A variational framework for modal estimation

Tâm Le Minh[1], Julyan Arbel[1] and Florence Forbes[1]

[1] *Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France*

Abstract. We approach multivariate mode estimation through Gibbs distributions and introduce GERVE (Gibbs-measure Entropy-Regularized Variational Estimation), a likelihood-free framework that approximates Gibbs measures directly from samples by maximizing an entropy-regularized variational objective with natural-gradient updates. GERVE brings together kernel density estimation, mean-shift, variational inference, and annealing in a single platform for mode estimation. It fits Gaussian mixtures that concentrate on high-density regions and yields cluster assignments from responsibilities, with reduced sensitivity to the chosen number of components. We provide theory in two regimes: as the Gibbs temperature approaches zero, mixture components converge to population modes; at fixed temperature, maximizers of the empirical objective exist, are consistent, and are asymptotically normal. We also propose a bootstrap procedure for per-mode confidence ellipses and stability scores. Simulation and real-data studies show accurate mode recovery and emergent clustering, robust to mixture overspecification. GERVE is a practical likelihood-free approach when the number of modes or groups is unknown and full density estimation is impractical.

Keywords. Mode estimation; Modal clustering; Variational methods; Gibbs measures; Annealing; Bootstrap.

1 Introduction

Understanding the structure of an unknown probability distribution p on $\mathbb{R}^d$, observed through samples, is a central task in statistics. Modes, the local maxima of p, reveal data high-density regions. Detecting multiple modes exposes heterogeneity, supports clustering and segmentation, and offers alternative explanations in inverse problems (Carreira-Perpiñan, 2006; Chacón, 2015). Recently, it has also emerged as a preliminary requirement for new efficient sampling methods of multimodal high-dimensional distributions (Noble et al., 2025).

Classical approaches to mode estimation are commonly grouped into three main families. The first comprises geometric methods that detect high-density regions from the

sample (Chernoff, 1964; Dalenius, 1965; Venter, 1967; Sager, 1978). These methods are simple and robust but tend to degrade in high dimensions. The second fits a smooth density and then locates its modes (Parzen, 1962; Loftsgaarden and Quesenberry, 1965; Kronmal and Tarter, 1968; Samworth, 2018; Liu and Ghosh, 2020). This route is principled but suffers from smoothing bias and high computational cost. The third consists of mean-shift algorithms that ascend the gradient of a kernel density estimate (KDE) without requiring fitting a global model (Fukunaga and Hostetler, 1975; Comaniciu and Meer, 2002; Carreira-Perpiñan, 2007; Arias-Castro et al., 2016). These methods are efficient but bandwidth sensitive and can miss well-separated modes. To our knowledge, no existing method simultaneously targets multiple density modes, provides valid uncertainty quantification, and scales with dimension via simple algorithmic updates, without requiring a full density estimate (Genovese et al., 2016; Ameijeiras-Alonso et al., 2021).

We propose a different perspective: we recast the task of mode estimation into that of approximating a Gibbs measure from samples. Let $f : \mathcal{S} \to \mathbb{R}$ be a measurable function and fix a temperature parameter $\omega > 0$. Define

$$g_\omega(d\boldsymbol{x}) = \frac{\exp(f(\boldsymbol{x})/\omega)}{Z_\omega(f)} d\boldsymbol{x}, \qquad \text{with } Z_\omega(f) := \int_{\mathcal{S}} \exp(f(\boldsymbol{y})/\omega) d\boldsymbol{y},$$

and assume $Z_\omega(f) < \infty$. We assume that $\mathcal{S} \subset \mathbb{R}^d$ has finite Lebesgue measure. The Gibbs measure g_ω preserves the modes of f and is the maximizer, among the unconstrained family of distributions, of the entropy-regularized variational objective (Donsker and Varadhan, 1976):

$$\mathcal{L}_\omega(q) = \mathbb{E}_q\left[f(\boldsymbol{X})\right] + \omega \mathcal{H}_{\mathcal{S}}(q), \qquad \text{with } \mathcal{H}_{\mathcal{S}}(q) = -\int_{\mathcal{S}} q(\boldsymbol{y}) \log q(\boldsymbol{y}) d\boldsymbol{y}. \tag{1}$$

We introduce **GERVE** (*Gibbs-measure Entropy-Regularized Variational Estimation*), a likelihood-free variational framework that approximates the Gibbs measure associated with a density $f := p$, in a Gaussian-mixture family from i.i.d. samples $\boldsymbol{X}_{1:N} \sim p$, where p is nonnegative, supported on a subset of $\mathcal{S}$. Since p is a density, but not analytically available, we leverage the identity $\mathbb{E}_q\left[p(\boldsymbol{X})\right] = \mathbb{E}_p\left[q(\boldsymbol{X})\right]$ to target, with natural-gradient updates, the maximization of an empirical version of $\mathcal{L}_\omega$ averaging the variational density q over the observed samples:

$$\widehat{\mathcal{L}}_{\omega,N}(q) = \frac{1}{N} \sum_{i=1}^{N} q(\boldsymbol{X}_i) + \omega \mathcal{H}_{\mathcal{S}}(q). \tag{2}$$

This estimation problem requires fundamentally different treatment from optimization settings where p (or f) can be evaluated: we develop asymptotic theory for maximizers, establish consistency and asymptotic normality, and introduce bootstrap-based inference for per-mode uncertainty quantification, extending recent computational techniques from

variational optimization (Khan and Rue, 2023; Le Minh et al., 2025) to a statistical estimation methodology.

By fitting Gaussian mixtures to the Gibbs density, GERVE unifies ideas from mean-shift, kernel density estimation, entropy annealing, and variational inference within one algorithmic paradigm. The variational formulation is well-suited to multimodal and multivariate settings, where geometric or purely nonparametric mode estimators often struggle. The entropy term induces repulsion between mixture components, enabling exploration of the modal landscape, while annealing ($\omega \to 0$) concentrates mass at the global modes. This allows overcomplete mixtures to automatically adapt to the effective number of modes without prior specification.

Our main contributions are:

1. *Variational formulation for mode estimation.* We cast mode estimation as Gibbs-measure approximation from samples, enabling mode finding and uncertainty quantification when the number of modes is unknown and density estimation is impractical.
2. *Asymptotic theory for multimodal recovery.* We establish two complementary theoretical regimes: As $\omega \to 0$, the variational mixture concentrates on the population modes. At fixed $\omega > 0$, empirical maximizers are consistent and asymptotically normal.
3. *Bootstrap uncertainty for modes.* We provide per-mode confidence ellipses and stability scores via a matched-mode bootstrap with theoretical guarantees.
4. *Algorithmic framework.* Natural-gradient updates with annealing recover multiple modes, estimate local curvature, and yield cluster assignments without pre-specifying the number of groups. A fixed-covariance special case recovers the mean-shift algorithm, while learned covariances adapt to anisotropy and reduce bandwidth sensitivity.
5. *Evidence on simulations and data.* We show accurate mode recovery and competitive clustering, robustness to mild overcompleteness, and a hotspot analysis with uncertainty on UK road-collision data.

Our approach is related to modal clustering (Menardi, 2016; Chacón, 2020), which defines clusters as basins of attraction of density modes. GERVE uses variational approximation, yielding clusters that locally agree with basins near modes under mild conditions (Suppl. S1.3), while providing computational efficiency and natural uncertainty quantification.

Outline. The remainder is organized as follows. Section 2 develops the variational foundations of Gibbs-measure approximation and identifies the behavior of Gaussian mixtures approximations under annealing (Thm. 2.1). Section 3 introduces the main principles of GERVE and establishes the asymptotics of the empirical maximizers (Thm. 3.1). Section 4 presents algorithmic details (Alg. 1). Section 5 gives the fixed-temperature bootstrap procedure (Alg. 2) with guarantees (Thm. 5.1 and 5.3). Sections 6 and 7 report simulations and a case study on UK road-collision hotspots. Section 8 discusses limitations and future

directions. All proofs and additional details are in Supplementary Material. A Python implementation for GERVE can be found at `github.com/tam-leminh/gerve`.

Notations. We consider that p (or f) has a finite number I of global modes $\{\boldsymbol{x}_i^\star\}_{i=1}^I$. When $f : \mathcal{S} \to \mathbb{R}$ and there is no ambiguity on the domain, f is $\mathcal{C}^k$ means $f \in \mathcal{C}^k(\mathcal{S})$, that is f is continuously differentiable on $\mathcal{S}$ to the order k. For a distribution q on $\mathcal{A} \subseteq \mathbb{R}^d$ and a set $\mathcal{B} \subset \mathcal{A}$, we write $q(\mathcal{B}) := \int_{\mathcal{B}} q(\boldsymbol{x})d\boldsymbol{x}$ and $q^{\mathcal{B}} := q\mathbb{1}_{\mathcal{B}}/q(\mathcal{B})$, its truncation (or conditional law) on $\mathcal{B}$. The set $\{q_{\boldsymbol{\vartheta}}, \boldsymbol{\vartheta} \in \Theta\}$ denotes a variational family of densities parameterized by $\boldsymbol{\vartheta}$ where $\boldsymbol{\vartheta} = \boldsymbol{\lambda}$ for single Gaussians $q_{\boldsymbol{\lambda}} = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, or $\boldsymbol{\vartheta} = \boldsymbol{\Lambda}$ for Gaussian mixtures $q_{\boldsymbol{\Lambda}} = \sum_{k=1}^K \pi_k \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. When there is no ambiguity, we write for the objective $\mathcal{L}_\omega(\boldsymbol{\vartheta}) := \mathcal{L}_\omega(q_{\boldsymbol{\vartheta}})$.

