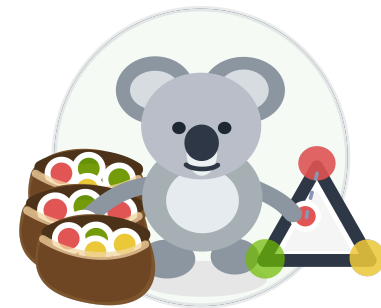
UNIVERSITY OF COPENHAGEN

Funded by the European Union

AIPHY

Multiclass Classification without Labels

Multiclass Classification without Labels via Posterior Simplex Geometry (arXiv:2607.24943)



Date: 14 September 2026

Venue: ML4Jets 2026, Vienna

By: Johann Ioannou-Nikolaides – Niels Bohr Institute, University of Copenhagen

See also Gregorio de la Fuente's talk on *Simplex Demixing: Disentangling Multiple Light-Flavor Jets at Colliders* (arXiv:2607.24921)

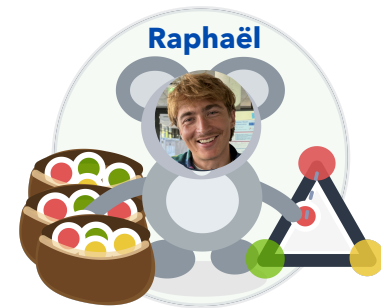
UNIVERSITY OF COPENHAGEN

Funded by the European Union

AIPHY

Multiclass Classification without Labels

Multiclass Classification without Labels via Posterior Simplex Geometry (arXiv:2607.24943)



Date: 14 September 2026

Venue: ML4Jets 2026, Vienna

By: Johann Ioannou-Nikolaides – Niels Bohr Institute, University of Copenhagen

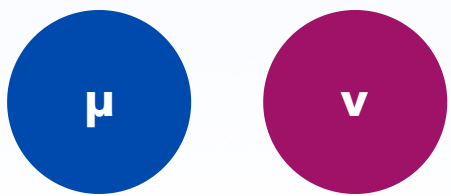
See also Gregorio de la Fuente's talk on *Simplex Demixing: Disentangling Multiple Light-Flavor Jets at Colliders* (arXiv:2607.24921)

conceptual intuition

CWoLa: two mixtures → two classes

Given two differently-enriched, unlabeled mixtures, CWoLa recovers the two underlying classes.

Pull ν events out of an atmospheric-muon-dominated stream – data-driven, no truth labels.

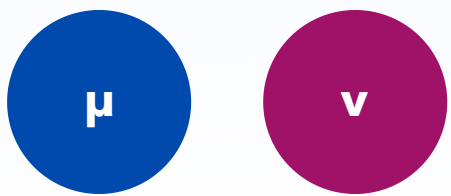


conceptual intuition

CWoLa: two mixtures → two classes

Given two differently-enriched, unlabeled mixtures, CWoLa recovers the two underlying classes.

Pull ν events out of an atmospheric-muon-dominated stream – data-driven, no truth labels.



But why stop at two?

Let's split the neutrino *flavors* too, directly out of the same kind of unlabeled mixtures.

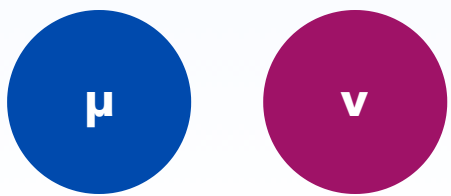
→ Needs more than two classes.

conceptual intuition

CWoLa: two mixtures → two classes

Given two differently-enriched, unlabeled mixtures, CWoLa recovers the two underlying classes.

Pull ν events out of an atmospheric-muon-dominated stream – data-driven, no truth labels.



But why stop at two?

Let's split the neutrino *flavors* too, directly out of the same kind of unlabeled mixtures.

→ Needs more than two classes.

CWoLa already solves the left problem. This talk is about solving the right one.

