Bent discriminant analysis:

a Bayesian nonparametric approach to discriminant analysis

Bernardo Nipoti

Università di Milano-Bicocca

Padova, SISBayes, 4/9/2025

- Joint project with Laura D'Angelo and Tommaso Rigon
- Preliminary results in this presentation

- Support from the Cariplo's grant "Data Science for Health", project "Joint Bayesian Learning of Related Clinical Studies"

Discriminant analysis

- Classical approach to supervised classification
- Very popular thanks to its simplicity

Variants:

- Linear discriminant analysis (LDA) [Fisher, 1936]
- Quadratic discriminant analysis (QDA)
- Various generalizations [see EOSL, Hastie et al, 2009]

"Both LDA and QDA perform well on an amazingly large and diverse set of classification tasks. [...] It seems that whatever exotic tools are the rage of the day, we should always have available these two simple tools." [Hastie et al, 2009]

Setup and notation

- Data: $(\boldsymbol{x}, \boldsymbol{y}) = \{(x_i, y_i) : i = 1, \ldots, n\}$, where:

  predictors: $x_i \in \mathbb{R}^d$

  categorical response: $y_i \in \mathcal{G} = \{1, \ldots, G\}$

- Assumption on the distribution of the predictors x_i:

$$x_i \mid y_i = g, \boldsymbol{\mu}, \boldsymbol{\Sigma} \overset{\text{ind}}{\sim} \mathrm{N}_d(\mu_g, \Sigma_g),$$

  where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_G)$ and $\boldsymbol{\Sigma} = (\Sigma_1, \ldots, \Sigma_G)$

- Bayes classifier: given a predictor x_*, a prediction $\hat{y}(x_*)$ is made based on the posterior probability of the response y_*:

$$\hat{y}(x_*) = \underset{g \in \mathcal{G}}{\operatorname{argmax}} \Pr(y_* = g \mid x_*)$$

Discriminant functions

$$\Pr(y_* = g \mid x_*) \overset{\text{Bayes}}{\propto} \Pr(y_* = g) p(x_* \mid y_* = g)$$
$$= \pi_g f_{\mathrm{N}_d}(x_*; \mu_g, \Sigma_g)$$

π_g: prior probability for category g

f_{N_d}: pdf of a d-dimensional normal

- *E.g.*, two categories (g_1, g_2) with $\pi_1 = \pi_2$: $\hat{y}(x_*) = g_1$ if

$$\log \Pr(y_* = g_1 \mid x_*) > \log \Pr(y_* = g_2 \mid x_*)$$

- The assumption of normality of the predictors leads to a discriminant inequality *quadratic* in x_* (QDA)
- Further assuming that $\Sigma_g = \Sigma$ for every $g \in \mathcal{G}$ leads to a discriminant inequality *linear* in x_* (LDA)

Parameters

The discriminant inequalities involve parameters to be estimated:

LDA: $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_G)$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_G)$, Σ

QDA: $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_G)$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_G)$, $\boldsymbol{\Sigma} = (\Sigma_1, \ldots, \Sigma_G)$

Approach is flexible in the choice of the estimation method:

- Maximum likelihood estimators
- Bayesian posterior estimators
- Other estimators, depending on the focus

LDA, QDA or other variants?

- QDA makes less assumptions than LDA but requires estimating a larger number of parameters
- Issue when the class-specific sizes n_g are small and d is large

Compromises studied in the literature, *e.g.*:

- Regularized discriminant analysis (RDA) [Friedman, 1989]