2 Motivation for Gibbs approximation: Annealing behavior of Gaussian mixtures

For a measurable function $f : \mathcal{S} \to \mathbb{R}$, the associated Gibbs measure g_ω preserves its modes. The classical variational identity

$$\mathcal{L}_\omega(q) = \omega \log Z_\omega(f) - \omega \mathrm{KL}\left(q \| g_\omega\right),$$

with $\mathrm{KL}\left(q\|g_\omega\right) := \int_{\mathcal{S}} q(\boldsymbol{x}) \log(q(\boldsymbol{x})/g_\omega(\boldsymbol{x}))d\boldsymbol{x}$, implies that g_ω uniquely maximizes the objective $\mathcal{L}_\omega$ when $Z_\omega(f) < \infty$ (Donsker and Varadhan, 1976). Therefore, restricting the search to a variational family gives an approximation for g_ω by minimization of $\mathrm{KL}\left(q\|g_\omega\right)$.

We study how Gaussian mixtures approximate g_ω when ω decreases to 0. Since g_ω is defined only on $\mathcal{S}$, admissible variational distributions are required to have support in $\mathcal{S}$. We therefore work with mixtures of truncated Gaussians. The next result, proved in Supplementary Material S2.2, describes how optimal mixtures concentrate on the global modes when the number of components K is at least the number of global modes I.

Theorem 2.1 (Gaussian mixture concentration on global modes)**.** *Assume $f \in \mathcal{C}^3(\mathcal{S})$ and bounded on $\mathcal{S}$, and for each mode in $\{\boldsymbol{x}_i^\star\}_{i=1}^I$, $\boldsymbol{x}_i^\star \in \mathrm{int}(\mathcal{S})$ and $\boldsymbol{H}_i := -\nabla^2 f(\boldsymbol{x}_i^\star) \succ 0$. Consider $\mathcal{Q}$ to be the variational family of Gaussian mixtures truncated on $\mathcal{S}$, with a fixed number of components $K \geq I$ and upper-bounded covariances. Let $q_\omega^\star$ denote any maximizer of $\mathcal{L}_\omega$ in $\mathcal{Q}$. Choose disjoint neighborhoods $U_1, \ldots, U_I$ of the modes and let $U_0 := S \backslash \bigcup_{i=1}^I U_i$.*

Then, as $\omega \to 0$:

(i) We have $\mathrm{KL}(q_\omega^\star \| g_\omega) \to 0$. In particular, $q_\omega^\star(U_0) \to 0$, and, for each neighborhood U_i, $i = 1, \ldots, I$,

$$q_\omega^\star(U_i) \to c_i, \qquad \text{where } c_i \propto (\det \boldsymbol{H}_i)^{-1/2}, \sum_{i=1}^I c_i = 1.$$

(ii) For all $i = 1, \dots, I$, the conditional laws of $q_\omega^\star$ and g_ω on U_i satisfy $\mathrm{KL}\big(q_\omega^{\star,U_i} \| g_\omega^{U_i}\big) \to 0$. *In particular, the mean and covariance of $q_\omega^{\star,U_i}$ satisfy*

$$\mathbb{E}_{q_\omega^{\star,U_i}}[\boldsymbol{X}] = \boldsymbol{x}_i^\star + o(1), \qquad \mathrm{Cov}_{q_\omega^{\star,U_i}}(\boldsymbol{X}) = o(1).$$

This result shows that Gaussian mixtures provide accurate approximations of the Gibbs measure and place their mass on all global modes of g_ω, thus of f, under annealing whenever $K \geq I$. The local shape of g_ω at a particular mode $\boldsymbol{x}_i^\star$ has been studied in Proposition 3 of Le Minh et al. (2025). There, the optimal Gaussian approximation in a neighborhood of $\boldsymbol{x}_i^\star$ is of the form $\mathcal{N}(\boldsymbol{\mu}_\omega, \boldsymbol{\Sigma}_\omega)$ satisfying:

$$\boldsymbol{\mu}_\omega = \boldsymbol{x}_i^\star + o(\sqrt{\omega}), \qquad \boldsymbol{\Sigma}_\omega = \omega \boldsymbol{H}_i^{-1} + o(\omega).$$

These expansions clarify how Gibbs measures can be used for mode seeking. As ω decreases, mixture components sharpen and their means drift toward the global modes of f. The temperature parameter governs the trade-off between exploration (through smooth and diffuse mixtures at large ω) and exploitation (concentrated components at small ω). A decreasing ω schedule allows starting in an exploration regime and gradually refining mode localization. To extend this mechanism to mode estimation from data, we next introduce a statistical formulation in which f is replaced by information available from samples.

3 Statistical framework and asymptotic properties

3.1 Statistical formulation of GERVE

We observe i.i.d. samples $\boldsymbol{X}_{1:N} \sim p$ supported on a bounded domain $\mathrm{supp}(p) := \{\boldsymbol{x} \in \mathbb{R}^d : p(\boldsymbol{x}) > 0\}$ and we choose $\mathcal{S}$ with finite Lebesgue measure such that $\mathrm{supp}(p) \subset \mathcal{S}$. Replacing f with p, we distinguish the *population* and *empirical* objectives respectively referring to $\mathcal{L}_\omega(q)$ and $\widehat{\mathcal{L}}_{\omega,N}(q)$ defined by (1), with $f := p$, and (2).

Instead of performing optimization on truncated distributions on $\mathcal{S}$, GERVE maximizes $\widehat{\mathcal{L}}_{\omega,N}$ over untruncated Gaussian mixture families on $\mathbb{R}^d$, while all integrals in the objective are kept over $\mathcal{S}$. Supplementary Material S1.2 discusses that if $\mathcal{S}$ is taken large enough, then the objective gap between the truncated and untruncated distributions is uniformly small, and the truncated and untruncated maximizers coincide asymptotically under the M-estimation theory (van der Vaart, 1998). This simplifies optimization, as it removes parameter-dependent normalizers in the natural-gradient updates (details in Section 4.1).

For a single Gaussian $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, we consider the compact parameter set

$$\Theta := \Big\{ \boldsymbol{\lambda} = (\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, -\frac{1}{2}\boldsymbol{\Sigma}^{-1}) : \|\boldsymbol{\mu}\|_\infty \leq \boldsymbol{\mu}_{\max}, \sigma_{\min}^2 \boldsymbol{I} \preceq \boldsymbol{\Sigma} \preceq \sigma_{\max}^2 \boldsymbol{I} \Big\}. \tag{3}$$

Proposition S1.1 shows that if the outside mass $\varepsilon(\boldsymbol{\lambda}) = 1 - \int_{\mathcal{S}} q_{\boldsymbol{\lambda}}(\boldsymbol{x}) d\boldsymbol{x}$ is uniformly small on Θ, then the objective gap is uniformly small. The same reasoning applies component-wise

to mixtures on the parameter set Θ_K

$$\Theta_K := \Big\{ \boldsymbol{\Lambda} = (v_1, \ldots, v_{K-1}, \boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_K) : |v_k| \leq v_{\max}, \boldsymbol{\lambda}_k \in \Theta \Big\}, \tag{4}$$

with bounded logits $v_k := \log(\pi_k/\pi_K)$, which prevent degeneracies while preserving flexibility (Suppl. S1.2, eq. (S3)).

In the multimodal optimization setting, the main source of bias is due to the entropy-induced repulsion in mixtures. We deliberately exploit this repulsion early in training to mitigate mode collapse, a known issue in variational Gaussian mixtures (Wu et al., 2019; Jerfel et al., 2021; Soletskyi et al., 2025). The induced bias progressively diminishes under annealing.

3.2 Fixed-temperature asymptotic theory

The following theorem establishes that GERVE remains statistically consistent at fixed $\omega = \omega_0 > 0$, as $N \to \infty$. An asymptotic normality result is also derived. These results are stated for Gaussian mixtures in Θ_K (eq. (4)), but can be applied to single Gaussians in Θ (eq. (3)) as a special case. The proof is provided in Supplementary Material S2.3.

Theorem 3.1 (Asymptotics of empirical maximizers)**.** *At fixed $\omega_0 > 0$, if p is bounded and continuous on $\mathcal{S}$, then a population maximizer $\boldsymbol{\Lambda}^\star_{\omega_0}$ of $\mathcal{L}_{\omega_0}$ exists on Θ_K. If it is unique, then any empirical maximizer $\widehat{\boldsymbol{\Lambda}}_{\omega_0,N} \in \arg\max \widehat{\mathcal{L}}_{\omega_0,N}$ satisfies, as $N \to \infty$:*

$$\widehat{\boldsymbol{\Lambda}}_{\omega_0,N} \xrightarrow{P} \boldsymbol{\Lambda}^\star_{\omega_0}.$$

Furthermore, if $\boldsymbol{H}^\star_{\omega_0} := -\nabla^2_{\boldsymbol{\Lambda}} \mathcal{L}_{\omega_0}(\boldsymbol{\Lambda}^\star_{\omega_0}) \succ 0$, then

$$\sqrt{N}\big(\widehat{\boldsymbol{\Lambda}}_{\omega_0,N} - \boldsymbol{\Lambda}^\star_{\omega_0}\big) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \boldsymbol{W}_{\omega_0}),$$

where $\boldsymbol{W}_{\omega_0} := \big(\boldsymbol{H}^\star_{\omega_0}\big)^{-1} \boldsymbol{V}_{\omega_0} \big(\boldsymbol{H}^\star_{\omega_0}\big)^{-1}$, and $\boldsymbol{V}_{\omega_0} := \mathrm{Var}\left(\nabla_{\boldsymbol{\Lambda}} q_{\boldsymbol{\Lambda}}(\boldsymbol{X})\right)\big|_{\boldsymbol{\Lambda}=\boldsymbol{\Lambda}^\star_{\omega_0}}$.