CWoLa recap

Two mixtures of the same two classes, different (unknown) proportions:

$$M_1 = f_1\, S + (1 - f_1)\, B, \qquad M_2 = f_2\, S + (1 - f_2)\, B, \qquad f_1 \neq f_2$$

The Bayes-optimal classifier for M_1 vs. M_2 is a monotonic function of $p_{M_1}(x)/p_{M_2}(x)$.

$$\frac{p_{M_1}(x)}{p_{M_2}(x)} = \frac{f_1\, p_S(x) + (1 - f_1)\, p_B(x)}{f_2\, p_S(x) + (1 - f_2)\, p_B(x)} = \frac{f_1\, p_S(x)/p_B(x) + (1 - f_1)}{f_2\, p_S(x)/p_B(x) + (1 - f_2)}$$

CWoLa recap

Two mixtures of the same two classes, different (unknown) proportions:

$$M_1 = f_1\, S + (1 - f_1)\, B, \qquad M_2 = f_2\, S + (1 - f_2)\, B, \qquad f_1 \neq f_2$$

The Bayes-optimal classifier for M_1 vs. M_2 is a monotonic function of $p_{M_1}(x)/p_{M_2}(x)$.

$$\frac{p_{M_1}(x)}{p_{M_2}(x)} = \frac{f_1\, p_S(x) + (1 - f_1)\, p_B(x)}{f_2\, p_S(x) + (1 - f_2)\, p_B(x)} = \frac{f_1\, p_S(x)/p_B(x) + (1 - f_1)}{f_2\, p_S(x)/p_B(x) + (1 - f_2)}$$

which is **monotonically increasing in** p_S/p_B whenever $f_1 > f_2$ and $f_1(1-f_2) - f_2(1-f_1) \neq 0$.

CWoLa recap

Two mixtures of the same two classes, different (unknown) proportions:

$$M_1 = f_1\, S + (1 - f_1)\, B, \qquad M_2 = f_2\, S + (1 - f_2)\, B, \qquad f_1 \neq f_2$$

The Bayes-optimal classifier for M_1 vs. M_2 is a monotonic function of $p_{M_1}(x)/p_{M_2}(x)$.

$$\frac{p_{M_1}(x)}{p_{M_2}(x)} = \frac{f_1\, p_S(x) + (1 - f_1)\, p_B(x)}{f_2\, p_S(x) + (1 - f_2)\, p_B(x)} = \frac{f_1\, p_S(x)/p_B(x) + (1 - f_1)}{f_2\, p_S(x)/p_B(x) + (1 - f_2)}$$

which is **monotonically increasing in** p_S/p_B whenever $f_1 > f_2$ and $f_1(1-f_2) - f_2(1-f_1) \neq 0$.

So the mixture classifier and the signal/background classifier rank every event identically. No labels needed – only two differently-enriched, unlabeled mixtures. (Metodiev, Nachman & Thaler, arXiv:1708.02949)

from likelihood ratio to 1D geometry

Define mixture posterior $g^\star(x) = P(\text{mixture} = 1 \mid x)$ trained with binary cross-entropy and substitute the mixture densities:

$$g^\star(x) = \frac{p_{M_1}(x)}{p_{M_1}(x) + p_{M_2}(x)} = \frac{f_1\, p_S(x) + (1 - f_1)\, p_B(x)}{\underbrace{(f_1 + f_2)}_{c_S}\, p_S(x) \;+\; \underbrace{(1 - f_1) + (1 - f_2)}_{c_B}\, p_B(x)}$$

$r/(1 + r)$ results in the same ordering as r, but bounded to $[0, 1]$. Here c_S is the signal abundance in the mixtures

from likelihood ratio to 1D geometry

Define mixture posterior $g^\star(x) = P(\mathbf{mixture} = 1 \mid x)$ trained with binary cross-entropy and substitute the mixture densities:

$$g^\star(x) = \frac{p_{M_1}(x)}{p_{M_1}(x) + p_{M_2}(x)} = \frac{f_1\, p_S(x) + (1 - f_1)\, p_B(x)}{\underbrace{(f_1 + f_2)}_{c_S}\, p_S(x) \;+\; \underbrace{(1 - f_1) + (1 - f_2)}_{c_B}\, p_B(x)}$$

