Yale University

Generative Classification: Bayes, LDA, and QDA

S&DS 265 · Introductory Machine Learning · Lecture 7

Zhuoran Yang

Fall 2026

AI usage: figures were generated, and these slides polished, with the help of AI tools.

Outline

1. Classification Problem Setup
2. Bayes Classification
3. Gaussian Discriminant Analysis
4. Fitted Boundaries on Real Data

Learning Objectives

In this lecture you will learn:

1. How a generative model differs from a discriminative one;
2. How priors and class-conditional densities combine into posterior probabilities;
3. Why unequal mistake costs shift the decision threshold;
4. How a Gaussian likelihood estimates class priors, means, and covariances;
5. Why the covariance assumption alone decides whether the boundary is linear or quadratic.

Outline

1. Classification Problem Setup
2. Bayes Classification
3. Gaussian Discriminant Analysis
4. Fitted Boundaries on Real Data

Motivating Example: Credit Default

A bank has records for 10,000 customers, of whom only 333 defaulted. For a new customer, it must estimate default probability and decide whether the account needs review.

How should a rare class, the observed feature values, and the costs of the two kinds of error combine?

Example and data: ISLP §4.1–4.4 (James, Witten, Hastie, Tibshirani, and Taylor, 2023).

The Default Data

default	student	balance	income
No	No	729.5	44,362
No	Yes	817.2	12,106
No	No	1,073.5	31,767
⋮	⋮	⋮	⋮
No	Yes	200.9	16,863

Source: The `Default` data, ISLP §4.3; available at `statlearning.com`.

10,000 customers

- 333 defaults
- 9,667 non-defaults
- Two numerical features
- One binary feature

Each row is one customer. `balance` is average credit-card balance and `income` is annual income, both measured in US dollars; `default` is the prediction target.

Reading One Row

customer	default	student	balance	income
1	No	No	729.5	44,362

Customer 1 was not a student, carried a balance of about $730, earned about $44,362, and did not default. Generative classification asks how plausible this feature vector is within each possible class.

Problem Setup

Training data. $\mathcal{D}_n = \{(x_i, y_i)\}_{i=1}^n$, drawn independently from an unknown population distribution P.

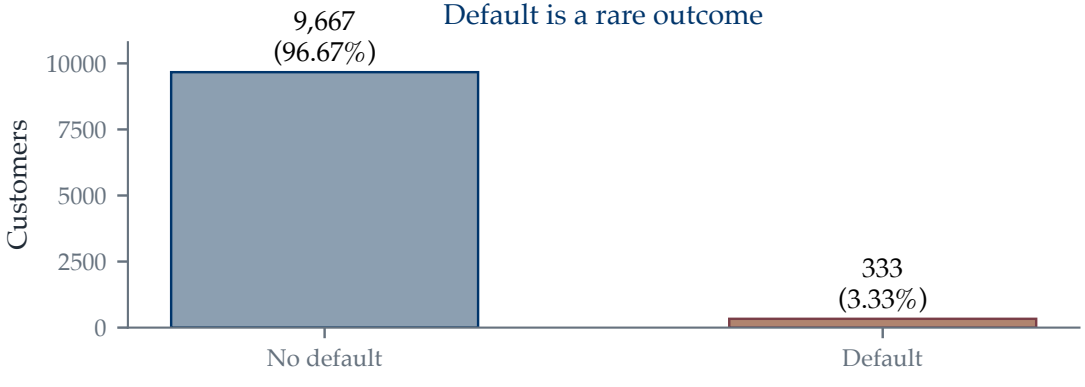
Features. $x_i \in \mathbb{R}^d$ contains the measured customer characteristics.

Response. $y_i \in \{0, 1\}$ records non-default or default.

Goal. Estimate $\eta(x^*) = \mathbb{P}(Y = 1 \mid X = x^*)$ for an unseen customer, then choose a decision using the costs of the two errors.

For `Default`: $n = 10{,}000$, $d = 3$, and $x_i = (\text{student}, \text{balance}, \text{income})^\top$. Predicting class 0 for everyone is 96.67% accurate but detects no defaults.

Class Prevalence in the `Default` Data



Course calculation from the ISLP `Default` data.

Regression and Classification Have Different Targets

Regression	Classification
numerical response Y	categorical response Y
target $\mathbb{E}[Y \mid X = x]$	target $\mathbb{P}(Y = k \mid X = x)$
squared error is common	mistakes can have unequal costs

The probability model, decision rule, loss, and evaluation metric are separate choices.

Two Routes to a Classifier

The target has not changed. For binary $Y \in \{0, 1\}$,

$$\mathbb{E}[Y \mid X = x] = \mathbb{P}(Y = 1 \mid X = x),$$

the same conditional-mean target as in regression; only the response is categorical.

Two Routes to a Classifier

The target has not changed. For binary $Y \in \{0, 1\}$,

$$\mathbb{E}[Y \mid X = x] = \mathbb{P}(Y = 1 \mid X = x),$$

the same conditional-mean target as in regression; only the response is categorical.

Two ways to reach it, differing in how much they model:

- **Discriminative.** Model only the conditional relation of Y given X.
- **Generative.** Model the joint distribution of (X, Y), then read the conditional distribution $\mathbb{P}(Y = 1 \mid X = x)$ from the joint.

Two Routes to a Classifier

The target has not changed. For binary $Y \in \{0, 1\}$,

$$\mathbb{E}[Y \mid X = x] = \mathbb{P}(Y = 1 \mid X = x),$$

the same conditional-mean target as in regression; only the response is categorical.

Two ways to reach it, differing in how much they model:

- **Discriminative.** Model only the conditional relation of Y given X.
- **Generative.** Model the joint distribution of (X, Y), then read the conditional distribution $\mathbb{P}(Y = 1 \mid X = x)$ from the joint.

This lecture takes the generative route; Lecture 8 takes the discriminative one on the same data.

The Generative Route

Model the joint by describing how a row is produced:

1. Draw a class $Y = k$ with probability $\pi_k = \mathbb{P}(Y = k)$, the **prior**;
2. Draw features from that class, $X \sim p_k$, the **class-conditional density**.

That is exactly the factorization $p(x, Y = k) = \pi_k \, p_k(x)$.