Remark. Maximizer uniqueness is subtle. Compactness of the parameter space is essential, because in mixture models a vanishing weight lets the associated mean and covariance drift without changing the mixture density, and if several components share the same $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ then redistributing weights within that tied group also leaves the density unchanged. As a result, convergence at the parameter level cannot be guaranteed in general, even when the induced densities converge to a maximizer.

To formalize consistency without uniqueness, we can replace the assumption in Theorem 3.1 by the weaker requirement that $\mathcal{D}_{\omega_0} = \arg\max \mathcal{L}_{\omega_0}$ is nonempty and compact, and invoke the set-valued argmax theorem (van der Vaart, 1998, Cor. 5.58) to obtain $\mathrm{dist}\left(\widehat{\boldsymbol{\Lambda}}_{\omega_0,N}, \mathcal{D}_{\omega_0}\right) \xrightarrow{P} 0$, where $\mathrm{dist}(\boldsymbol{x}, \mathcal{A}) = \inf_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{x} - \boldsymbol{a}\|$.

The asymptotic normality statement, however, holds only under local identifiability: $\boldsymbol{H}^{\star}_{\omega_0} \succ 0$, which excludes duplicated components at the maximizer.

In practice, non-uniqueness can help: redundant components are absorbed as some weights vanish or as means coalesce near the modes, so an overcomplete mixture adapts to the effective number of modes without the user pre-specifying K.

3.3 Application to mode estimation and modal clustering

As demonstrated by Theorem 2.1, optimal variational mixtures concentrate on the modes of p as $\omega \to 0$. With the consistency guaranties of Theorem 3.1, GERVE therefore defines a mode estimator: Gaussian mixtures $q_{\boldsymbol{\Lambda}} = \sum_{k=1}^{K} \pi_k q_{\boldsymbol{\lambda}_k}$ simultaneously capture multiple modes under annealing.

Beyond mode estimation, GERVE naturally induces a clustering structure. With possibly overcomplete mixtures, components are attracted toward distinct modes as entropy decreases. Redundant components either vanish or collapse, so the effective number of clusters adapts automatically. Unlike likelihood-based Gaussian mixtures clustering, GERVE targets density modes under annealing.

In practice, once fitted, cluster assignments are straightforward via responsibilities (posterior component probabilities). This connects to *modal clustering* (Menardi, 2016; Chacón, 2020), where clusters are attraction basins of density modes. Classical approaches rely on nonparametric density estimates and gradient flows, while GERVE uses a parametric variational approximation. The resulting clusters need not coincide exactly with attraction basins, but they agree locally near each mode under mild conditions, see Supplementary Material S1.3 for a statement and a counterexample. This yields adaptive modal clustering without tuning K. A caveat is that annealing emphasizes dominant modes: secondary modes may disappear as $\omega \to 0$, so retaining them requires stopping at a positive temperature or imposing weight lower bounds.

4 GERVE algorithm and variants

4.1 Variational optimization with natural gradients

The empirical objective $\widehat{\mathcal{L}}_{\omega,N}$ (eq. (2)) can be optimized over Θ or Θ_K using natural gradients $\widetilde{\nabla}_{\boldsymbol{\vartheta}} \widehat{\mathcal{L}}_{\omega,N}(\boldsymbol{\vartheta}) := \boldsymbol{F}(\boldsymbol{\vartheta})^{-1} \nabla_{\boldsymbol{\vartheta}} \widehat{\mathcal{L}}_{\omega,N}(\boldsymbol{\vartheta})$, which incorporates the geometry of the variational parameter space through the Fisher information matrix $\boldsymbol{F}(\boldsymbol{\vartheta})$ (Amari, 1998). When $q_{\boldsymbol{\vartheta}}$ belongs to an exponential family with sufficient statistic $\boldsymbol{T}(\boldsymbol{X})$, the natural gradient equals the gradient in expectation parameters $\boldsymbol{M} = \mathbb{E}_{q_{\boldsymbol{\vartheta}}}[\boldsymbol{T}(\boldsymbol{X})]$:

$$\widetilde{\nabla}_{\boldsymbol{\vartheta}} \widehat{\mathcal{L}}_{\omega,N}(\boldsymbol{\vartheta}) = \nabla_{\boldsymbol{M}} \widehat{\mathcal{L}}_{\omega,N}(\boldsymbol{\vartheta}). \tag{5}$$

This identity allows computation of natural gradients without explicit estimation and inversion of the Fisher matrix. It also holds for untruncated Gaussian mixtures via the

Minimal Conditional Exponential Family representation (MCEF; Lin et al., 2019).

Algorithm 1 describes natural gradient ascent on $\widehat{\mathcal{L}}_{\omega,N}(\boldsymbol{\vartheta})$ with a temperature schedule ω_t over iterations $t = 1, \ldots, T$. Convergence guarantees are stated in Supplementary Material S1.4: at fixed-ω, updates converge to the stationary set under standard Robbins–Monro conditions and bounded iterates (Thm. S1.11); with $\omega_t \to 0$, stationary sets can be tracked with slow schedules, i.e. $|\omega_{t+1} - \omega_t| = o(\rho_t)$ where ρ_t is a gradient stepsize (Cor. S1.12).

Algorithm 1: GERVE (general form)

```
INPUT: samples X_{1:N}.
SET: initial ϑ_1, stepsizes (ρ_t)_t, schedule (ω_t)_t, (optional) batch size B.
for t = 1:T do
    (Optional) SAMPLE a mini-batch {X_i^(t)}_{i=1}^B from X_{1:N} (or use full data).
    COMPUTE an (unbiased) natural-gradient estimate ∇̃ L̂_{ω_t,N}(ϑ_t).
    UPDATE ϑ_{t+1} ← ϑ_t + ρ_t ∇̃ L̂_{ω_t,N}(ϑ_t).
    (Optional) PROJECT to a compact set.
end
RETURN: ϑ_{T+1}.
```

4.2 Variant A: Equivalence to Gaussian mean-shift

Consider the family of Gaussian distributions with a fixed isotropic covariance $\sigma^2 \boldsymbol{I}$, $\{q_{\boldsymbol{\mu}} = \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}), \boldsymbol{\mu} \in \mathbb{R}^d\}$. For a large enough $\mathcal{S}$, we have $\nabla_{\boldsymbol{\mu}} \mathcal{H}_{\mathcal{S}}(q_{\boldsymbol{\mu}}) \approx 0$, so the natural-gradient step can be simplified into:

$$\boldsymbol{\mu}_{t+1} = \boldsymbol{\mu}_t + \frac{\rho_t}{N} \sum_{i=1}^{N} (\boldsymbol{X}_i - \boldsymbol{\mu}_t) q_{\boldsymbol{\lambda}_t}(\boldsymbol{X}_i). \tag{6}$$

Using a batch $\{\boldsymbol{X}_i^{(t)}\}_{i=1}^B$ gives an unbiased estimator of the RHS (see Suppl. S3.2 for the derivation). Denoting by $p_h(\boldsymbol{\mu}) = \frac{1}{N} \sum_{i=1}^{N} \varphi_h(\boldsymbol{X}_i - \boldsymbol{\mu})$ the KDE with a Gaussian kernel φ_h with bandwidth h, the following result states that GERVE recovers the mean-shift algorithm when $h = \sigma^2$.

Proposition 4.1 (Equivalence to Gaussian mean-shift)**.** *Update rule (6) performs gradient ascent on p_h, and normalizing the step $\rho_t = p_h(\boldsymbol{\mu})^{-1}$ yields the classical mean-shift fixed-point update:*

$$\boldsymbol{\mu}_{t+1} = \frac{\sum_{i=1}^{N} \varphi_h(\boldsymbol{X}_i - \boldsymbol{\mu}_t) \boldsymbol{X}_i}{\sum_{j=1}^{N} \varphi_h(\boldsymbol{X}_j - \boldsymbol{\mu}_t)}.$$

Viewing fixed-covariance GERVE as mean-shift puts us under classical KDE theory. Operationally, $\boldsymbol{\mu}_t$ is a particle ascending the smoothed density p_h, and the (fixed) covariance

$h\boldsymbol{I}$ is the bandwidth: large h leads to smooth and global but biased updates, while small h yields sharp but noisy updates, sensitive to sampling.

This regime is thus a well-understood nonparametric baseline that grounds richer GERVE variants. This can typically be used to set the covariance bounds in learned-covariance variants of GERVE. For example, if p is $\mathcal{C}^3$, the gradient of the Gaussian KDE has bias $O(h^2)$ and fluctuation $O_p\big((Nh^{d+2})^{-1/2}\big)$, yielding the optimal plug-in rate $h \asymp N^{-1/(d+6)}$ (Parzen, 1962; Romano, 1988; Tsybakov, 1990; Genovese et al., 2016).