$r/(1 + r)$ results in the same ordering as r, but bounded to $[0, 1]$. Here c_S is the signal abundance in the mixtures

$$\alpha(x) := \frac{c_S\, p_S(x)}{c_S\, p_S(x) + c_B\, p_B(x)}$$

from likelihood ratio to 1D geometry

Define mixture posterior $g^\star(x) = P(\text{mixture} = 1 \mid x)$ trained with binary cross-entropy and substitute the mixture densities:

$$g^\star(x) = \frac{p_{M_1}(x)}{p_{M_1}(x) + p_{M_2}(x)} = \frac{f_1\, p_S(x) + (1 - f_1)\, p_B(x)}{\underbrace{(f_1 + f_2)}_{c_S}\, p_S(x) \;+\; \underbrace{(1 - f_1) + (1 - f_2)}_{c_B}\, p_B(x)}$$

$r/(1+r)$ results in the same ordering as r, but bounded to $[0, 1]$. Here c_S is the signal abundance in the mixtures

$$\alpha(x) := \frac{c_S\, p_S(x)}{c_S\, p_S(x) + c_B\, p_B(x)}$$

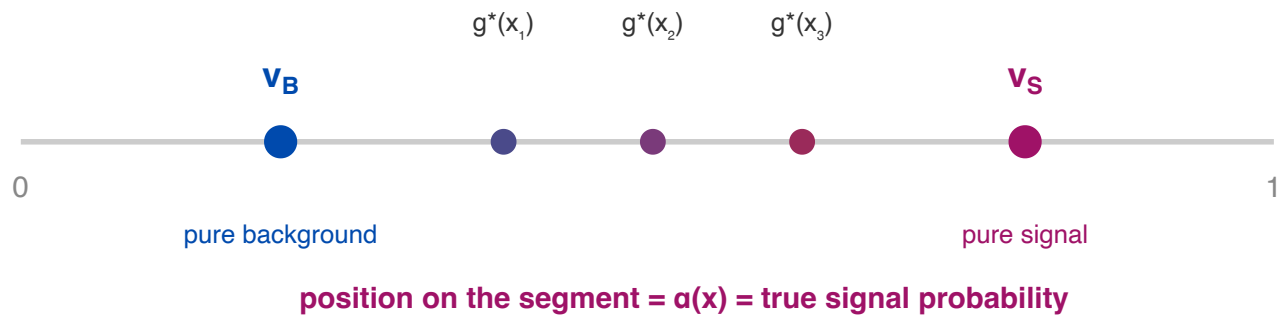
$$g^\star(x) = \alpha(x) \cdot \underbrace{\tfrac{f_1}{c_S}}_{v_S} \;+\; (1 - \alpha(x)) \cdot \underbrace{\tfrac{1-f_1}{c_B}}_{v_B}$$

v_S, v_B are built only from the (unknown, but fixed) mixing fractions – independent of x.

2-mixture posterior is confined to a 1-simplex

$g^\star(x)$ is confined to a straight line between two fixed points, and where you land on it is exactly $\alpha(x)$, the true signal probability:

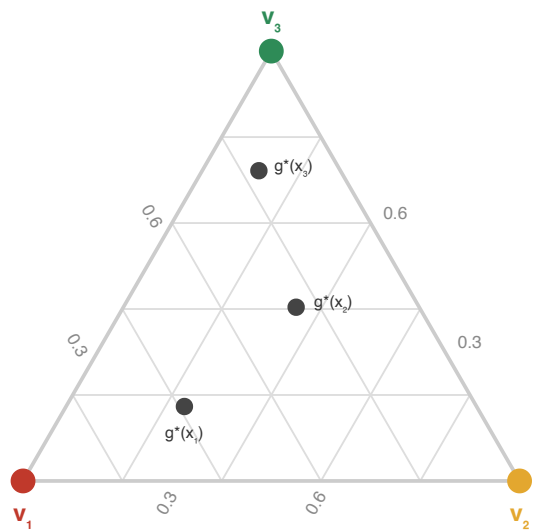
$$g^\star(x) = \alpha(x)\, v_S + (1 - \alpha(x))\, v_B$$