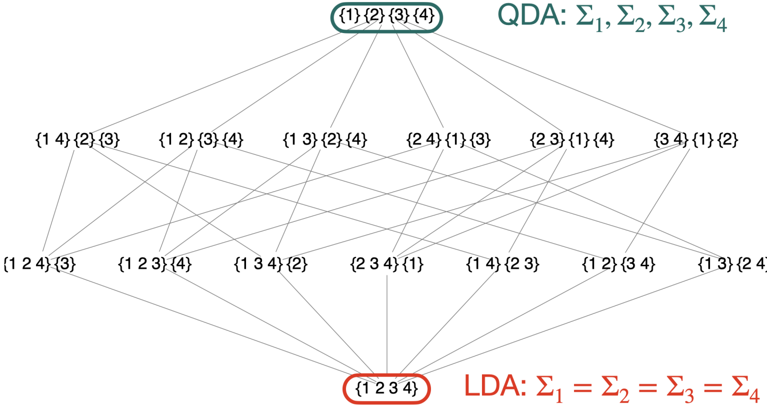
$$\hat{\Sigma}_g(\alpha) = \alpha \hat{\Sigma}_g + (1 - \alpha)\hat{\Sigma}$$

$\hat{\Sigma}_g(\alpha)$ combines the assumptions of LDA and QDA

We instead explore methods that lie in between LDA and QDA

Between LDA and QDA

Example with $G = 4$. Hasse diagram:



- LDA and QDA extreme cases of a rich collection of models

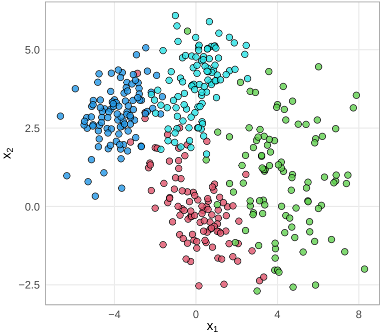
Bent discriminant analysis

- Problem of selecting among B_G (Bell number) models
- $B_G = 15$ when $G = 4$ as in the example
- $\{\text{models}\} \longleftrightarrow \{\text{partitions of } \mathcal{G} = \{1, \ldots, G\}\}$

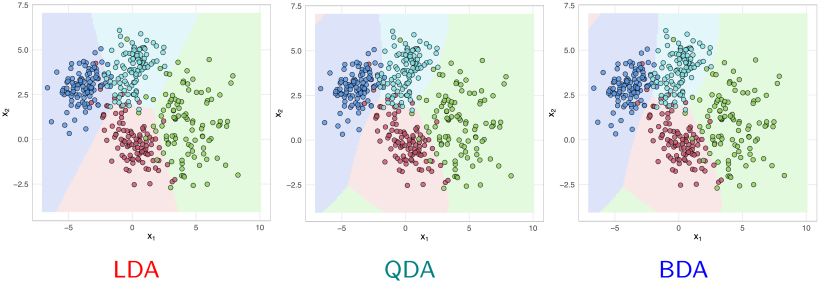
  e.g. $\{1, 2\}, \{3, 4\}$ corresponds to model $\Sigma_1 = \Sigma_2$, $\Sigma_3 = \Sigma_4$
- Idea: assigning positive prior probability to all possible partitions of $\mathcal{G}$
- Our proposal: bent discriminant analysis (BDA)
- LDA and QDA are special cases of BDA

Toy example with $G = 4$ classes

- Simulate $n_g = 100$ points per class from a normal distribution
- Means: $\mu_1 = (0, 0)$, $\mu_2 = (4, 1)$, $\mu_3 = (-4, 3)$, $\mu_4 = (0, 4)$
- Covariances: Σ_1, Σ_2, $\Sigma_3 = \Sigma_4$ (*i.e.* $\{1\}, \{2\}, \{3, 4\}$)



Toy example: classification regions

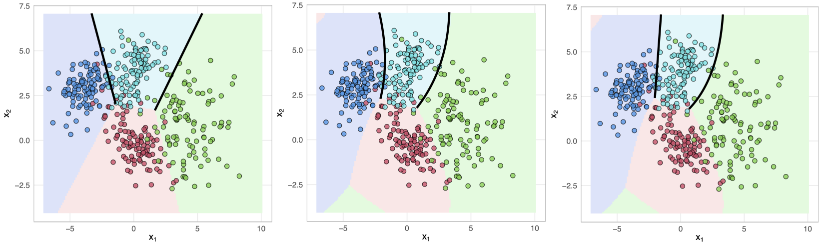


LDA: $\Sigma_1 = \Sigma_2 = \Sigma_3 = \Sigma_4$

QDA: $\Sigma_1, \Sigma_2, \Sigma_3, \Sigma_4$

BDA: selected model: $\{1\}, \{2\}, \{3, 4\}$, i.e. $\Sigma_1, \Sigma_2, \Sigma_3 = \Sigma_4$