4.3 Variant B: Gaussian mixture GERVE

When considering Gaussian mixtures represented by Θ_K, it is convenient to adopt the following parametrization. Let $q_{\boldsymbol{\Lambda}} = \sum_{k=1}^K \pi_k \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{S}_k^{-1})$, where $\boldsymbol{S}_k := \boldsymbol{\Sigma}_k^{-1}$ with logits $v_k = \log(\pi_k/\pi_K)$ and $\boldsymbol{\Lambda} = (v_{1:K-1}, \boldsymbol{\lambda}_{1:K}) \in \Theta_K$. Natural-gradient steps lead to component-wise updates coupled via the entropy term (full derivation in Suppl. S3.3):

$$
\begin{aligned}
\boldsymbol{S}_{k,t+1} &= \boldsymbol{S}_{k,t}\Big(1 - \frac{\rho_t}{N}\sum_{i=1}^N \Big((\boldsymbol{X}_i - \boldsymbol{\mu}_{k,t})(\boldsymbol{X}_i - \boldsymbol{\mu}_{k,t})^T \boldsymbol{S}_{k,t} - \boldsymbol{I}\Big) q_{\boldsymbol{\lambda}_{k,t}}(\boldsymbol{X}_i)\Big) \\
&\quad - \rho_t \omega_t \nabla_{\boldsymbol{S}_k^{-1}} \mathcal{H}_{\mathcal{S}}(q_{\boldsymbol{\Lambda}})|_{\boldsymbol{\Lambda}=\boldsymbol{\Lambda}_t}, \\
\boldsymbol{\mu}_{k,t+1} &= \boldsymbol{\mu}_{k,t} + \frac{\rho_t}{N} \boldsymbol{S}_{k,t+1}^{-1} \boldsymbol{S}_{k,t} \sum_{i=1}^N (\boldsymbol{X}_i - \boldsymbol{\mu}_{k,t}) q_{\boldsymbol{\lambda}_{k,t}}(\boldsymbol{X}_i) + \rho_t \omega_t \nabla_{\boldsymbol{\mu}_k} \mathcal{H}_{\mathcal{S}}(q_{\boldsymbol{\Lambda}})|_{\boldsymbol{\Lambda}=\boldsymbol{\Lambda}_t}, \\
v_{k,t+1} &= v_{k,t} + \frac{\rho_t}{N} \sum_{i=1}^N \big(q_{\boldsymbol{\lambda}_{k,t}}(\boldsymbol{X}_i) - q_{\boldsymbol{\lambda}_{K,t}}(\boldsymbol{X}_i)\big) - \rho_t \omega_t \nabla_{\pi_k} \mathcal{H}_{\mathcal{S}}(q_{\boldsymbol{\Lambda}})|_{\boldsymbol{\Lambda}=\boldsymbol{\Lambda}_t}.
\end{aligned}
$$

Learned covariances adapt to anisotropy and converge to positive limits at fixed temperature. As $\omega \to 0$, covariances shrink faster. Furthermore, the entropy term induces repulsion between components at large ω that vanishes as $\omega \to 0$. Therefore, a slower decreasing annealing schedule allows for longer exploration by delaying mode collapse and improving mode capture in rugged and multimodal landscapes. Overcomplete mixtures (when K is larger than the number of modes) naturally share components per mode (see Thm. 2.1).

Gradient derivations and Monte Carlo estimators for the entropy terms are in Supplementary Material S3.3. Diagonal, isotropic, or fixed-covariance variants are obtained by modifying Θ_K, and derivations follow the same pattern.

4.4 Complexity and practical guidelines

With T iterations, batch size B, and a Gaussian mixture with K components, the computational cost is $O(d^3BKT)$. The d^3 factor reflects the matrix multiplications and inversions that arise when updating full precision matrices. Restricting the variational family reduces this cost to $O(dBKT)$ for diagonal or fixed covariances. These restricted families often preserve the qualitative behavior of the algorithm and they improve scalability, which brings

the cost in line with classical mode-seeking and clustering procedures. Further details appear in Supplementary Material S3.4.

We now summarize practical guidance for hyperparameter selection. For the temperature schedule, the initial value of ω should be large so that mixture components repel and cover $\mathcal{S}$. The temperature should then decrease to a target value $\omega^\dagger$ as slowly as possible so that the iterates continue to track the relevant stationary sets, with $\omega^\dagger = 0$ for mode estimation and $\omega^\dagger = \omega_0$ with $\omega_0 > 0$ for clustering, as formalized in Corollary S1.12.

For stepsizes, a safe choice satisfies the Robbins–Monro conditions $\sum_t \rho_t = \infty$ and $\sum_t \rho_t^2 < \infty$. In practice, when the temperature ω_t decreases over time, we find that mean updates are stabilized by the scaling $\rho_t \propto \omega_t^{-\beta}$ with $\beta \in (0, 1)$. This choice often accelerates convergence, although it can violate the Robbins–Monro conditions.

Compactness constraints require explicit bounds for the covariance parameters in Θ and for the mixture parameters in Θ_K. The upper bound $\sigma^2_{\max}$ should be large enough so that a single component can cover the entire dataset. For mode estimation, a suitable rule for the lower bound is $\sigma^2_{\min}$ of the order of magnitude of $N^{-1/(d+6)}$, which follows classical mean-shift bandwidth theory. For clustering, $\sigma^2_{\min}$ should be comparable to intra-class variances. For the logits that define the weights in Θ_K (eq. (4)), setting the upper bound $v_{\max} \geq 6$ allows any subset of weights to be smaller than 10^{-5}, which makes it possible to identify vanishing components while maintaining compactness (see Suppl. S1.2, eq. (S3)).

5 Bootstrap uncertainty quantification for mode estimation

Mode-level uncertainty can be provided at a fixed stopping temperature $\omega_0 > 0$, using a bootstrap principle. A GERVE baseline fit is performed on the initial sample (constant $\omega_t = \omega_0$). GERVE is then refit L times on resampled data. Each set of resulting modes is matched to the baseline modes, providing some empirical spread as confidence ellipses and a stability score.

5.1 Bootstrap procedure

The bootstrap procedure is summarized in Algorithm 2 for Gaussian mixture GERVE. Each fit, $\ell = 0, \ldots, L$, uses an overcomplete K-component mixture followed by pruning (for some threshold $\epsilon \ll 1$) and merging of near-duplicates. The first fit identifies $\widehat{K}$ "baseline" modes. Each bootstrap fit $\ell = 1, \ldots, L$ produces its own set of $\widehat{K}^{(\ell)}$ modes, which are matched to the baseline modes via minimum-cost assignment (Hungarian algorithm, Kuhn, 1955). Then, we form confidence intervals from the empirical distribution of matched modes, and monitor mode recovery with the stability score

$$s_k = \frac{1}{L} \sum_{\ell=1}^{L} \mathbb{1}\{\text{mode } k \text{ is matched in replicate } \ell\}. \tag{7}$$

Algorithm 2: Bootstrap uncertainty quantification for mode estimation

INPUT: samples $\boldsymbol{X}_{1:N}$, temperature ω_0, number of replicates L.
FIT GERVE at ω_0 on $\boldsymbol{X}_{1:N}$, prune/merge to obtain baseline modes $\{\widehat{\boldsymbol{\mu}}_k^{(0)}\}_{k=1}^{\widehat{K}}$.
for $\ell = 1:L$ **do**
    SAMPLE $J_{1:N}^{(\ell)} \overset{\text{i.i.d.}}{\sim} \text{Unif}([N])$, set $\boldsymbol{X}_i^{(\ell)} \leftarrow \boldsymbol{X}_{J_i^{(\ell)}}$ for $i = 1:N$.
    FIT GERVE at ω_0 on $\boldsymbol{X}_{1:N}^{(\ell)}$, prune/merge to obtain $\{\widehat{\boldsymbol{\mu}}_j^{(\ell)}\}_{j=1}^{\widehat{K}^{(\ell)}}$.
    MATCH $\{\widehat{\boldsymbol{\mu}}_j^{(\ell)}\}_{j=1}^{\widehat{K}^{(\ell)}}$ to $\{\widehat{\boldsymbol{\mu}}_k^{(0)}\}_{k=1}^{\widehat{K}}$ (Hungarian).
    RECORD for each k: the matched $\widehat{\boldsymbol{\mu}}_k^{(\ell)}$.
end
RETURN: per-mode confidence ellipses from $\{\widehat{\boldsymbol{\mu}}_k^{(\ell)}\}_{\ell=1}^{L}$ and stability scores $(s_k)_{k=1}^{\widehat{K}}$.

5.2 Statistical validity

Let $\widehat{\mathcal{L}}_{\omega_0,N}$ be the empirical objective at fixed $\omega_0 > 0$ on a compact parameter set Θ_K (eq. (4)), and let $\widehat{\boldsymbol{\Lambda}}_{\omega_0,N} \in \arg\max_{\boldsymbol{\Lambda}\in\Theta_K} \widehat{\mathcal{L}}_{\omega_0,N}(\boldsymbol{\Lambda})$ with a unique population maximizer $\boldsymbol{\Lambda}^\star_{\omega_0}$. Write P^* for the conditional probability given the observed sample $\boldsymbol{X}_{1:N}$. Bootstrap limits are understood conditionally on $\boldsymbol{X}_{1:N}$, in P-probability. Define the bootstrap criterion and maximizer

$$\widehat{\mathcal{L}}^*_{\omega_0,N}(\boldsymbol{\Lambda}) = E_{P_N^*}[q_{\boldsymbol{\Lambda}}(\boldsymbol{X})] + \omega_0 \mathcal{H}_{\mathcal{S}}(q_{\boldsymbol{\Lambda}}), \qquad \widehat{\boldsymbol{\Lambda}}^*_{\omega_0,N} \in \underset{\boldsymbol{\Lambda}\in\Theta_K}{\arg\max}\, \widehat{\mathcal{L}}^*_{\omega_0,N}(\boldsymbol{\Lambda}).$$

In addition to unicity, we assume that the population maximizer satisfies $\boldsymbol{H}^\star_{\omega_0} = -\nabla^2_{\boldsymbol{\Lambda}}\mathcal{L}_{\omega_0}(\boldsymbol{\Lambda}^\star_{\omega_0}) \succ 0$, so that Theorem 3.1 fully applies.

The following statements guarantee the validity of bootstrap for parameters (Thm. 5.1), matched modes (Thm. 5.3), and their robustness to inexact maximization (Prop. 5.4). All results are proved in Supplementary Material S2.4. In Supplementary Material S1.5, we also investigate the consistency of stability scores and propose strategies to handle nonconvex objectives.