The Generative Route

Model the joint by describing how a row is produced:

1. Draw a class $Y = k$ with probability $\pi_k = \mathbb{P}(Y = k)$, the **prior**;
2. Draw features from that class, $X \sim p_k$, the **class-conditional density**.

That is exactly the factorization $p(x, Y = k) = \pi_k \, p_k(x)$.

Both factors can be estimated from data. The target itself cannot be, directly: no two rows share an x.

Why the Factors Are Estimable

- $\mathbb{P}(Y = 1 \mid X = x)$ at one x has nothing to average: with continuous features that x occurs once, or never.
- π_k is a proportion — the share of rows in class k. Ordinary estimation, with all n rows behind it.
- p_k is a density fitted to the n_k rows of class k alone, a separate and self-contained estimation problem.

Why the Factors Are Estimable

- $\mathbb{P}(Y = 1 \mid X = x)$ at one x has nothing to average: with continuous features that x occurs once, or never.
- π_k is a proportion — the share of rows in class k. Ordinary estimation, with all n rows behind it.
- p_k is a density fitted to the n_k rows of class k alone, a separate and self-contained estimation problem.

Bayes' rule then reassembles the two into the target, which is where we go next.

Source: Same difficulty as Lecture 6: a conditional at a point is not directly estimable, so something must be assumed.

Outline

1. Classification Problem Setup
2. Bayes Classification
3. Gaussian Discriminant Analysis
4. Fitted Boundaries on Real Data

Bayes' Rule Combines Prevalence and Feature Evidence

The joint density factors in two directions:

$$p(x, Y = k) = \mathbb{P}(Y = k)\, p(x \mid Y = k) = p(x)\, \mathbb{P}(Y = k \mid X = x).$$

Solving for the posterior gives

$$\boxed{\mathbb{P}(Y = k \mid X = x) = \frac{\pi_k p_k(x)}{\sum_{j=1}^{K} \pi_j p_j(x)}}.$$

π_k measures prevalence; $p_k(x)$ measures how typical x is within class k.

A Rare Class Needs Strong Evidence

Take one customer with card balance $x = \$1{,}800$. Fitting a density to each class gives

$$p_1(x) = 11.3 \times 10^{-4}, \qquad p_0(x) = 0.99 \times 10^{-4},$$

so this balance is about 11× more typical of defaulters. Yet

$$\mathbb{P}(Y = 1 \mid x) = \frac{0.033 \times 11.3}{0.033 \times 11.3 + 0.967 \times 0.99} = \frac{3.75}{3.75 + 9.58} = 0.28.$$

A Rare Class Needs Strong Evidence

Take one customer with card balance $x = \$1{,}800$. Fitting a density to each class gives

$$p_1(x) = 11.3 \times 10^{-4}, \qquad p_0(x) = 0.99 \times 10^{-4},$$

so this balance is about 11× more typical of defaulters. Yet

$$\mathbb{P}(Y = 1 \mid x) = \frac{0.033 \times 11.3}{0.033 \times 11.3 + 0.967 \times 0.99} = \frac{3.75}{3.75 + 9.58} = 0.28.$$

Being 11 times more typical of defaulters is not enough, because defaulters are 29 times rarer to begin with.

A Rare Class Needs Strong Evidence

Take one customer with card balance $x = \$1{,}800$. Fitting a density to each class gives

$$p_1(x) = 11.3 \times 10^{-4}, \qquad p_0(x) = 0.99 \times 10^{-4},$$

so this balance is about 11× more typical of defaulters. Yet

$$\mathbb{P}(Y = 1 \mid x) = \frac{0.033 \times 11.3}{0.033 \times 11.3 + 0.967 \times 0.99} = \frac{3.75}{3.75 + 9.58} = 0.28.$$

Being 11 times more typical of defaulters is not enough, because defaulters are 29 times rarer to begin with.

At $x = \$2{,}000$ the ratio reaches 40, and the probability crosses a half at 0.58.

Source: Course calculation on the ISLP `Default` data.

The Bayes Classifier

derive

Suppose a rule answers a at the query x. Its error there is

$$\mathbb{P}(Y \neq a \mid X = x) = 1 - \mathbb{P}(Y = a \mid X = x).$$

The Bayes Classifier

derive

Suppose a rule answers a at the query x. Its error there is

$$\mathbb{P}(Y \neq a \mid X = x) = 1 - \mathbb{P}(Y = a \mid X = x).$$

To make this small, pick the a with the largest posterior. Doing so at every x defines the **Bayes classifier**

$$\boxed{h^\star(x) = \arg\max_k \ \mathbb{P}(Y = k \mid X = x) = \arg\max_k \ \pi_k p_k(x)}$$

the second form because Bayes' denominator is the same for every class.

The Bayes Classifier

derive

Suppose a rule answers a at the query x. Its error there is

$$\mathbb{P}(Y \neq a \mid X = x) = 1 - \mathbb{P}(Y = a \mid X = x).$$

To make this small, pick the a with the largest posterior. Doing so at every x defines the **Bayes classifier**

$$h^\star(x) = \arg\max_k \ \mathbb{P}(Y = k \mid X = x) = \arg\max_k \ \pi_k p_k(x)$$

the second form because Bayes' denominator is the same for every class.

It is optimal because it minimizes the error separately at each x: no rule can beat it anywhere, hence not on average.

The Two-Class Rule Is a Likelihood-Ratio Test

derive

Predict class 1 when it is the more probable class. The common denominator of the two posteriors cancels, so

$$\mathbb{P}(Y = 1 \mid x) > \mathbb{P}(Y = 0 \mid x)$$

$$\pi_1 p_1(x) > \pi_0 p_0(x)$$

$$\underbrace{\frac{p_1(x)}{p_0(x)}}_{\text{likelihood ratio}} > \underbrace{\frac{\pi_0}{\pi_1}}_{\text{prior odds against}} .$$

The Two-Class Rule Is a Likelihood-Ratio Test

derive

Predict class 1 when it is the more probable class. The common denominator of the two posteriors cancels, so

$$\mathbb{P}(Y = 1 \mid x) > \mathbb{P}(Y = 0 \mid x)$$

$$\pi_1 p_1(x) > \pi_0 p_0(x)$$

$$\underbrace{\frac{p_1(x)}{p_0(x)}}_{\text{likelihood ratio}} > \underbrace{\frac{\pi_0}{\pi_1}}_{\text{prior odds against}} .$$

The decision compares one number against one number, whatever d is: all of x enters only through the ratio.