3-mixture posterior latent geometry

For $K = 3$ classes and $M = 3$ mixtures, $g^\star(x)$ is confined to the triangle spanned by three fixed points

$$g^\star(x) = \alpha_1(x)\, v_1 + \alpha_2(x)\, v_2 + \alpha_3(x)\, v_3, \qquad \sum_i \alpha_i = 1$$



The posterior geometry is constrained by the latent class geometry $h^\star(x) = (\alpha_1(x), \alpha_2(x), \alpha_3(x))$ and the mixture composition $\Pi = (v_1/c_1, v_2/c_2, v_3/c_3)$.

mixture identity is sufficient supervision

mixture identity is sufficient supervision

Training on mixture identity with $M \geq K$ mixtures, g^* lives on Δ^{M-1}, but is confined to Δ^{K-1}.

mixture identity is sufficient supervision

Training on mixture identity with $M \geq K$ mixtures, g^* lives on Δ^{M-1}, but is confined to Δ^{K-1}.

The K vertices $v_1, \ldots, v_K$ are **fixed** by mixture composition and affinely independent.

Recovering the true class posterior from the mixture posterior means inverting a fixed linear map:

$$\bar{h}^\star(x) = (V^\top V)^{-1} V^\top \, g^\star(x)$$

with $V = [v_1, ..., v_K]$.

mixture identity is sufficient supervision

Training on mixture identity with $M \geq K$ mixtures, g^* lives on Δ^{M-1}, but is confined to Δ^{K-1}.

The K vertices $v_1, \ldots, v_K$ are **fixed** by mixture composition and affinely independent.

Recovering the true class posterior from the mixture posterior means inverting a fixed linear map:

$$\bar{h}^\star(x) = (V^\top V)^{-1} V^\top \, g^\star(x)$$

with $V = [v_1, ..., v_K]$.

g*(x) is a sufficient statistic for the latent classes. Training on mixture identity and training on class labels are *the same problem*. The M-way mixture classifier is exactly as Bayes-optimal as a supervised K-class classifier would be.

recovering the simplex

The theorem guarantees that $\hat{V}$ exists. We propose two label-free ways to find it:

R1 — Post-hoc vertex hunting

Train an ordinary, unconstrained M-way classifier $g_\theta(x)$. Fit a vertex-hunting algorithm to the resulting posterior cloud $\{g_\theta(x_i)\}$ to estimate $\hat{V}$, then decode each point by its barycentric coordinates against $\hat{V}$.

Trying different fitters is **cheap** and requires no retraining.

recovering the simplex

The theorem guarantees that $\hat{V}$ exists. We propose two label-free ways to find it:

R1 – Post-hoc vertex hunting

Train an ordinary, unconstrained M-way classifier $g_\theta(x)$. Fit a vertex-hunting algorithm to the resulting posterior cloud $\{g_\theta(x_i)\}$ to estimate $\hat{V}$, then decode each point by its barycentric coordinates against $\hat{V}$.

Trying different fitters is **cheap** and requires no retraining.

R2 – Architectural bottleneck

Bake the factorisation into the network itself: $g_\theta(x) = \hat{V}\,\alpha_\theta(x)$, with $\hat{V}$ column-stochastic and $\alpha_\theta(x) \in \Delta^{K-1}$. Train end-to-end; the columns of $\hat{V}$ *are* the learned vertices.

Strong inductive bias helps with weak backbones.

recovering the simplex

The theorem guarantees that $\hat{V}$ exists. We propose two label-free ways to find it:

R1 – Post-hoc vertex hunting

Train an ordinary, unconstrained M-way classifier $g_\theta(x)$. Fit a vertex-hunting algorithm to the resulting posterior cloud $\{g_\theta(x_i)\}$ to estimate $\hat{V}$, then decode each point by its barycentric coordinates against $\hat{V}$.

Trying different fitters is **cheap** and requires no retraining.