Theorem 5.1 (Parameter-level bootstrap validity)**.** *We have, as* $N \to \infty$*:*

$$\sqrt{N}\big(\widehat{\boldsymbol{\Lambda}}^*_{\omega_0,N} - \widehat{\boldsymbol{\Lambda}}_{\omega_0,N}\big) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \boldsymbol{W}_{\omega_0}) \quad \textit{conditionally on } \boldsymbol{X}_{1:N},$$

where $\boldsymbol{W}_{\omega_0} := \big(\boldsymbol{H}^\star_{\omega_0}\big)^{-1}\boldsymbol{V}_{\omega_0}\big(\boldsymbol{H}^\star_{\omega_0}\big)^{-1}$*, and* $\boldsymbol{V}_{\omega_0} := \text{Var}\left(\nabla_{\boldsymbol{\Lambda}} q_{\boldsymbol{\Lambda}}(\boldsymbol{X})\right)\big|_{\boldsymbol{\Lambda}=\boldsymbol{\Lambda}^\star_{\omega_0}}$.

For the theoretical analysis, we fix deterministic target locations $\boldsymbol{u}_{1:K_0}$, which act as reference anchors for defining the matching map. In practice, these targets correspond to the subset of baseline mode estimates we intend to track, so K_0 is taken no larger than the number $\widehat{K}$ of resolved baseline modes returned by the initial fit. To move from parameters

to matched modes, we then define the matching map $\mathcal{M}$ that stacks the K_0 component means closest to their respective targets $\boldsymbol{u}_{1:K_0}$ (see Suppl. S1.5.1 for more details). The following guarantees only require a local separation margin.

Assumption 5.2 (Local separation for matching)**.** *There exists $\delta > 0$ such that near $\boldsymbol{\Lambda}^\star_{\omega_0}$, each target $\boldsymbol{u}_j$ has a unique nearest component mean and the nearest index is locally constant (see Suppl. S1.5.1 for a formal definition).*

When separation fails, bootstrap ellipses inflate and the stability score s_k drops, which is informative about ambiguous modes (see Suppl. S1.5 for a consistency result for s_k).

Theorem 5.3 (Matched-modes bootstrap validity and confidence ellipses)**.** *Under Assumption 5.2, $\mathcal{M}$ is $\mathcal{C}^1$ at $\boldsymbol{\Lambda}^\star_{\omega_0}$ with Jacobian $\boldsymbol{J}$, and as $N \to \infty$:*

$$\sqrt{N}\big(\mathcal{M}(\widehat{\boldsymbol{\Lambda}}_{\omega_0,N}) - \mathcal{M}(\boldsymbol{\Lambda}^\star_{\omega_0})\big) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \boldsymbol{C}_{\mathcal{M}}), \qquad \sqrt{N}\big(\mathcal{M}(\widehat{\boldsymbol{\Lambda}}^*_{\omega_0,N}) - \mathcal{M}(\widehat{\boldsymbol{\Lambda}}_{\omega_0,N})\big) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \boldsymbol{C}_{\mathcal{M}}),$$

where $\boldsymbol{C}_{\mathcal{M}} = \boldsymbol{J}\boldsymbol{W}_{\omega_0}\boldsymbol{J}^T$. For each matched mode j, the marginal covariance is $\boldsymbol{C}_j$, so percentile (and, with a consistent $\widehat{\boldsymbol{C}}_j$, studentized) bootstrap ellipses are asymptotically valid.

Finally, we allow inexact baseline and bootstrap fits. Let $\widetilde{\boldsymbol{\Lambda}}_{\omega_0,N}$ and $\widetilde{\boldsymbol{\Lambda}}^*_{\omega_0,N}$ denote any approximate solutions that are first-order stationary up to $o_P(N^{-1/2})$ and $o_{P^*}(N^{-1/2})$, respectively, that is $\|\nabla_{\boldsymbol{\Lambda}}\widehat{\mathcal{L}}_{\omega_0,N}(\widetilde{\boldsymbol{\Lambda}}_{\omega_0,N})\| = o_P(N^{-1/2})$ and $\|\nabla_{\boldsymbol{\Lambda}}\widehat{\mathcal{L}}^*_{\omega_0,N}(\widetilde{\boldsymbol{\Lambda}}^*_{\omega_0,N})\| = o_{P^*}(N^{-1/2})$.

Proposition 5.4 (Robustness to inexact maximization)**.** *The conclusions of Theorems 5.1 and 5.3 hold with $\widetilde{\boldsymbol{\Lambda}}_{\omega_0,N}$ and $\widetilde{\boldsymbol{\Lambda}}^*_{\omega_0,N}$ in place of $\widehat{\boldsymbol{\Lambda}}_{\omega_0,N}$ and $\widehat{\boldsymbol{\Lambda}}^*_{\omega_0,N}$.*

6 Simulation studies

We show that GERVE recovers modes and yields useful clusterings without tuning K. Furthermore, on simulated data, we illustrate (i) how clustering emerges as a byproduct of variational mode estimation and (ii) the consistency of mode recovery.

6.1 Clustering example

To illustrate the clustering properties of GERVE, we simulate $N = 6000$ points from a 2D, 3-component Gaussian mixture with means at the nodes of an equilateral triangle, with isotropic covariance $\sigma^2\boldsymbol{I}$, $\sigma^2 = 0.25$. The data exhibits $J = 3$ high-density regions that we aim to recover. Details on the setup and a density plot can be found in Supplementary Material S4.1.

We run GERVE with Gaussian mixtures (full covariances) under two regimes: matched $K = 3$ and overcomplete $K = 7$. To showcase GERVE's robustness to component mean

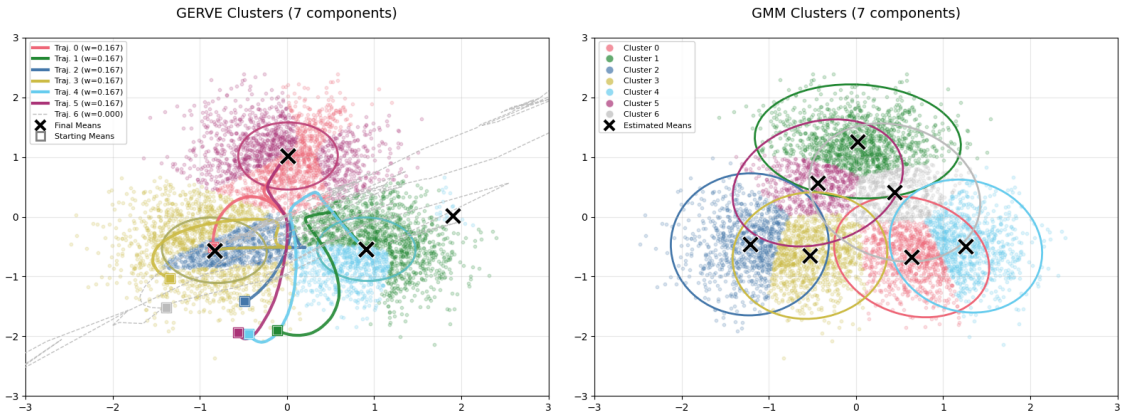

Figure 1: Overspecified clustering ($K = 7$) of a triangle Gaussian mixture sample. Data points are colored according to the component with the highest posterior responsibility in the fitted mixture. Ellipses represent the component covariances. Left: GERVE returns 3 effective clusters (the component with dashed trajectory has vanishing weight, the remaining six components form three clusters by grouping equivalent components). Right: GMM-EM returns a partition into 7 clusters.

initialization, we initialize means in $[-2, 0]^2$. We start with broad covariances and equal weights. We use a continuously decreasing annealing schedule, and compensate with step size $\rho_t \propto \omega_t^{-0.7}$ (full hyperparameters and schedules in Suppl. S4.1). After training, assign by highest responsibility and prune weights $< 10^{-3}$. Baselines are GMM-EM (Gaussian mixture model fit by the EM algorithm) with full covariances and k-means for $K \in \{3, 7\}$.

Figure 1 shows the obtained partitions when $K = 7$. With this overcomplete mixture, GERVE still yields 3 effective clusters: redundant components coalesce at modes or are pruned, whereas GMM-EM returns 7 clusters, so does k-means (Suppl. S4.1, Fig. S3). Because GERVE targets modal structure rather than a clustering likelihood, it is more robust to over-specification. With $K = 3$, all methods recover 3 clusters (Suppl. S4.1, Fig. S3).

In Supplementary Material S5, we assess robustness against classical methods with a benchmark on standard UCI datasets. We also provide an ablation study and report runtimes.

6.2 Mode-finding performance

We use the same triangle setting as previously, with sharper components ($\sigma^2 = 0.1$). True global modes practically coincide with means.