Reading the Likelihood Ratio as Evidence

- $p_k(x)$ is not the probability of class k. It says how typical the observed x is of the rows in class k.
- Their ratio $p_1(x)/p_0(x)$ is therefore **evidence**: how much more typical x is of one class than the other. It ignores prevalence entirely.
- The threshold π_0/π_1 is the **base rate**, and carries prevalence alone. It ignores x entirely.

Reading the Likelihood Ratio as Evidence

- $p_k(x)$ is not the probability of class k. It says how typical the observed x is of the rows in class k.
- Their ratio $p_1(x)/p_0(x)$ is therefore **evidence**: how much more typical x is of one class than the other. It ignores prevalence entirely.
- The threshold π_0/π_1 is the **base rate**, and carries prevalence alone. It ignores x entirely.

So the rule splits cleanly: evidence from this customer against the base rate of the population. On `Default` the threshold is $0.967/0.033 = 29$, so a balance must be 29 times more typical of defaulters before the decision flips.

Unequal Costs Change the Decision Threshold

derive

Let C_{10} be the cost of a false positive and C_{01} the cost of a false negative. Predicting a at x costs, on average,

$$r(1 \mid x) = C_{10}\,\mathbb{P}(Y = 0 \mid x), \qquad r(0 \mid x) = C_{01}\,\mathbb{P}(Y = 1 \mid x).$$

Unequal Costs Change the Decision Threshold

derive

Let C_{10} be the cost of a false positive and C_{01} the cost of a false negative. Predicting a at x costs, on average,

$$r(1 \mid x) = C_{10}\,\mathbb{P}(Y = 0 \mid x), \qquad r(0 \mid x) = C_{01}\,\mathbb{P}(Y = 1 \mid x).$$

Predict 1 when $r(1 \mid x) < r(0 \mid x)$. Writing the posteriors as $\pi_k p_k(x)$ over a common denominator, that denominator cancels:

$$C_{10}\,\pi_0 p_0(x) < C_{01}\,\pi_1 p_1(x) \quad \Longleftrightarrow \quad \boxed{\frac{p_1(x)}{p_0(x)} > \frac{C_{10}}{C_{01}} \cdot \frac{\pi_0}{\pi_1}}$$

Unequal Costs Change the Decision Threshold

derive

Let C_{10} be the cost of a false positive and C_{01} the cost of a false negative. Predicting a at x costs, on average,

$$r(1 \mid x) = C_{10}\,\mathbb{P}(Y = 0 \mid x), \qquad r(0 \mid x) = C_{01}\,\mathbb{P}(Y = 1 \mid x).$$

Predict 1 when $r(1 \mid x) < r(0 \mid x)$. Writing the posteriors as $\pi_k p_k(x)$ over a common denominator, that denominator cancels:

$$C_{10}\,\pi_0 p_0(x) < C_{01}\,\pi_1 p_1(x) \quad \Longleftrightarrow \quad \boxed{\frac{p_1(x)}{p_0(x)} > \frac{C_{10}}{C_{01}} \cdot \frac{\pi_0}{\pi_1}}$$

The same likelihood-ratio test, with the threshold multiplied by the cost ratio. Costs and prevalence enter in exactly the same way.

What the Threshold Does on `Default`

cost of a missed default	threshold	flag balances above	customers flagged
equal, $C_{01} = C_{10}$	29.0	\$1,950	1.3%
5× a false alarm	5.8	\$1,667	5.0%
10× a false alarm	2.9	\$1,510	8.7%

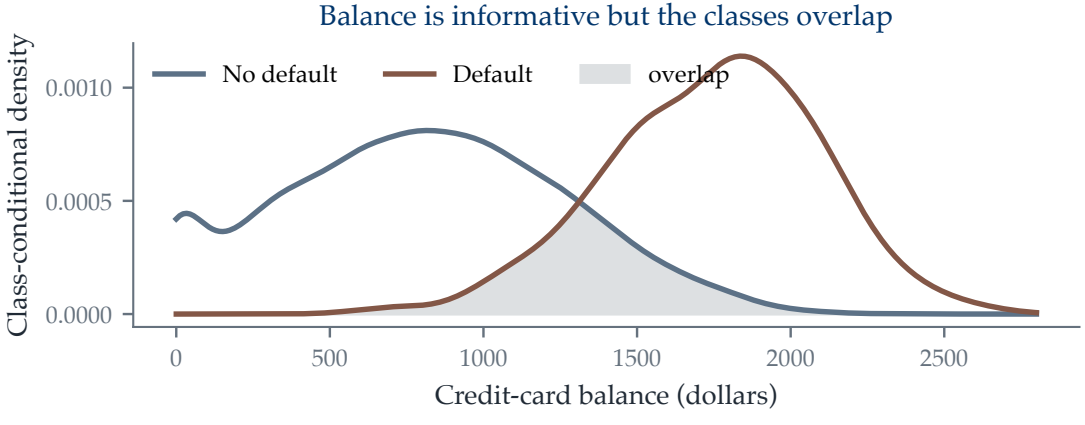
What the Threshold Does on `Default`

cost of a missed default	threshold	flag balances above	customers flagged
equal, $C_{01} = C_{10}$	29.0	\$1,950	1.3%
5× a false alarm	5.8	\$1,667	5.0%
10× a false alarm	2.9	\$1,510	8.7%

- The model is fitted once. Only the threshold moves, so a change of policy needs no refitting.
- Deciding what a missed default costs relative to a false alarm is a business question, not a statistical one — but it fully determines where the cut falls.

Source: Course calculation on the ISLP `Default` data.

Class Overlap Creates Irreducible Classification Error



Course calculation from the ISLP `Default` data.