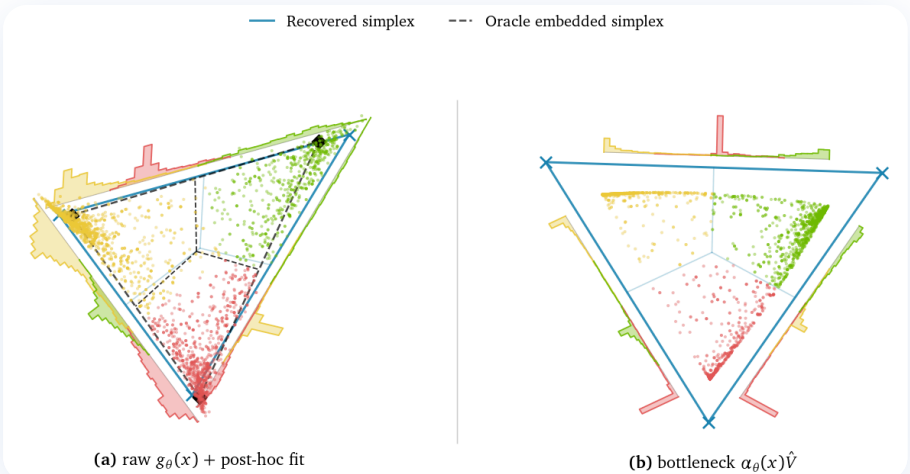
R2 – Architectural bottleneck

Bake the factorisation into the network itself: $g_\theta(x) = \hat{V}\,\alpha_\theta(x)$, with $\hat{V}$ column-stochastic and $\alpha_\theta(x) \in \Delta^{K-1}$. Train end-to-end; the columns of $\hat{V}$ *are* the learned vertices.

Strong inductive bias helps with weak backbones.

Simplex recovery methods can be exchanged → engineering challenge.