Over $n_{\text{rep}} = 100$ replicates, each method outputs K mode estimates. We report three metrics (see Suppl. S4.2): (i) *Mode recovery* (MR_ϵ), the number of true modes recovered

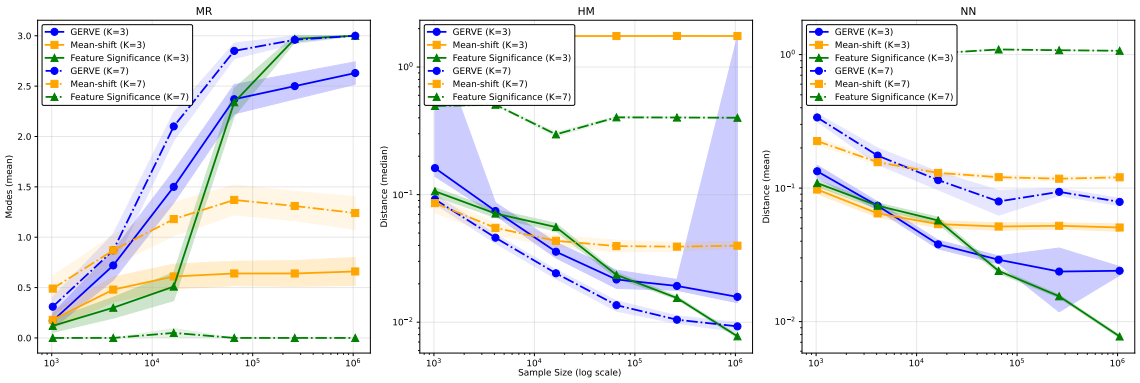

Figure 2: Mode estimation vs. sample size N for the triangle mixture ($I = 3$) example. Curves: means (MR_ϵ, NN) or medians (HM) over $n_{\text{rep}} = 100$; bands: 95% confidence intervals.

within a tolerance ϵ, we set $\epsilon = 10^{-2}$; (ii) *Hungarian matching* sum (HM), minimum-assignment cost between between true and estimated modes; (iii) *Nearest-neighbor* sum (NN), aggregate distance from each estimate to the closest mode.

We run GERVE (Gaussian mixture with full covariances) with $K \in \{3, 7\}$ and a vanishing annealing schedule. Baselines are Gaussian mean-shift with fixed-point updates, modes founds aggregated from K initializations, and the Feature Significance method via the `feature` R package, which detects significant curvature points, then estimates modes as the centroids of k-means. For each sample size $N \in \{2^{10}, 2^{12}, \ldots, 2^{20}\}$ and value of K, we perform a light hyperparameter grid search (Suppl. S4.2). We average MR_ϵ, NN and report HM medians with 95% confidence intervals.

Fig. 2 summarizes performance with respect to sample size N for $K \in \{3, 7\}$. GERVE's MR_ϵ increases towards the number of true modes and HM decreases as N grows, illustrating the consistency of mode recovery (this matches the conclusions of Thm 3.1). With $K = 3$, GERVE's NN is comparable to that of mean-shift and Feature Significance, showcasing its mode-location accuracy. With $K = 7$, GERVE improves MR_ϵ and HM while keeping NN competitive. This exhibits GERVE's robustness to overspecification, whereas Feature Significance is degraded. A broader sweep over $K \in \{3, \ldots, 7\}$ appears in Supplementary Material S4.2.

To summarize the key aspects highlighted in this simulation study: (i) Clustering emerges from mode estimation, with robustness to overspecification; (ii) For mode estimation, GERVE is statistically consistent and competitive in location accuracy, and mild overspecification helps GERVE where it hinders the other baselines.

7 Application: UK road collision hotspots detection

Identifying collision hotspots (spatial intensity modes) and quantify their uncertainty for intervention planning is crucial in road safety studies. Count-based metrics and KDE are common (Xie and Yan, 2008; Okabe et al., 2009). Spatial scans add significance testing and Bayesian models add uncertainty, but at the cost of parametric assumptions and heavier computation (Kulldorff, 1997; Aguero-Valverde and Jovanis, 2006). GERVE offers a likelihood-free alternative that estimates mode locations with calibrated localization uncertainty.

We analyze UK `STATS19` collisions from the Department for Transport Road Safety Data portal (*www.gov.uk/government/statistics/road-safety-data*), restricting to A-roads in Greater London (longitudes $[-0.54, 0.33]$, latitudes $[51.28, 51.70]$) and severity labeled "fatal" or "serious" from 2020 to 2024.

We fit an overcomplete mixture, targeting 10 hotspots and using $K = 20$ components to enable overcompleteness, and resolve modes by pruning and merging (details in Suppl. S6). The stopping temperature $\omega^\dagger$ is chosen at the elbow, the largest ω where the resolved-mode count peaks (Suppl. S6, Fig. S10). At $\omega^\dagger$, we compute $L = 500$ bootstrap resamples, match bootstrap modes to the baseline by assignment with adaptive gates (Suppl. S6), and project to a metric system (EPSG:27700). For each mode k, we report a stability score s_k and a 95% confidence ellipse from the empirical covariance of matched centers. Under the fixed-temperature theory in Section 5, ellipses have nominal coverage and the stability score s_k (eq. (7)) estimates the recovery probability.

Figure 3 maps 2020-2024 hotspots with 95% ellipses and s_k in the zoomed-in window $[-0.15, 0.15] \times [-0.075, 0.075]$. High-stability locations ($s_k \geq 0.7$, arbitrary threshold) such as Shoreditch (ID: 12) and Elephant & Castle (ID: 1) have ellipses of order 100-150 m, which supports intersection-level action. Medium-stability locations ($0.4 \leq s_k < 0.7$) such as Aldgate (ID: 8) and Clapham HS (ID: 16) show wider ellipses that span multiple junctions, which suggests area-wide measures. Table S5 in Supplementary Material S6 lists the entries, ranked by stability score.

To compare like-for-like, we use mean-shift, operating on the same object as GERVE. With a flat kernel and a bandwidth sweep to a data-driven reference h_0 (see Suppl. S6), mean-shift enumerates many local maxima at small bandwidths and merges aggressively at larger ones. In the zoomed-in window, mean-shift at $0.05h_0$ yields 122 centers, while GERVE reports 17. The contrast reflects design: entropy-regularized mixtures impose global competition for support, so weaker peaks are absorbed as temperature falls. This yields a coarser and more stable partition that aligns with major junctions. A visual comparison appears in Supplementary Material S6, Figure S14.

For temporal stability, we fit GERVE on 2014-2019 and 2020-2024 and declare a hotspot persistent if 95% ellipses overlap across periods. Eight of the 17 hotspots persist, which indicates structural risk that is stable across years. Non-persistent hotspots (Piccadilly Circus, Gunter Grove, Clapham HS) may reflect temporally varying factors such as pandemic-

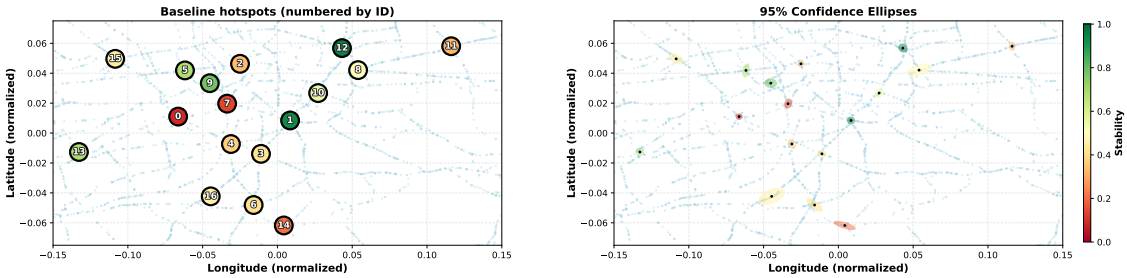

Figure 3: Baseline collision hotspots identified by GERVE for the Greater London Area between 2020 and 2024, with stability scores. The normalized coordinates are bounded within $[-0.7, 0.7] \times [-0.35, 0.35]$ (normalized window) box, but all hotspots are in the $[-0.15, 0.15] \times [-0.075, 0.075]$ box (zoomed-in window), see Supplementary Material S6, Figure S12 for the full study area. Left: with hotspot IDs for reference to Table S5, Supplementary Material S6. Right: 95% confidence ellipses, in normalized coordinates.

related traffic shifts, local construction, or policy changes. Supplementary Material S6, Figure S16 overlays the two periods.

Stability and ellipses can be used to target actions. High s_k and tight ellipses support redesign at a named junction, whereas moderate s_k and wide ellipses reveal corridor-level risks, which are more suitable for monitoring.

8 Conclusion

GERVE is a likelihood-free variational approach to mode estimation and modal clustering. It optimizes an entropy-regularized objective with natural gradients under annealing, fitting mixtures that concentrate on high-density regions and induce clusters without density estimation. As temperature goes to zero, components align with population modes and inherit local curvature. At fixed temperature, empirical maximizers exist, are consistent, and satisfy a central limit theorem. A matched-mode bootstrap procedure yields per-mode confidence sets.

Empirically, GERVE recovers modal structure and adapts the effective number of groups as temperature decreases, with redundant components merging or vanishing. Diagonal-covariance variants preserve most of the behavior at lower cost, and an ablation study shows that annealing, moderate component overspecification, and short burn-in phases are useful levers.

GERVE is most suitable when uncertainty in mode locations is important, when the cluster count is misspecified, or when modal clustering is preferable to partition-based grouping. For problems where the number of clusters is known and computational speed is paramount, established task-specific clustering algorithms remain preferable. Three lim-

itations merit attention. First, modal clustering may misalign with the group structure that matters in an application. Modes define clusters by attraction basins, but scientific groupings may depend on other criteria, so agreement is not guaranteed. Second, the optimization procedure can converge to local rather than global maxima. Careful initialization, temperature schedules, and overcomplete mixtures reduce this risk but do not completely eliminate it. Third, computation can be costly with full covariance matrices in high dimensions. In our experiments, using diagonal covariances allows one to control the cost while preserving mode separation.

The theory and experimental results position GERVE as a principled statistical framework for likelihood-free modal inference. By unifying modal clustering, mean-shift, variational inference, and annealing, GERVE provides both algorithmic tools and theoretical guaranties within one platform. Looking ahead, three directions are natural. First, develop data-driven annealing schedules with principled stopping. Second, establish pathwise guaranties for empirical maximizers and for the validity of the bootstrap procedure under annealing. Third, characterize mode-component correspondence and label stability as temperature varies. Pursuing these directions will link the annealing procedure to stronger statistical guaranties and broaden the method's practical reach.