Bayes Risk

Averaging the pointwise error over the distribution of X gives the population error rate of a rule h,

$$R(h) = \mathbb{P}\{Y \neq h(X)\} = \mathbb{E}[1 - \mathbb{P}\{Y = h(X) \mid X\}].$$

Bayes Risk

Averaging the pointwise error over the distribution of X gives the population error rate of a rule h,

$$R(h) = \mathbb{P}\{Y \neq h(X)\} = \mathbb{E}[1 - \mathbb{P}\{Y = h(X) \mid X\}].$$

Evaluating this at $h^\star$ defines the **Bayes risk**

$$\boxed{R^\star = R(h^\star) = \mathbb{E}[1 - \max_k \mathbb{P}(Y = k \mid X)]}$$

the smallest error rate any rule can achieve on this population.

Why $R^\star$ Cannot Be Driven to Zero

- At $x = \$1{,}800$ the posterior is 0.28: some such customers default and some do not.
- Knowing the true posterior exactly does not help. A decision must name one class, and it is wrong for the other 28%.
- Averaging that unavoidable error over all balances gives $R^\star \approx 2.7\%$, against 3.3% for always predicting no default.

Why $R^\star$ Cannot Be Driven to Zero

- At $x = \$1{,}800$ the posterior is 0.28: some such customers default and some do not.
- Knowing the true posterior exactly does not help. A decision must name one class, and it is wrong for the other 28%.
- Averaging that unavoidable error over all balances gives $R^\star \approx 2.7\%$, against 3.3% for always predicting no default.

$R^\star$ belongs to the population, not the model or the sample. It plays the role σ^2 played in regression; finite data and a wrong model class add avoidable error on top.

What Is Left to Estimate

Bayes' rule has reduced classification to two unknowns,

$$\pi_k = \mathbb{P}(Y = k), \qquad p_k(x) = p(x \mid Y = k),$$

and they are not equally hard.

What Is Left to Estimate

Bayes' rule has reduced classification to two unknowns,

$$\pi_k = \mathbb{P}(Y = k), \qquad p_k(x) = p(x \mid Y = k),$$

and they are not equally hard.

- π_k is a distribution over K values. Estimating it is counting, and with $n = 10{,}000$ rows it is accurate.
- p_k is a density on $\mathbb{R}^d$. Estimating a density with no assumptions meets the same wall as Lecture 6: neighbours vanish as d grows.

What Is Left to Estimate

Bayes' rule has reduced classification to two unknowns,

$$\pi_k = \mathbb{P}(Y = k), \qquad p_k(x) = p(x \mid Y = k),$$

and they are not equally hard.

- π_k is a distribution over K values. Estimating it is counting, and with $n = 10{,}000$ rows it is accurate.
- p_k is a density on $\mathbb{R}^d$. Estimating a density with no assumptions meets the same wall as Lecture 6: neighbours vanish as d grows.

So we do what we did for regression: restrict p_k to a parametric class and estimate finitely many numbers. The next section takes that class to be Gaussian.

Outline

1. Classification Problem Setup
2. Bayes Classification
3. Gaussian Discriminant Analysis
4. Fitted Boundaries on Real Data

One Gaussian Cloud per Class

$$X \mid Y = k \sim \mathcal{N}(\mu_k, \Sigma_k), \qquad \mathbb{P}(Y = k) = \pi_k.$$

Center μ_k

typical feature vector for class k

Shape Σ_k

scale, correlation, and orientation within class k

For this derivation, X contains the numerical balance and income variables. A binary feature such as student status can instead receive a class-conditional Bernoulli model and contribute an additional log-score term.

The Gaussian Generative Model Class

The model assigns a joint density to a feature–label pair:

$$p_\theta(x, Y = k) = \pi_k \, \phi_d(x; \mu_k, \Sigma_k), \qquad \theta = \{\pi_k, \mu_k, \Sigma_k\}_{k=1}^{K}.$$

Different restrictions on Σ_k define different model classes:

model	restriction	consequence
LDA	$\Sigma_1 = \cdots = \Sigma_K$	linear boundaries
QDA	separate full Σ_k	quadratic boundaries
Gaussian naive Bayes	diagonal Σ_k	additive feature evidence

The assumptions come before estimation: data choose parameters within the selected class, not the class itself.

The Fitted Estimators

Maximum likelihood on the joint model separates into one problem per class, and every answer is what you would have guessed:

$$\widehat{\pi}_k = \frac{n_k}{n}, \qquad \widehat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i, \qquad \widehat{\Sigma}_k = \frac{1}{n_k} \sum_{i:y_i=k} (x_i - \widehat{\mu}_k)(x_i - \widehat{\mu}_k)^\top.$$

The Fitted Estimators

Maximum likelihood on the joint model separates into one problem per class, and every answer is what you would have guessed:

$$\widehat{\pi}_k = \frac{n_k}{n}, \qquad \hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i, \qquad \widehat{\Sigma}_k = \frac{1}{n_k} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^\top.$$

Class share, class mean, class scatter. For `Default`, $\widehat{\pi}_1 = 333/10{,}000 = 0.033$.

The Fitted Estimators

Maximum likelihood on the joint model separates into one problem per class, and every answer is what you would have guessed:

$$\widehat{\pi}_k = \frac{n_k}{n}, \qquad \hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i, \qquad \widehat{\Sigma}_k = \frac{1}{n_k} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^\top.$$

Class share, class mean, class scatter. For `Default`, $\widehat{\pi}_1 = 333/10{,}000 = 0.033$.

LDA differs in one place only: it pools the scatter across classes into a single $\widehat{\Sigma}$, since it assumes there is only one.

Source: Derived in the appendix.

Covariance Estimation Determines the Model Family

A class-specific covariance estimate is

$$\widehat{\Sigma}_k = \frac{1}{n_k - 1} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^\top.$$

A pooled estimate is

$$\widehat{\Sigma} = \frac{1}{n - K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^\top.$$

QDA uses $\widehat{\Sigma}_k$; LDA uses the shared $\widehat{\Sigma}$. These conventional estimates include degrees-of-freedom corrections. Exact Gaussian MLE uses denominators n_k and n; software conventions should be reported because small classes make the difference visible.

