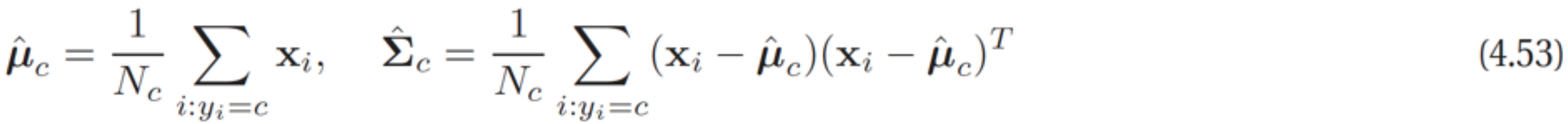
$$\log p(\mathcal{D}|\boldsymbol{\theta}) \quad = \quad \left[\sum_{i=1}^{N}\sum_{c=1}^{C}\mathbb{I}(y_i = c)\log \pi_c\right] + \sum_{c=1}^{C}\left[\sum_{i:y_i=c}\log \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)\right] \tag{4.52}$$

We see that this factorizes into a term for $\boldsymbol{\pi}$, and C terms for each $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$. Hence we can estimate these parameters separately. For the class prior, we have $\hat{\pi}_c = \frac{N_c}{N}$, as with naive Bayes. For the class-conditional densities, we just partition the data based on its class label, and compute the MLE for each Gaussian:

$$\hat{\boldsymbol{\mu}}_c = \frac{1}{N_c}\sum_{i:y_i=c}\mathbf{x}_i, \quad \hat{\boldsymbol{\Sigma}}_c = \frac{1}{N_c}\sum_{i:y_i=c}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_c)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_c)^T \tag{4.53}$$

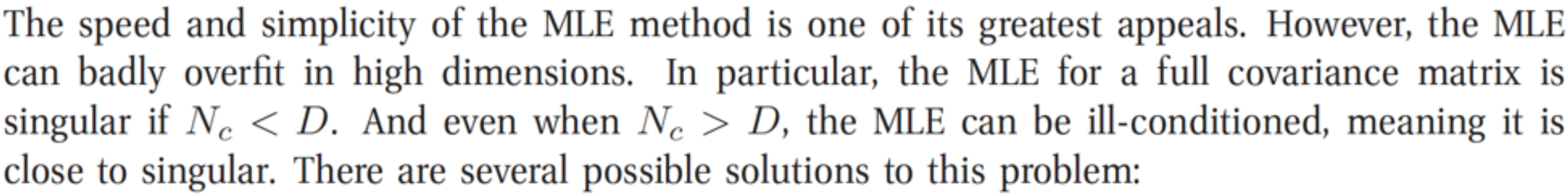
See `discrimAnalysisFit` for a Matlab implementation. Once the model has been fit, you can make predictions using `discrimAnalysisPredict`, which uses a plug-in approximation.