Acknowledgements

This work is supported by the French ANR and JST-CREST Bayes-Duality grant (ANR-21-JSTM-0001).

References

Aguero-Valverde, J. and Jovanis, P. P. (2006). Spatial analysis of fatal and injury crashes in Pennsylvania. *Accident Analysis & Prevention*, 38(3):618–625.

Amari, S.-I. (1998). Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276.

Ameijeiras-Alonso, J., Crujeiras, R. M., and Rodriguez-Casal, A. (2021). Multimode: an R package for mode assessment. *Journal of Statistical Software*, 97:1–32.

Arias-Castro, E., Mason, D., and Pelletier, B. (2016). On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. *Journal of Machine Learning Research*, 17:1–28.

Borkar, V. S. (2008). *Stochastic Approximation: a Dynamical Systems Viewpoint*. Springer.

Carreira-Perpiñan, M. Á. (2006). Acceleration strategies for Gaussian mean-shift image segmentation. *Conference on Computer Vision and Pattern Recognition*, 1:1160–1167.

Carreira-Perpiñan, M. Á. (2007). Gaussian mean-shift is an EM algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(5):767–776.

Chacón, J. E. (2015). A population background for nonparametric density-based clustering. *Statistical Science*, 30(4):518–532.

Chacón, J. E. (2020). The modal age of statistics. *International Statistical Review*, 88(1):122–141.

Chernoff, H. (1964). Estimation of the mode. *Annals of the Institute of Statistical Mathematics*, 16(1):31–41.

Comaniciu, D. and Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):603–619.

Dalenius, T. (1965). The mode–A neglected statistical parameter. *Journal of the Royal Statistical Society Series A: General*, 128(1):110–117.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Donsker, M. D. and Varadhan, S. R. S. (1976). Asymptotic evaluation of certain Markov process expectations for large time—III. *Communications on Pure and Applied Mathematics*, 29(4):389–461.

Dudley, R. M. (2002). *Real Analysis and Probability*. Cambridge University Press.

Fukunaga, K. and Hostetler, L. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1):32–40.

Genovese, C. R., Perone-Pacifico, M., Verdinelli, I., and Wasserman, L. (2016). Non-parametric inference for density modes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(1):99–126.

Hwang, C.-R. (1980). Laplace's method revisited: weak convergence of probability measures. *The Annals of Probability*, 8(6):1177–1182.

Jerfel, G., Wang, S., Wong-Fannjiang, C., Heller, K. A., Ma, Y., and Jordan, M. I. (2021). Variational refinement for importance sampling using the forward Kullback–Leibler divergence. *Uncertainty in Artificial Intelligence*, pages 1819–1829.

Khan, M. E. and Rue, H. (2023). The Bayesian learning rule. *Journal of Machine Learning Research*, 24(281):1–46.

Kronmal, R. and Tarter, M. (1968). The estimation of probability densities and cumulatives by Fourier series methods. *Journal of the American Statistical Association*, 63(323):925–952.

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics-Theory and Methods*, 26(6):1481–1496.

Kushner, H. J. and Yin, G. G. (2003). *Stochastic Approximation and Recursive Algorithms and Applications*. Springer.

Le Minh, T., Arbel, J., Möllenhoff, T., Khan, M. E., and Forbes, F. (2025). Natural variational annealing for multimodal optimization. *arXiv preprint arXiv:2501.04667*.

Lin, W., Khan, M. E., and Schmidt, M. (2019). Fast and simple natural-gradient variational inference with mixture of exponential-family approximations. *International Conference on Machine Learning*.

Liu, B. and Ghosh, S. K. (2020). On empirical estimation of mode based on weakly dependent samples. *Computational Statistics & Data Analysis*, 152:107046.

Loftsgaarden, D. O. and Quesenberry, C. P. (1965). A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics*, 36(3):1049–1051.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. John Wiley & Sons.

Menardi, G. (2016). A review on modal clustering. *International Statistical Review*, 84(3):413–433.

Noble, M., Grenioux, L., Gabrié, M., and Durmus, A. O. (2025). Learned reference-based diffusion sampler for multi-modal distributions. *International Conference on Learning Representations*.

Okabe, A., Satoh, T., and Sugihara, K. (2009). A kernel density estimation method for networks, its computational method and a GIS-based tool. *International Journal of Geographical Information Science*, 23(1):7–32.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076.

Romano, J. P. (1988). On weak convergence and optimality of kernel density estimates of the mode. *The Annals of Statistics*, 16(2):629–647.

Sager, T. W. (1978). Estimation of a multivariate mode. *The Annals of Statistics*, 6(4):802–812.

Samworth, R. J. (2018). Recent progress in log-concave density estimation. *Statistical Science*, 33(4):493–509.

Scott, D. W. (1992). *Multivariate density estimation: Theory, Practice, and Visualization*. John Wiley & Sons.

Soletskyi, R., Gabrié, M., and Loureiro, B. (2025). A theoretical perspective on mode collapse in variational inference. *Machine Learning: Science and Technology*, 6(2):025056.

Tsybakov, A. B. (1990). Recursive estimation of the mode of a multivariate distribution. *Problems of Information Transmission*, 26(1):38–45.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: with Applications to Statistics*. Springer.

Venter, J. H. (1967). On estimation of the mode. *The Annals of Mathematical Statistics*, 38(5):1446–1455.

Wu, D., Wang, L., and Zhang, P. (2019). Solving statistical mechanics using variational autoregressive networks. *Physical Review Letters*, 122(8):080602.

Xie, Z. and Yan, J. (2008). Kernel density estimation of traffic accidents in a network space. *Computers, Environment and Urban Systems*, 32(5):396–406.

Supplementary material

S1 Theory details

S1.1 Notation and setup

We work on a domain $\mathcal{S} \subset \mathbb{R}^d$, with finite Lebesgue measure. Bold symbols denote vectors/matrices (e.g., $\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$).

Given a measurable $f : \mathcal{S} \to \mathbb{R}$ and a temperature $\omega > 0$, define a normalized Gibbs measure as

$$g_\omega(d\boldsymbol{x}) = \frac{\exp(f(\boldsymbol{x})/\omega)}{Z_\omega(f)} d\boldsymbol{x},$$

with partition function

$$Z_\omega(f) = \int_{\mathcal{S}} \exp(f(\boldsymbol{y})/\omega) d\boldsymbol{y}$$

assumed finite when invoked.

We approximate g_ω on $\mathcal{S}$ with a tractable q (restricted to $\mathcal{S}$): (i) a single Gaussian $q_{\boldsymbol{\lambda}} = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, (ii) a K-component Gaussian mixture

$$q_{\boldsymbol{\Lambda}}(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \qquad \sum_{k=1}^{K} \pi_k = 1, \pi_k > 0.$$

Mixture weights are parameterized by logits $v_k = \log(\pi_k/\pi_K)$. Unless stated otherwise, covariance eigenvalues are constrained to $[\sigma^2_{\min}, \sigma^2_{\max}]$.

For any density q, $\mathbb{E}_q[\cdot]$ denotes the expectation under q and $\mathrm{KL}(q\|p)$ the Kullback–Leibler divergence. For q on $\mathcal{S}$, define the entropy on $\mathcal{S}$:

$$\mathcal{H}_{\mathcal{S}}(q) = -\int_{\mathcal{S}} q(\boldsymbol{x}) \log q(\boldsymbol{x}) d\boldsymbol{x},$$

and the variational objective

$$\mathcal{L}_\omega(q) = \mathbb{E}_q[f(\boldsymbol{X})] + \omega \mathcal{H}_{\mathcal{S}}(q).$$

By the classical variational identity,

$$\mathcal{L}_\omega(q) = \omega \log Z_\omega(f) - \omega \mathrm{KL}\left(q || g_\omega\right),$$

so g_ω uniquely maximizes $\mathcal{L}_\omega$ whenever $Z_\omega(f) < \infty$.

For variational parameters $\boldsymbol{\vartheta}$, let $\boldsymbol{F}(\boldsymbol{\vartheta})$ be the Fisher information matrix. The natural gradient is

$$\widetilde{\nabla}_{\boldsymbol{\vartheta}} \mathcal{L}_\omega(\boldsymbol{\vartheta}) = \boldsymbol{F}(\boldsymbol{\vartheta})^{-1} \nabla_{\boldsymbol{\vartheta}} \mathcal{L}_\omega(\boldsymbol{\vartheta}).$$

If $q_{\boldsymbol{\vartheta}}$ is in an exponential family with sufficient statistic $\boldsymbol{T}(\boldsymbol{X})$ (or in an MCEF for mixtures (Lin et al., 2019)), the natural gradient equals the standard gradient with respect to the expectation parameters $\boldsymbol{M} = \mathbb{E}_{q_{\boldsymbol{\vartheta}}}[\boldsymbol{T}(\boldsymbol{X})]$:

$$\widetilde{\nabla}_{\boldsymbol{\vartheta}} \mathcal{L}_\omega(\boldsymbol{\vartheta}) = \nabla_{\boldsymbol{M}} \mathcal{L}_\omega(\boldsymbol{\vartheta})] \ .$$

Let $\{\boldsymbol{x}_i^\star\}_{i=1}^I$ denote the set of global maximizers ("modes") of f (or of p when $f = p$) on $\mathcal{S}$. At a nondegenerate mode $\boldsymbol{x}_i^\star$ (i.e. $\nabla^2 f(\boldsymbol{x}_i^\star) \prec 0$), define the positive-definite curvature

$$\boldsymbol{H}_i := -\nabla^2 f(\boldsymbol{x}_i^\star) \succ 0.$$

For symmetric matrix A, we write $\mathrm{eig}_{\min}(A), \mathrm{eig}_{\max}(A)$ for extremal eigenvalues. Norms are Euclidean for vectors and operator norms for matrices unless noted.