The Gaussian Log Posterior Is a Discriminant Score

Ignoring the shared normalizing denominator in Bayes' rule,

$$\log\{\pi_k p_k(x)\} = \log \pi_k - \frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^\top \Sigma_k^{-1}(x-\mu_k).$$

Dropping only terms common to all classes gives

$$\delta_k(x) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)^\top \Sigma_k^{-1}(x-\mu_k) + \log \pi_k.$$

Three Terms Determine a Gaussian Score

$$\delta_k(x) = \underbrace{\log \pi_k}_{\text{prevalence}} \underbrace{-\frac{1}{2}\log|\Sigma_k|}_{\text{cloud volume}} \underbrace{-\frac{1}{2}(x-\mu_k)^\top \Sigma_k^{-1}(x-\mu_k)}_{\text{Mahalanobis distance}}.$$

- A larger prior raises the score before observing any features.
- A point far from μ_k relative to the cloud's scale lowers the score.
- The log-determinant prevents a diffuse cloud from assigning high evidence everywhere merely because its distances are small.

Prediction compares these terms across classes; an isolated density value has no class meaning by itself.

LDA Follows from a Shared Covariance

When $\Sigma_k = \Sigma$, expand the quadratic:

$$\begin{aligned}(x - \mu_k)^\top \Sigma^{-1} (x - \mu_k) \\ = x^\top \Sigma^{-1} x - 2x^\top \Sigma^{-1} \mu_k + \mu_k^\top \Sigma^{-1} \mu_k.\end{aligned}$$

The terms $-\frac{1}{2} \log |\Sigma|$ and $-\frac{1}{2} x^\top \Sigma^{-1} x$ are common to every class, leaving

$$\boxed{\delta_k^{\mathrm{LDA}}(x) = x^\top \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^\top \Sigma^{-1} \mu_k + \log \pi_k}.$$

LDA Produces a Linear Boundary

derive

Compare classes 1 and 0. With a shared Σ the term $x^\top \Sigma^{-1} x$ appears in both scores and cancels:

$$\begin{aligned}\delta_1(x) - \delta_0(x) &= x^\top \Sigma^{-1}(\mu_1 - \mu_0) - \tfrac{1}{2}(\mu_1^\top \Sigma^{-1}\mu_1 - \mu_0^\top \Sigma^{-1}\mu_0) + \log \tfrac{\pi_1}{\pi_0} \\ &= w^\top x + b.\end{aligned}$$

LDA Produces a Linear Boundary

derive

Compare classes 1 and 0. With a shared Σ the term $x^\top \Sigma^{-1} x$ appears in both scores and cancels:

$$\begin{aligned}\delta_1(x) - \delta_0(x) &= x^\top \Sigma^{-1}(\mu_1 - \mu_0) - \tfrac{1}{2}(\mu_1^\top \Sigma^{-1}\mu_1 - \mu_0^\top \Sigma^{-1}\mu_0) + \log \tfrac{\pi_1}{\pi_0} \\ &= w^\top x + b.\end{aligned}$$

Everything that depended on x quadratically has gone. The boundary $\delta_1(x) = \delta_0(x)$ is the set where $w^\top x + b = 0$.

Why $w^\top x + b = 0$ Is a Hyperplane

Take any two points x and x' on the set. Subtracting $w^\top x + b = 0$ from $w^\top x' + b = 0$ gives

$$w^\top (x' - x) = 0,$$

so every direction lying inside the set is perpendicular to w.

Why $w^\top x + b = 0$ Is a Hyperplane

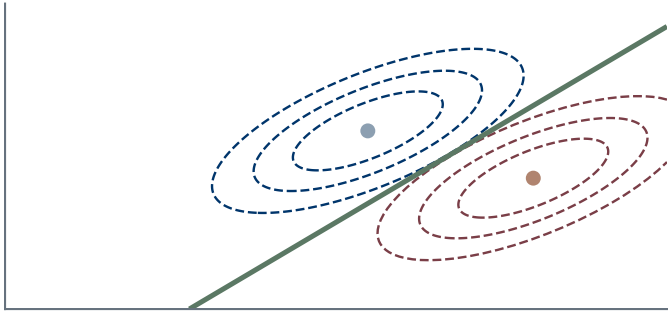
Take any two points x and x' on the set. Subtracting $w^\top x + b = 0$ from $w^\top x' + b = 0$ gives

$$w^\top (x' - x) = 0,$$

so every direction lying inside the set is perpendicular to w.

- That is what a **hyperplane** is: a flat set with one normal direction w — a line in $\mathbb{R}^2$, a plane in $\mathbb{R}^3$.
- Moving along w changes the score fastest; moving perpendicular to w does not change it at all.
- The offset b slides the plane without rotating it, which is what a change of prior or cost does.

LDA: One Shared Shape



shared Σ: the quadratic terms cancel, leaving a line

Two declared Gaussians with the same Σ, equal priors. Course illustration; no data are fitted.

QDA Keeps the Quadratic Terms

With separate covariances nothing cancels, because $x^\top \Sigma_1^{-1} x$ and $x^\top \Sigma_0^{-1} x$ differ:

$$\delta_1(x) - \delta_0(x) = x^\top A x + b^\top x + c, \qquad A = \tfrac{1}{2}(\Sigma_0^{-1} - \Sigma_1^{-1}).$$

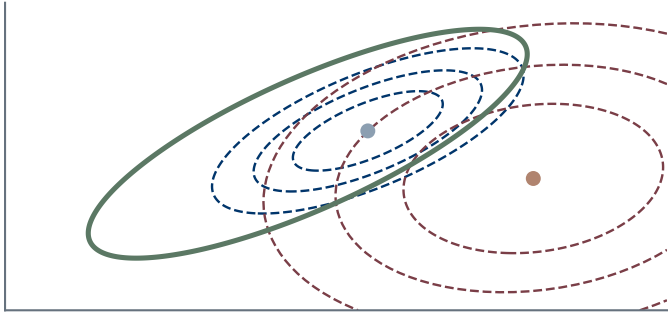
QDA Keeps the Quadratic Terms

With separate covariances nothing cancels, because $x^\top \Sigma_1^{-1} x$ and $x^\top \Sigma_0^{-1} x$ differ:

$$\delta_1(x) - \delta_0(x) = x^\top A x + b^\top x + c, \qquad A = \tfrac{1}{2}(\Sigma_0^{-1} - \Sigma_1^{-1}).$$

A quadratic in x set to zero is a **conic**: an ellipse, a parabola, or a hyperbola, which can have two branches. If $\Sigma_0 = \Sigma_1$ then $A = 0$ and QDA reduces exactly to LDA.