latent classes in practice

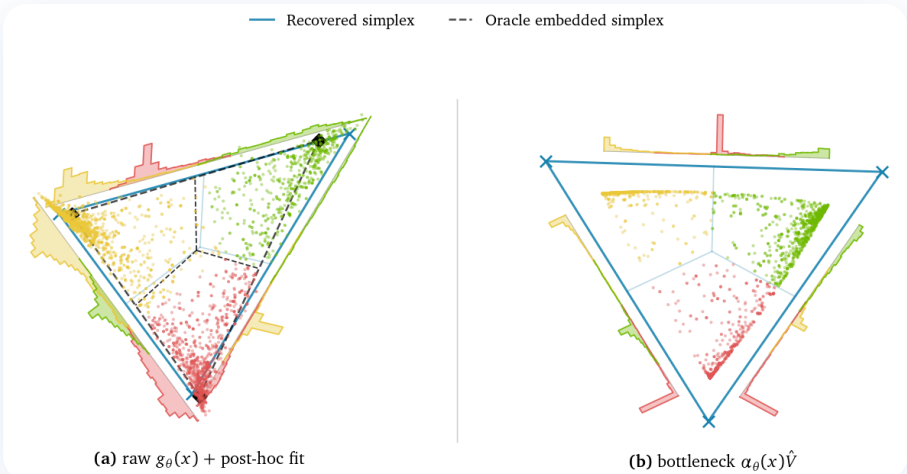


(a) raw $g_\theta(x)$ + post-hoc fit

(b) bottleneck $\alpha_\theta(x)\hat{V}$

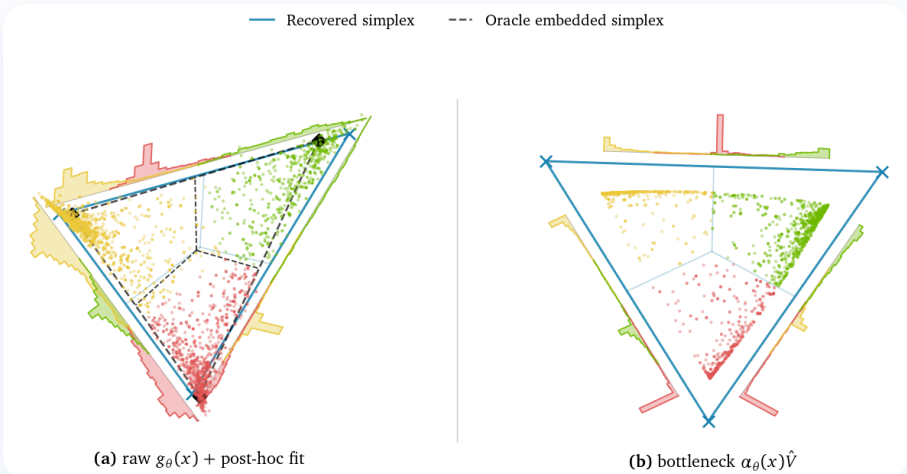
latent classes in practice

The unconstrained mixture classifier can easily overtrain and find differences in mixtures beyond class → early stopping on held-out mixture NLL.



latent classes in practice

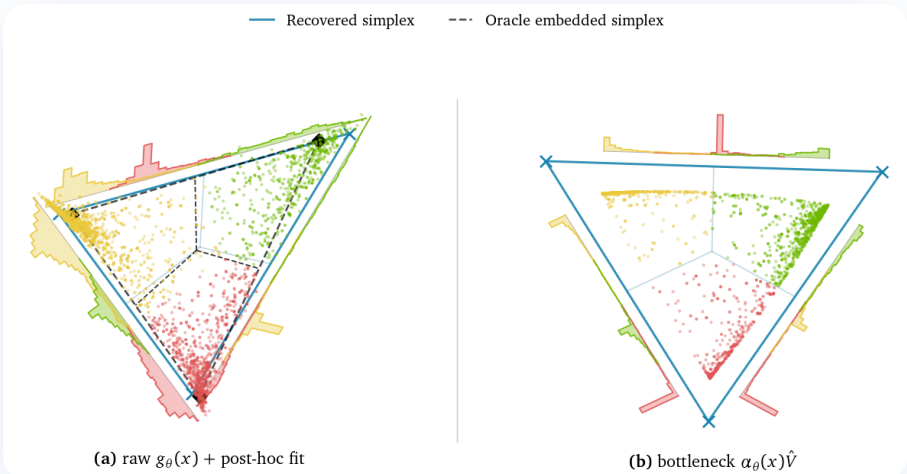
The unconstrained mixture classifier can easily overtrain and find differences in mixtures beyond class → early stopping on held-out mixture NLL.



Identifiability comes from variation across mixtures → **a latent class is whatever varies independently across the mixtures**.

latent classes in practice

The unconstrained mixture classifier can easily overtrain and find differences in mixtures beyond class → early stopping on held-out mixture NLL.



Identifiability comes from variation across mixtures → **a latent class is whatever varies independently across the mixtures**.

What is identifiable depends on three assumptions.

assumptions and ways to break them

A1: shared class-conditionals

$p_k(x)$ must not depend on the mixture m.

Broken: recovered "classes" can be mixture artefacts.

assumptions and ways to break them

A1: shared class-conditionals

$p_k(x)$ must not depend on the mixture m.

Broken: recovered "classes" can be mixture artefacts.

A2: full rank

$\mathrm{rank}(\Pi) = K$, so $M \geq K$.

Broken: the simplex collapses below $K - 1$ dimensions. At most $\mathrm{rank}(\Pi)$ recoverable latent classes.

assumptions and ways to break them

A1: shared class-conditionals

$p_k(x)$ must not depend on the mixture m.

Broken: recovered "classes" can be mixture artefacts.

A2: full rank

$\mathrm{rank}(\Pi) = K$, so $M \geq K$.

Broken: the simplex collapses below $K - 1$ dimensions. At most $\mathrm{rank}(\Pi)$ recoverable latent classes.

A3: separability

Classes need to be sufficiently scattered to reach the boundary of Δ^{K-1}.

Broken: the data never reach the vertices; $h^\star$ is not identifiable.

assumptions and ways to break them

A1: shared class-conditionals

$p_k(x)$ must not depend on the mixture m.