Prevent overfitting



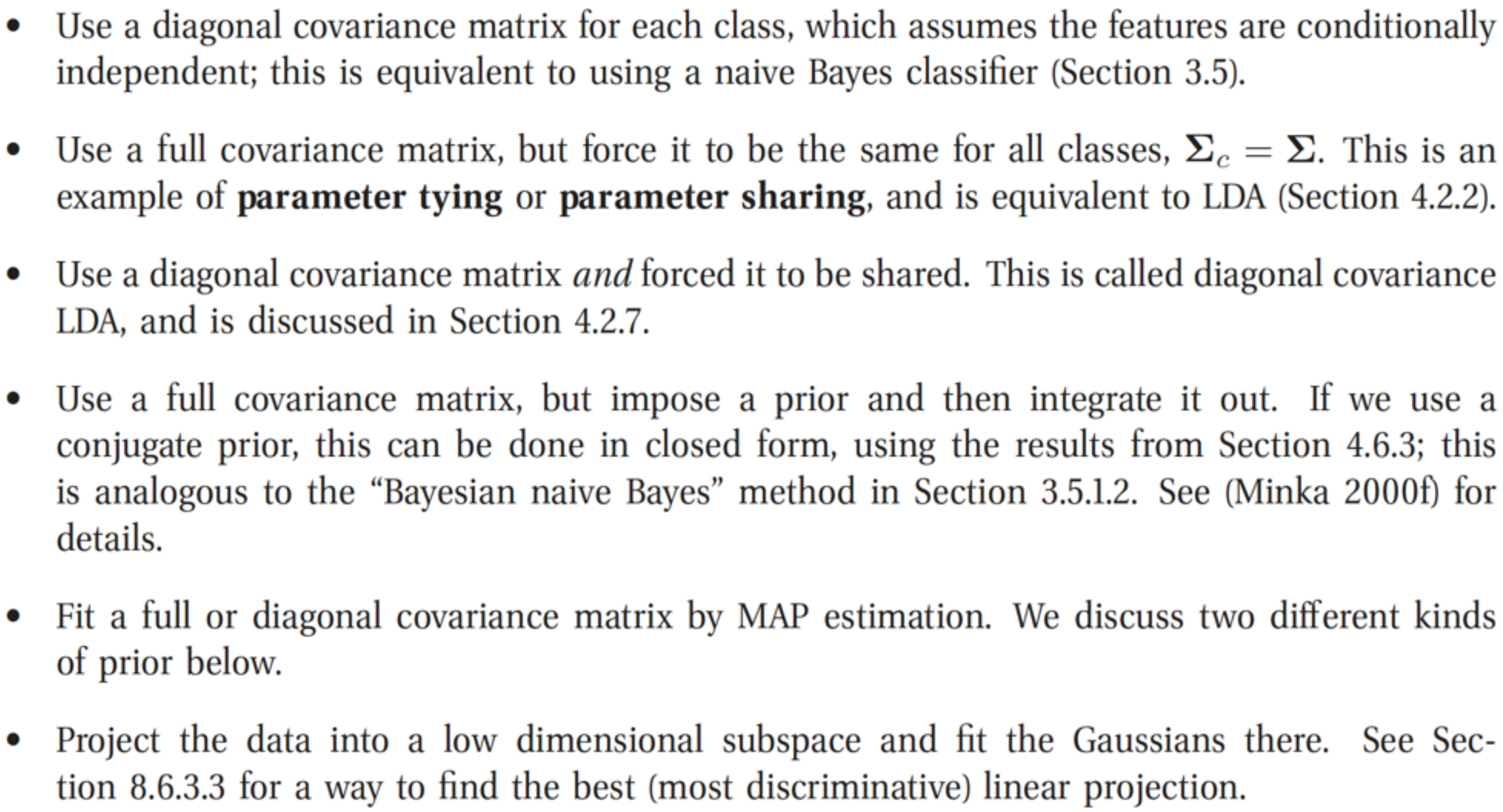
$$\hat{\boldsymbol{\mu}}_c = \frac{1}{N_c} \sum_{i:y_i=c} \mathbf{x}_i, \quad \hat{\boldsymbol{\Sigma}}_c = \frac{1}{N_c} \sum_{i:y_i=c} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_c)(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_c)^T \qquad (4.53)$$

$$rank(AB) \leqslant min(rank(A), rank(B))$$
$$rank(A + B) \leqslant rank(A) + rank(B)$$