QDA: A Shape per Class



separate Σ_k: the quadratic terms survive, leaving a conic

The same two means with different Σ_k. Course illustration; no data are fitted.

Naive Bayes Factorizes the Class-Conditional Density

Conditional independence assumes

$$p(x \mid Y = k) = \prod_{j=1}^{d} p(x_j \mid Y = k).$$

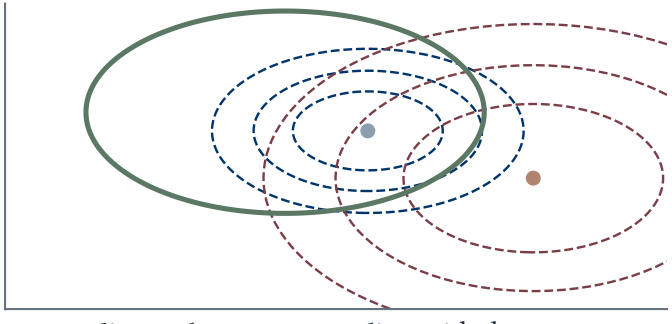
Hence the log score becomes additive:

$$\delta_k(x) = \log \pi_k + \sum_{j=1}^{d} \log p(x_j \mid Y = k).$$

Bias–variance tradeoff

Ignoring conditional correlations adds bias, but estimating one-dimensional densities can sharply reduce variance.

Naive Bayes: Axis-Aligned Shapes



diagonal Σ_k: contours align with the axes

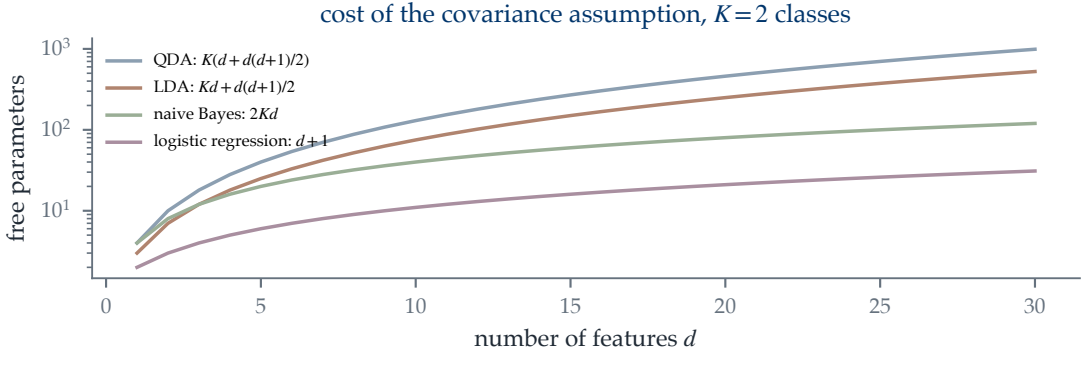
Independence within a class forces Σ_k diagonal, so no contour tilts. Course illustration; no data are fitted.

Covariance Flexibility Has a Parameter Cost

model	covariance assumption	covariance parameters
LDA	one shared full Σ	$d(d+1)/2$
QDA	one full Σ_k per class	$Kd(d+1)/2$
Gaussian naive Bayes	diagonal Σ_k	Kd

When d is large relative to each n_k, full covariance estimates can be unstable or singular.

The Cost Grows Quadratically in d



At $d = 30$ and $K = 2$: QDA must estimate 990 numbers, LDA 525, naive Bayes 120, and logistic regression — which models the boundary directly and is the subject of Lecture 8 — only 31.

Exact parameter counts; no data are involved.

The Fitted Classifier Is a Plug-In Rule

Training replaces every unknown population quantity by its estimate:

$$\hat{\eta}_k(x) = \frac{\widehat{\pi}_k \, \phi_d(x; \hat{\mu}_k, \widehat{\Sigma}_k)}{\sum_{j=1}^{K} \widehat{\pi}_j \, \phi_d(x; \hat{\mu}_j, \widehat{\Sigma}_j)}, \qquad \hat{h}(x) = \arg\max_k \hat{\eta}_k(x).$$

1. Choose LDA, QDA, or naive Bayes before fitting;
2. Estimate its priors and class-conditional parameters;
3. Evaluate one fitted score for each class at a new x;
4. Normalize scores for probabilities, then apply the decision costs.

This is called plug-in classification because estimated parameters are plugged into the population Bayes rule.

Summary: From Population Target to Fitted Rule

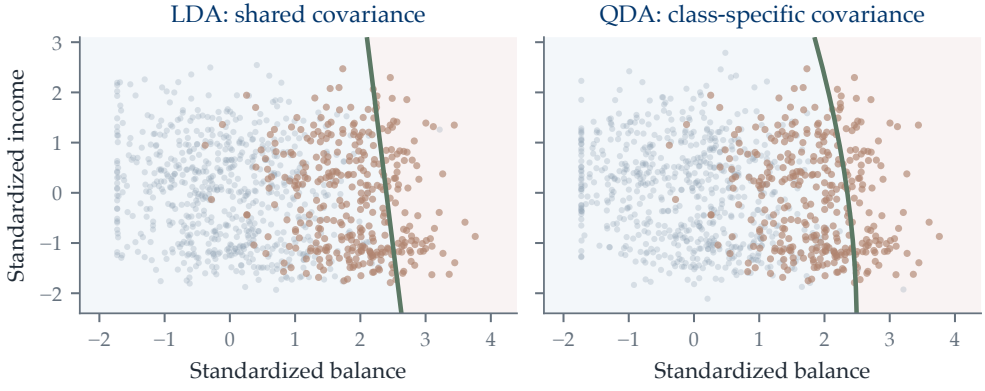
population target	posterior $\eta_k(x) = \mathbb{P}(Y = k \mid X = x)$
model class	$p(x, Y = k; \theta) = \pi_k p_k(x; \theta_k)$
fitting principle	choose $\hat{\theta}$ making the observed pairs most likely
posterior estimate	$\hat{\eta}_k(x) \propto \widehat{\pi}_k p_k(x; \hat{\theta}_k)$
decision	choose the class with smallest fitted conditional cost

Every arrow in this chain has now been derived: Bayes' rule gives the third and fourth, maximum likelihood the middle one, and the covariance choice fixes which family sits in the second.