Broken: recovered "classes" can be mixture artefacts.

A2: full rank

$\mathrm{rank}(\Pi) = K$, so $M \geq K$.

Broken: the simplex collapses below $K - 1$ dimensions. At most $\mathrm{rank}(\Pi)$ recoverable latent classes.

A3: separability

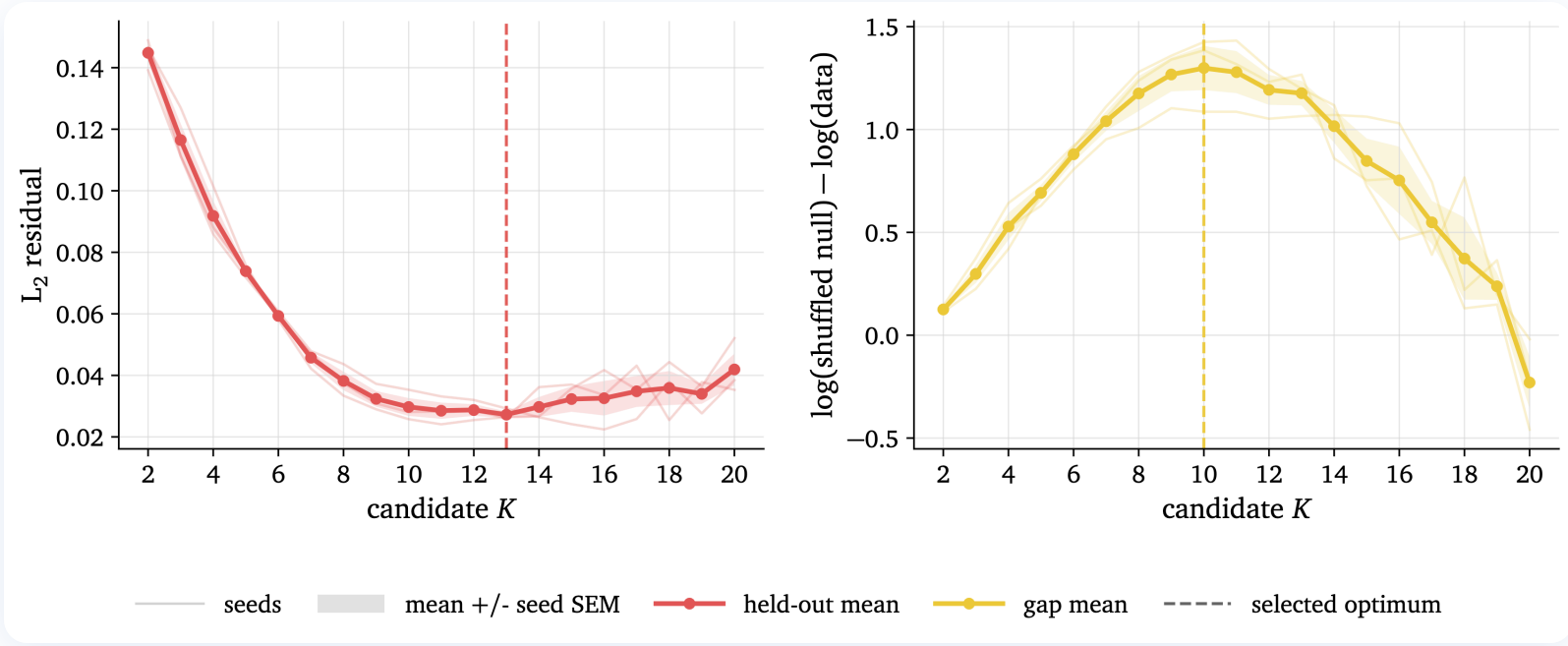
Classes need to be sufficiently scattered to reach the boundary of Δ^{K-1}.

Broken: the data never reach the vertices; $h^\star$ is not identifiable.

When (multiclass) CWoLa is the right method: Labels don't exist and mixtures are too noisy, so majority doesn't solve the task.

detection of the number of latent classes

Reconstruction error $\mathscr{L}_2$ yields a lower error beyond K_{true} because it can fit on noise.