Toy example: classification regions



LDA QDA BDA

LDA: $\Sigma_1 = \Sigma_2 = \Sigma_3 = \Sigma_4$

QDA: $\Sigma_1,\ \Sigma_2,\ \Sigma_3,\ \Sigma_4$

BDA: selected model: $\{1\}, \{2\}, \{3, 4\}$, *i.e.* $\Sigma_1,\ \Sigma_2,\ \Sigma_3 = \Sigma_4$

Latent partition model

- BDA exploits a latent partition model over $\mathcal{G} = \{1, \ldots, G\}$
- A random partition $\mathcal{S}$ of $\mathcal{G}$ is implied by a nonparametric mixture with only $\{\Sigma_1, \ldots, \Sigma_G\}$ modeled nonparametrically

Hierarchical model on $\{(\mu_g, \Sigma_g) : g = 1, \ldots, G\}$:

$$\mu_g \mid \Sigma_g \overset{\text{ind}}{\sim} \mathrm{N}_d\left(\mu_{0,g}, \frac{\Sigma_g}{\tau_{0,g}}\right)$$

$$\Sigma_g \mid P \overset{\text{iid}}{\sim} P$$

$$P \sim Q$$

- Q: distribution of discrete nonparametric random measure (*e.g.* DP, Gibbs) on the space of positive-definite matrices
- Inverse-Wishart(Λ_0, ν_0) as base measure of Q

Complete Bayesian model for BDA

$$\Pr(y_* = g \mid x_*) \propto \pi_g f_{\mathsf{N}_d}(x_*; \mu_g, \Sigma_g)$$

We define a Bayesian model with two components:

1. Scale-only mixture model for $\{(\mu_g, \Sigma_g) : g = 1, \ldots, G\}$
2. Prior for prior probabilities

$$(\pi_1, \ldots, \pi_G) \sim \text{Dirichlet}(\beta_1, \ldots, \beta_G)$$

Recall: the EPPF of a Gibbs-type prior Q [De Blasi et al., 2013] with $\sigma \leq 1$ and weights $\{V_{n,k} : n \geq 1, 1 \leq k \leq n\}$ is given by

$$p(\mathcal{S}) = \Pi_k^{(G)}(\tilde{n}_1, \ldots, \tilde{n}_k) = V_{n,k} \prod_{j=1}^{k} (1 - \sigma)_{\tilde{n}_j - 1}$$

with $\mathcal{S}$ partition of $\mathcal{G} = \{1, \ldots, G\}$ with k blocks of size $\tilde{n}_1, \ldots, \tilde{n}_k$

Posterior over the space of partitions (I)

Key for model selection is the posterior distribution of $\mathcal{S}$:

$$p(\mathcal{S} \mid \boldsymbol{x}, \boldsymbol{y}) \propto V_{n,k} \frac{|\Lambda_0|^{k\nu_0/2}}{\Gamma_d(\nu_0/2)^k} \prod_{j=1}^{k} \left\{ (1-\sigma)_{\tilde{n}_j - 1} \frac{\Gamma_d(\nu_{n,j}/2)}{|\Lambda_{n,j}|^{\nu_{n,j}/2}} \right\},$$

obtained after *marginalizing* with respect to parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$

- In classification problems, G is typically of moderate size
- $p(\mathcal{S} \mid \boldsymbol{x}, \boldsymbol{y})$ can be evaluated over the whole space of models

How many evaluations?