Outline

1. Classification Problem Setup
2. Bayes Classification
3. Gaussian Discriminant Analysis
4. Fitted Boundaries on Real Data

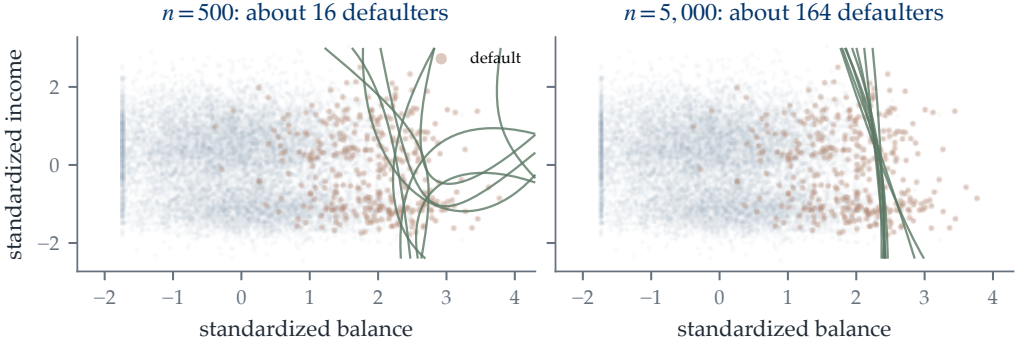
LDA and QDA on the `Default` Data



Fitted on all 10,000 rows. The two boundaries nearly coincide: the class covariances really are similar here, so QDA's extra flexibility buys almost nothing. The geometric difference the algebra promised need not show up in a given data set.

Course fits to the ISLP `Default` data.

What QDA's Flexibility Costs



The same procedure refitted on eight balanced subsamples. With 20 rows per class the covariance estimates are noisy and the boundary swings wildly; with 250 the eight curves nearly coincide. Flexibility is paid for in variance, and the rare class is what limits it — `Default` has only 333 defaulters.

Course resampling of the ISLP `Default` data.

Which Model, and When

All three sit on one axis: how many covariance parameters to estimate.

- **QDA** only when the classes visibly differ in shape and each has the rows to support $d(d+1)/2$ numbers of its own.
- **LDA** when the shapes are similar, or the rare class is small: pooling lets every row help estimate the one Σ.
- **Naive Bayes** when even that is too many. It assumes away all within-class correlation, buying low variance by accepting bias.

Which Model, and When

All three sit on one axis: how many covariance parameters to estimate.

- **QDA** only when the classes visibly differ in shape and each has the rows to support $d(d+1)/2$ numbers of its own.
- **LDA** when the shapes are similar, or the rare class is small: pooling lets every row help estimate the one Σ.
- **Naive Bayes** when even that is too many. It assumes away all within-class correlation, buying low variance by accepting bias.

At $d = 30$ one covariance costs 465 numbers. `Default` has 333 defaulters, so QDA there would be estimating more parameters than the rare class has rows.

Summary: The Generative Recipe

- **Assume.** A row is produced in two steps: a class from the prior π_k, then features from p_k. Choosing a family for p_k is the modelling step.
- **Fit.** $\widehat{\pi}_k$ is the class share, $\hat{\mu}_k$ the class mean, and $\widehat{\Sigma}$ a pooled or per-class scatter.
- **Predict.** Bayes' rule inverts the generation. With K classes, compute one score $\delta_k(x) = \log \widehat{\pi}_k + \log p_k(x; \hat{\theta}_k)$ per class and take the largest — shifted by log of the cost ratio when mistakes differ in cost.

Summary: The Generative Recipe

- **Assume.** A row is produced in two steps: a class from the prior π_k, then features from p_k. Choosing a family for p_k is the modelling step.
- **Fit.** $\widehat{\pi}_k$ is the class share, $\hat{\mu}_k$ the class mean, and $\widehat{\Sigma}$ a pooled or per-class scatter.
- **Predict.** Bayes' rule inverts the generation. With K classes, compute one score $\delta_k(x) = \log \widehat{\pi}_k + \log p_k(x; \hat{\theta}_k)$ per class and take the largest — shifted by log of the cost ratio when mistakes differ in cost.

The covariance assumption alone fixes the boundary: shared gives a line (LDA), per-class a conic (QDA), diagonal an axis-aligned one (naive Bayes).

Summary: What the Assumption Buys and Costs

model	parameters at $d = 30$	boundary
QDA	990	conic
LDA	525	linear
Gaussian naive Bayes	120	from axis-aligned densities
logistic regression	31	linear, fitted directly

Summary: What the Assumption Buys and Costs

model	parameters at $d = 30$	boundary
QDA	990	conic
LDA	525	linear
Gaussian naive Bayes	120	from axis-aligned densities
logistic regression	31	linear, fitted directly

- A density model is a strong assumption. It pays off when each class has few rows, and costs bias when the assumed shape is wrong.
- Flexibility is paid for in variance, and it is the rarest class that sets the price — as the eight QDA boundaries showed.

Next Lecture

We have modelled the whole joint distribution in order to classify. The last row of that table asks an obvious question: if the boundary is all we need, why estimate 990 numbers to find 31?

Lecture 8 — Discriminative Classification takes the other route on the same `Default` data: model $\mathbb{P}(Y = 1 \mid X = x)$ directly, derive cross-entropy, and fit logistic regression with Lecture 5's gradient methods.

Reading for this lecture: ISLP Chapter 4.