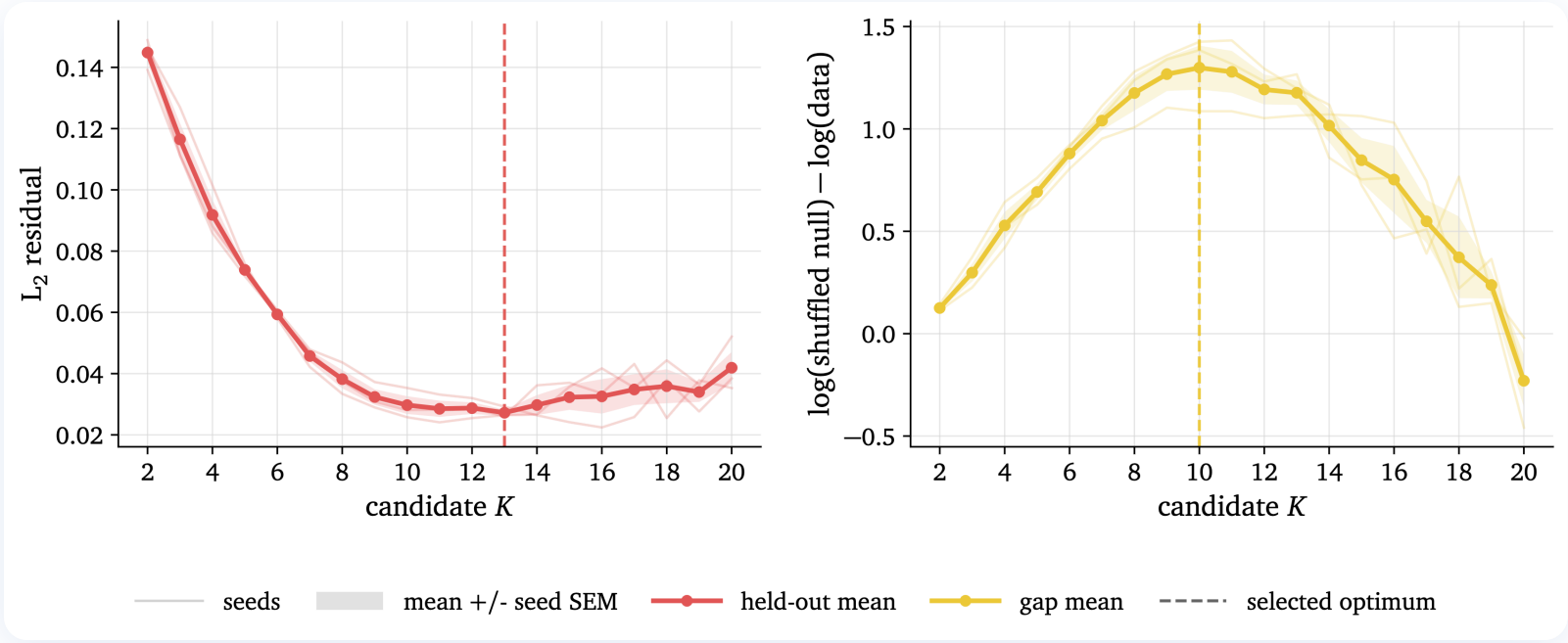


detection of the number of latent classes

Reconstruction error $\mathscr{L}_2$ yields a lower error beyond K_{true} because it can fit on noise.

Metric: Compare $\mathscr{L}_2$ to the error after shuffling each column of the posterior matrix independently across data points.

→ Shuffling destroys the joint geometric correlations while preserving the marginal distributions of individual coordinates.



Thank You

Multiclass Classification without Labels via Posterior Simplex Geometry

Johann Ioannou-Nikolaides

Niels Bohr Institute, University of Copenhagen

johann.nikolaides@nbi.ku.dk

arXiv:2607.24943

Assumptions, details, and all the numbers

Questions? Ideas? Natural mixtures you would like to try?

we're happy to collaborate on applications!