G	2	3	4	5	6	7	8	9	10
$\mathbf{B_G}$	2	5	15	52	203	877	4140	21147	115975

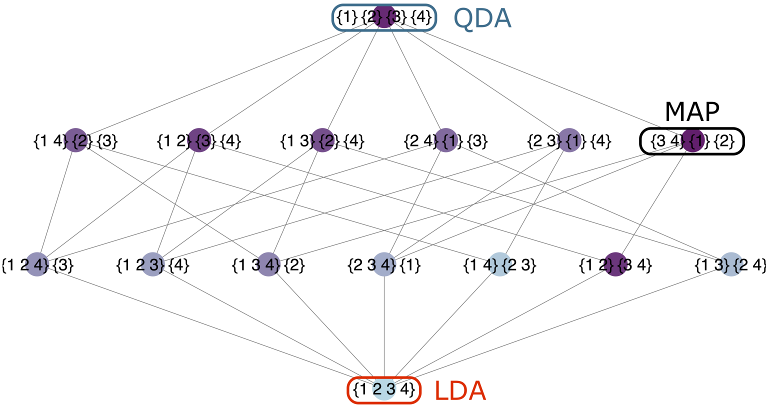
Posterior over the space of partitions (II)

The evaluation of $p(\mathcal{S} \mid \boldsymbol{x}, \boldsymbol{y})$ allows us to:

- Compute the normalizing costant
- Identify the MAP $\hat{\mathcal{S}}$
- Identify a set of likely partitions/models
  [Wade & Ghahramani, 2018; Balocchi & Wade, 2025]
- Sample exactly from $p(\mathcal{S} \mid \boldsymbol{x}, \boldsymbol{y})$
- Evaluate functionals of interest, *e.g.* the posterior distribution of the number of blocks $|\mathcal{S}|$:

$$\Pr(|\mathcal{S}| = k \mid \boldsymbol{X}, \boldsymbol{y}) \propto V_{n,k} \frac{|\Lambda_0|^{k\nu_0/2}}{\Gamma_d(\nu_0/2)^k} \sum_{\mathcal{S}:|\mathcal{S}|=k} \left\{ \prod_{j=1}^{k} (1-\sigma)_{\tilde{n}_j - 1} \frac{\Gamma_d(\nu_{n,j}/2)}{|\Lambda_{n,j}|^{\nu_{n,j}/2}} \right\}$$

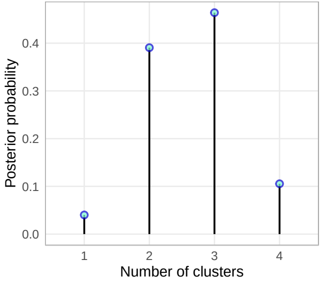
Back to toy example with $G = 4$ classes



purple: high posterior; light blue: low posterior.

Toy example: posterior inference on $\mathcal{S}$

- MAP: $\hat{\mathcal{S}} = \{\{1\},\{2\},\{3,4\}\}$
- Posterior distribution of number of blocks in $\mathcal{S}$:



Classification via conditional posterior predictive

We want to classify a new statistical unit with predictors x_*:

- Conditionally on an estimated partition/model $\hat{\mathcal{S}}$:

$$\Pr(y_* = g \mid x_*, \boldsymbol{x}, \boldsymbol{y}, \hat{\mathcal{S}}) \propto \\ \propto \frac{\beta_g + n_g}{\sum_{h=1}^{G}(\beta_h + n_h)} t_{\nu_{n,j}-d+1}\left(x_*;\ \mu_{j,n}, \frac{\Lambda_{n,j}(\tau_{n,g}+1)}{\tau_{n,g}(\nu_{n,j}-d+1)}\right)$$

- Bayes classifier:

$$\hat{y}(x_*) = \underset{g \in \mathcal{G}}{\operatorname{argmax}} \Pr(y_* = g \mid x_*, \boldsymbol{x}, \boldsymbol{y}, \hat{\mathcal{S}})$$

Classification via marginal posterior predictive

We want to classify a new statistical unit with predictors x_*:

- By marginalizing with respect to $\mathcal{S}$:

$$\Pr(y_* = g \mid x_*, X, y) = \sum_{\mathcal{S} \in \mathcal{P}_{\mathcal{G}}} \Pr(y_* = g \mid x_*, \boldsymbol{x}, \boldsymbol{y}, \mathcal{S}) p(\mathcal{S} \mid x_*, \boldsymbol{x}, \boldsymbol{y})$$

  whose evaluation is possible but computationally intensive

- If G is not small, via Monte Carlo:

$$\mathbb{E}_{\mathcal{S} \mid x_*, \boldsymbol{x}, \boldsymbol{y}}[\Pr(y_* = g \mid x_*, \boldsymbol{x}, \boldsymbol{y}, \mathcal{S})]$$

- Bayes classifier:

$$\hat{y}(x_*) = \operatorname*{argmax}_{g \in \mathcal{G}} \mathbb{E}_{\mathcal{S} \mid x_*, \boldsymbol{x}, \boldsymbol{y}}[\Pr(y_* = g \mid x_*, \boldsymbol{x}, \boldsymbol{y}, \mathcal{S})]$$

Simulation experiment

Synthetic data:

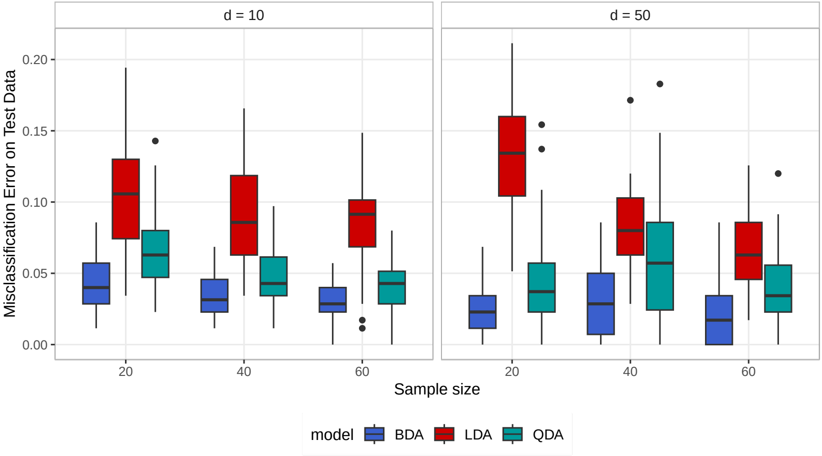
- $d \in \{10, 50\}$
- $G = 7$, with $n_g \in \{20, 40, 60\}$
- $\mathcal{S} = \{\{1, 2, 3\}, \{4, 5, 6\}, \{7\}\}$

  i.e. $\Sigma_1 = \Sigma_2 = \Sigma_3$; $\Sigma_4 = \Sigma_5 = \Sigma_6$; Σ_7
- 50 replicated datasets per scenario

Data analyzed with:

- LDA
- QDA
- BDA (via conditional posterior predictive)

Simulation experiment



What's next?

- Algorithmic approach to find the MAP $\hat{\mathcal{S}}$ when $G > 10$
- BDA's performance on challenging scenarios
- Classification of real data, *e.g.* in the field of clinical studies
- Study the impact of the choice nonparametric prior for P
- Incorporate class-specific covariates by modeling P with a PPMx (product partition model with regression on covariates) [Müller et al., 2011]
- Comments and suggestions are welcome!

Some references

- Balocchi, C., & Wade, S. (2025). Understanding uncertainty in Bayesian cluster analysis. *arXiv preprint*.
- De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I., & Ruggiero, M. (2013). Are Gibbs-type priors the most natural generalization of the Dirichlet process?. *IEEE transactions on pattern analysis and machine intelligence*.
- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*.
- Friedman, J. H. (1989). Regularized discriminant analysis. *Journal of the American statistical association*.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*.
- Müller, P., Quintana, F., & Rosner, G. L. (2011). A product partition model with regression on covariates. *Journal of Computational and Graphical Statistics*.
- Wade, S., & Ghahramani, Z. (2018). Bayesian cluster analysis: Point estimation and credible balls. *Bayesian Analysis*.