Part A

Why the Bayes Rule Is Optimal

A.1 The Bayes Classifier Minimizes Risk

derive

Write $\eta_k(x) = \mathbb{P}(Y = k \mid X = x)$. For any rule h, condition on X:

$$R(h) = \mathbb{P}\{Y \neq h(X)\} = \mathbb{E}\big[\mathbb{P}\{Y \neq h(X) \mid X\}\big] = \mathbb{E}\big[1 - \eta_{h(X)}(X)\big].$$

A.1 The Bayes Classifier Minimizes Risk

derive

Write $\eta_k(x) = \mathbb{P}(Y = k \mid X = x)$. For any rule h, condition on X:

$$R(h) = \mathbb{P}\{Y \neq h(X)\} = \mathbb{E}\big[\mathbb{P}\{Y \neq h(X) \mid X\}\big] = \mathbb{E}\big[1 - \eta_{h(X)}(X)\big].$$

At every x, by the definition of a maximum,

$$\eta_{h(x)}(x) \;\leq\; \max_k \eta_k(x) \qquad \Longrightarrow \qquad 1 - \eta_{h(x)}(x) \;\geq\; 1 - \max_k \eta_k(x).$$

A.1 The Bayes Classifier Minimizes Risk

derive

Write $\eta_k(x) = \mathbb{P}(Y = k \mid X = x)$. For any rule h, condition on X:

$$R(h) = \mathbb{P}\{Y \neq h(X)\} = \mathbb{E}\big[\mathbb{P}\{Y \neq h(X) \mid X\}\big] = \mathbb{E}\big[1 - \eta_{h(X)}(X)\big].$$

At every x, by the definition of a maximum,

$$\eta_{h(x)}(x) \;\leq\; \max_k \eta_k(x) \qquad \Longrightarrow \qquad 1 - \eta_{h(x)}(x) \;\geq\; 1 - \max_k \eta_k(x).$$

The right-hand side does not involve h, and expectation preserves the inequality, so for every h

$$\boxed{R(h) \;\geq\; \mathbb{E}\big[1 - \max_k \eta_k(X)\big] = R(h^\star) = R^\star}$$

A.2 What a Wrong Decision Costs

Subtracting the two lines gives the **excess risk**

$$R(h) - R^\star = \mathbb{E}[\max_k \eta_k(X) - \eta_{h(X)}(X)] \;\geq\; 0,$$

the average posterior gap between the class chosen and the best class.

A.2 What a Wrong Decision Costs

Subtracting the two lines gives the **excess risk**

$$R(h) - R^\star = \mathbb{E}[\max_k \eta_k(X) - \eta_{h(X)}(X)] \;\geq\; 0,$$

the average posterior gap between the class chosen and the best class.

- A rule is penalized only where it disagrees with $h^\star$, and then only by how close the call was.
- Disagreeing where the posterior is near $1/2$ costs almost nothing; disagreeing where one class is nearly certain costs a lot.
- So a fitted rule can differ from $h^\star$ over a wide region and still have risk close to $R^\star$, provided it errs where the classes overlap.

Part B

Maximum Likelihood for the Gaussian Models

B.1 The Joint Likelihood Splits by Class

derive

Rows are drawn independently, so the likelihood is a product, and the generative factorization splits each term:

$$L(\theta) = \prod_{i=1}^{n} p_\theta(x_i, y_i) = \prod_{i=1}^{n} \pi_{y_i}\, \phi_d(x_i; \mu_{y_i}, \Sigma_{y_i}).$$

B.1 The Joint Likelihood Splits by Class

derive

Rows are drawn independently, so the likelihood is a product, and the generative factorization splits each term:

$$L(\theta) = \prod_{i=1}^{n} p_\theta(x_i, y_i) = \prod_{i=1}^{n} \pi_{y_i}\, \phi_d(x_i; \mu_{y_i}, \Sigma_{y_i}).$$

Taking logarithms turns the product into a sum, and collecting the rows of each class together gives

$$\ell(\theta) = \sum_{k=1}^{K} \Big[\, n_k \log \pi_k + \sum_{i:y_i=k} \log \phi_d(x_i; \mu_k, \Sigma_k) \Big].$$

B.1 The Joint Likelihood Splits by Class

derive

Rows are drawn independently, so the likelihood is a product, and the generative factorization splits each term:

$$L(\theta) = \prod_{i=1}^{n} p_\theta(x_i, y_i) = \prod_{i=1}^{n} \pi_{y_i}\, \phi_d(x_i; \mu_{y_i}, \Sigma_{y_i}).$$

Taking logarithms turns the product into a sum, and collecting the rows of each class together gives

$$\ell(\theta) = \sum_{k=1}^{K} \Big[\, n_k \log \pi_k + \sum_{i:y_i=k} \log \phi_d(x_i; \mu_k, \Sigma_k) \Big].$$

No parameter appears in two of these brackets, so each class is fitted from its own rows alone — except for an LDA shared Σ, which is the one term that pools them.

B.2 The Prior and the Means

derive

Only the first term involves π, so maximize it subject to $\sum_k \pi_k = 1$:

$$\max_{\pi} \sum_k n_k \log \pi_k \quad \Longrightarrow \quad \boxed{\widehat{\pi}_k = \frac{n_k}{n}}$$

the class share, as counting would suggest.

B.2 The Prior and the Means

derive

Only the first term involves π, so maximize it subject to $\sum_k \pi_k = 1$:

$$\max_{\pi} \sum_k n_k \log \pi_k \quad \Longrightarrow \quad \boxed{\widehat{\pi}_k = \frac{n_k}{n}}$$

the class share, as counting would suggest.

With Σ_k held fixed, only the exponent of the Gaussian involves μ_k, so maximizing the likelihood minimizes a within-class distance:

$$\min_{\mu_k} \sum_{i:y_i=k} (x_i - \mu_k)^\top \Sigma_k^{-1} (x_i - \mu_k) \quad \Longrightarrow \quad \boxed{\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i}$$

the class mean, whatever Σ_k is.

B.3 The Covariance

derive

Substituting $\hat{\mu}_k$ and maximizing over Σ_k gives the within-class scatter,

$$\widehat{\Sigma}_k = \frac{1}{n_k} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^\top$$

B.3 The Covariance

derive

Substituting $\hat{\mu}_k$ and maximizing over Σ_k gives the within-class scatter,

$$\boxed{\widehat{\Sigma}_k = \frac{1}{n_k} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^\top}$$

LDA assumes a single Σ, so the same maximization runs over all rows at once and produces the pooled estimate

$$\widehat{\Sigma} = \frac{1}{n} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^\top = \frac{1}{n} \sum_{k=1}^{K} n_k \widehat{\Sigma}_k.$$

Naive Bayes keeps only the diagonal of whichever estimate it uses.