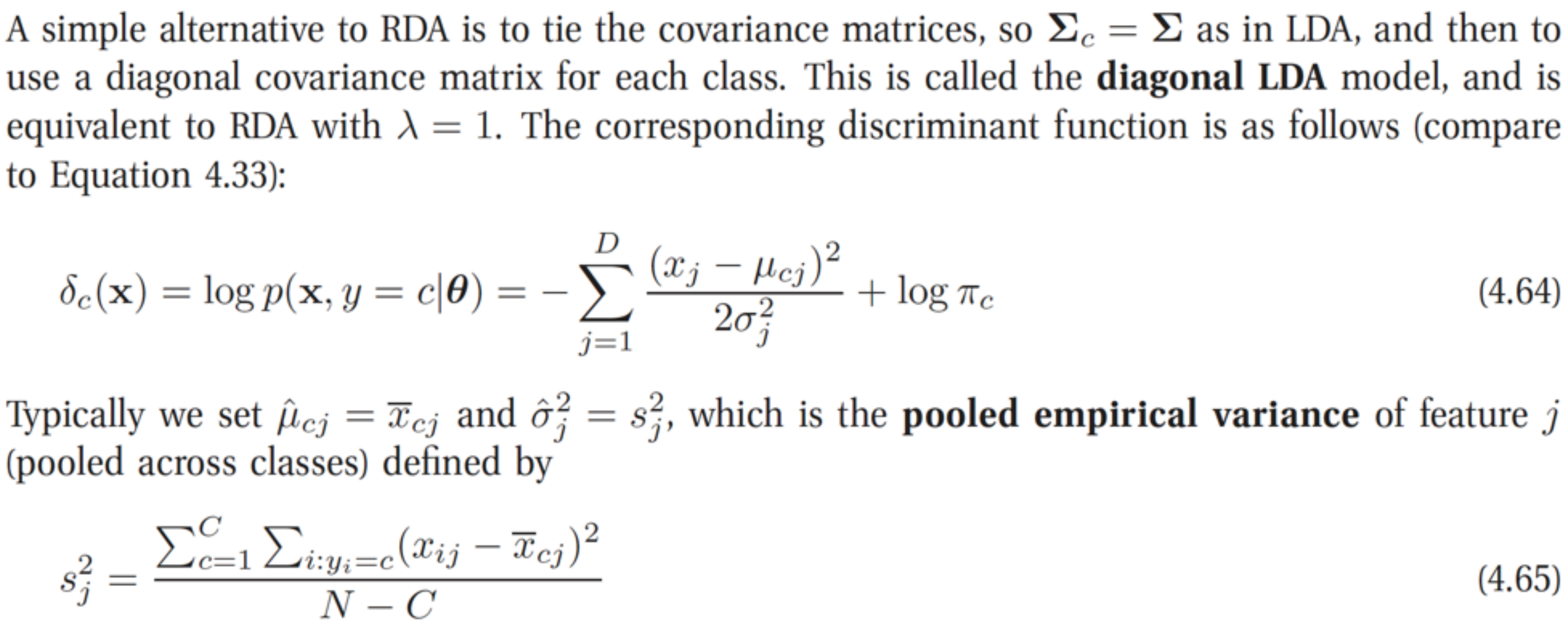
The speed and simplicity of the MLE method is one of its greatest appeals. However, the MLE can badly overfit in high dimensions. In particular, the MLE for a full covariance matrix is singular if $N_c < D$. And even when $N_c > D$, the MLE can be ill-conditioned, meaning it is close to singular. There are several possible solutions to this problem:

Prevent overfitting



- Use a diagonal covariance matrix for each class, which assumes the features are conditionally independent; this is equivalent to using a naive Bayes classifier (Section 3.5).
- Use a full covariance matrix, but force it to be the same for all classes, $\boldsymbol{\Sigma}_c = \boldsymbol{\Sigma}$. This is an example of **parameter tying** or **parameter sharing**, and is equivalent to LDA (Section 4.2.2).
- Use a diagonal covariance matrix *and* forced it to be shared. This is called diagonal covariance LDA, and is discussed in Section 4.2.7.
- Use a full covariance matrix, but impose a prior and then integrate it out. If we use a conjugate prior, this can be done in closed form, using the results from Section 4.6.3; this is analogous to the "Bayesian naive Bayes" method in Section 3.5.1.2. See (Minka 2000f) for details.
- Fit a full or diagonal covariance matrix by MAP estimation. We discuss two different kinds of prior below.
- Project the data into a low dimensional subspace and fit the Gaussians there. See Section 8.6.3.3 for a way to find the best (most discriminative) linear projection.

Diagonal LDA



A simple alternative to RDA is to tie the covariance matrices, so $\mathbf{\Sigma}_c = \mathbf{\Sigma}$ as in LDA, and then to use a diagonal covariance matrix for each class. This is called the **diagonal LDA** model, and is equivalent to RDA with $\lambda = 1$. The corresponding discriminant function is as follows (compare to Equation 4.33):

$$\delta_c(\mathbf{x}) = \log p(\mathbf{x}, y = c|\boldsymbol{\theta}) = -\sum_{j=1}^{D} \frac{(x_j - \mu_{cj})^2}{2\sigma_j^2} + \log \pi_c \tag{4.64}$$

Typically we set $\hat{\mu}_{cj} = \overline{x}_{cj}$ and $\hat{\sigma}_j^2 = s_j^2$, which is the **pooled empirical variance** of feature j (pooled across classes) defined by

$$s_j^2 = \frac{\sum_{c=1}^{C} \sum_{i:y_i=c} (x_{ij} - \overline{x}_{cj})^2}{N - C} \tag{4.65}$$

Pooled variance: estimate coming variance of populations with different means