For a set $\mathcal{A} \subset \mathbb{R}^d$, $\mathrm{int}(\mathcal{A})$ denotes its interior, $\partial\mathcal{A}$ its boundary, and $\mathcal{A}^c := \mathbb{R}^d \setminus \mathcal{A}$ its complement. For an element $\boldsymbol{x} \in \mathbb{R}^d$, we write $\mathrm{dist}(\boldsymbol{x}, \mathcal{A}) := \inf_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{x} - \boldsymbol{a}\|_2$. For $r > 0$, let $B(\boldsymbol{x}, r)$ denote the ball centered on $\boldsymbol{x}$ with radius r.

S1.2 Truncation equivalence and compact sets

In the optimization setting (Le Minh et al., 2025), the target f is known or directly evaluable, so $\mathbb{E}_{q_{\boldsymbol{\vartheta}}}[f(\boldsymbol{x})]$ is computable. In GERVE, f is replaced by an unknown density p and we observe only i.i.d. samples $\boldsymbol{X}_{1:N} \sim p$. We suppose that p is smooth, bounded, and supported on a bounded domain $\mathrm{supp}(p) \subset \mathcal{S}$.

Using symmetry,

$$\mathcal{L}_\omega(q_{\boldsymbol{\vartheta}}) = \int_{\mathcal{S}} p(\boldsymbol{x}) q_{\boldsymbol{\vartheta}}(\boldsymbol{x}) d\boldsymbol{x} + \omega \mathcal{H}_{\mathcal{S}}(q_{\boldsymbol{\vartheta}}) = \mathbb{E}_p\left[q_{\boldsymbol{\vartheta}}(\boldsymbol{X})\right] + \omega \mathcal{H}_{\mathcal{S}}(q_{\boldsymbol{\vartheta}}),$$

so the first term is an expectation under p. Let $P_N := \frac{1}{N}\sum_{i=1}^N \delta_{\boldsymbol{X}_i}$ be the empirical measure. A natural plug-in estimator replaces $\mathbb{E}_p[q_{\boldsymbol{\vartheta}}(\boldsymbol{X})]$ by $\mathbb{E}_{P_N}[q_{\boldsymbol{\vartheta}}(\boldsymbol{X})]$, yielding the empirical variational objective

$$\widehat{\mathcal{L}}_{\omega,N}(\boldsymbol{\vartheta}) = \mathbb{E}_{P_N}\left[q_{\boldsymbol{\vartheta}}(\boldsymbol{X})\right] + \omega \mathcal{H}_{\mathcal{S}}(q_{\boldsymbol{\vartheta}}) = \frac{1}{N}\sum_{i=1}^N q_{\boldsymbol{\vartheta}}(\boldsymbol{X}_i) + \omega \mathcal{H}_{\mathcal{S}}(q_{\boldsymbol{\vartheta}}). \tag{S1}$$

This is the learning objective optimized by GERVE.

Optimization on untruncated Gaussians. Since p is supported on $\mathcal{S}$ rather than on $\mathbb{R}^d$, the corresponding Gibbs measure is defined only on $\mathcal{S}$, and one might a priori restrict q to $\mathcal{S}$. $q_{\boldsymbol{\lambda}}$ denotes the untruncated Gaussian on $\mathbb{R}^d$ with natural parameter $\boldsymbol{\lambda}$. Define

$$\psi_{\mathcal{S}}(\boldsymbol{\lambda}) := \int_{\mathcal{S}} q_{\boldsymbol{\lambda}}(\boldsymbol{x}) d\boldsymbol{x}.$$

The $\mathcal{S}$-truncated version is $q_{\boldsymbol{\lambda}}^{\mathcal{S}}(\boldsymbol{x}) := q_{\boldsymbol{\lambda}}(\boldsymbol{x}) \mathbb{1}_{\mathcal{S}}(\boldsymbol{x}) / \psi_{\mathcal{S}}(\boldsymbol{\lambda})$.

Optimizing directly over $q_{\boldsymbol{\lambda}}^{\mathcal{S}}$ is cumbersome because $\psi_{\mathcal{S}}(\boldsymbol{\lambda})$ varies with the parameter. The next result (proven in Section S2.5.1) shows that on a large $\mathcal{S}$, optimizing over the untruncated family while integrating over $\mathcal{S}$ incurs only a controlled error on a compact set.

To define such a compact parameter space for Gaussians, fix a mean bound $\boldsymbol{\mu}_{\max} < \infty$ and eigenvalue bounds $0 < \sigma_{\min}^2 < \sigma_{\max}^2 < \infty$, and consider the compact set of natural parameters

$$\Theta := \Big\{ \boldsymbol{\lambda} = (\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}, -\frac{1}{2}\boldsymbol{\Sigma}^{-1}) : \boldsymbol{\mu} \in \mathbb{R}^d, \sigma_{\min}^2 \boldsymbol{I} \preceq \boldsymbol{\Sigma} \preceq \sigma_{\max}^2 \boldsymbol{I}, \|\boldsymbol{\mu}\|_\infty \le \boldsymbol{\mu}_{\max} \Big\}.$$

For $\boldsymbol{\lambda} \in \Theta$, let $\boldsymbol{\mu}(\boldsymbol{\lambda})$ and $\boldsymbol{\Sigma}(\boldsymbol{\lambda})$ be the mean and covariance matrix encoded by $\boldsymbol{\lambda}$.

For notational convenience, let

$$F_\omega(\boldsymbol{\lambda}) := \mathcal{L}_\omega(q_{\boldsymbol{\lambda}}), \qquad G_\omega(\boldsymbol{\lambda}) := \mathcal{L}_\omega(q_{\boldsymbol{\lambda}}^{\mathcal{S}}),$$

for $\boldsymbol{\lambda} \in \Theta$. Let $\varepsilon(\boldsymbol{\lambda}) := 1 - \psi_{\mathcal{S}}(\boldsymbol{\lambda})$, $\varepsilon_{\max} := \sup_{\boldsymbol{\lambda} \in \Theta} \varepsilon(\boldsymbol{\lambda}) < 1$, and $p_{\max} = \sup_{\boldsymbol{x} \in \mathcal{S}} p(\boldsymbol{x})$. In the following proposition, we bound $|G_\omega(\boldsymbol{\lambda}) - F_\omega(\boldsymbol{\lambda})|$ uniformly on Θ.

Proposition S1.1 (Truncation gap)**.** *Let* $C_q = \sup_{\boldsymbol{\lambda} \in \Theta} \sup_{\boldsymbol{x} \in \mathcal{S}} \big| \log q_{\boldsymbol{\lambda}}(\boldsymbol{x}) \big| < \infty$. *Define, for* $t \in [0, 1)$,

$$\delta_\omega(t) := p_{\max} t + \omega \big| \log(1 - t) \big| + \omega C_q t.$$

Then for all $\boldsymbol{\lambda} \in \Theta$,

$$\big| G_\omega(\boldsymbol{\lambda}) - F_\omega(\boldsymbol{\lambda}) \big| \le \delta_\omega\big(\varepsilon(\boldsymbol{\lambda})\big), \tag{S2}$$

and

$$\sup_{\boldsymbol{\lambda} \in \Theta} \big| G_\omega(\boldsymbol{\lambda}) - F_\omega(\boldsymbol{\lambda}) \big| \le \delta_\omega(\varepsilon_{\max}).$$

This bound implies that if $\boldsymbol{\mu}$ stays away from $\partial\mathcal{S}$, the error from using untruncated Gaussians is negligible. Now, we derive a tail bound and, combined with an argmax stability argument, we obtain the next theorem, proven in Section S2.5.2.

Theorem S1.2 (Truncation equivalence on a bounded support)**.** *Let* $\delta_\omega(\cdot)$ *be as in Proposition S1.1. For any fixed* $\omega > 0$, *if there is a margin* $M_0 > 0$ *such that* $\operatorname{dist}(\boldsymbol{\mu}(\boldsymbol{\lambda}), \partial\mathcal{S}) \ge M_0 \sqrt{\operatorname{eig}_{\max}(\boldsymbol{\Sigma}(\boldsymbol{\lambda}))}$ *for all* $\boldsymbol{\lambda} \in \Theta$, *then for some* $C > 0$,

$$\varepsilon_{\max} \le C e^{-M_0^2/2}, \qquad \sup_{\boldsymbol{\lambda} \in \Theta} \big| G_\omega(\boldsymbol{\lambda}) - F_\omega(\boldsymbol{\lambda}) \big| \le \delta_\omega(\varepsilon_{\max}).$$

Moreover, let $m_\omega(\rho) := F_\omega(\boldsymbol{\lambda}^\star_\omega) - \sup_{\|\boldsymbol{\lambda}-\boldsymbol{\lambda}^\star_\omega\|\geq\rho} F_\omega(\boldsymbol{\lambda})$ *be the value gap. Consequently, if* $m_\omega(\rho) > 0$ *for some* $\rho > 0$*, then every maximizer of* G_ω *lies in* $B(\boldsymbol{\lambda}^\star_\omega, \rho)$ *whenever* $\delta_\omega(\varepsilon_{\max}) < m(\rho)/3$*. In particular, if the maximizer of* F_ω *is unique and in* $\mathrm{int}(\Theta)$*, then under the same conditions, the maximizer of* G_ω *is also unique and* $\arg\max_{\boldsymbol{\lambda}\in\Theta} G_\omega = \boldsymbol{\lambda}^\star_\omega$.

Case of mixtures. Theorem S1.2 is stated for single Gaussians. For mixtures, we have to apply Proposition S1.1 componentwise under the same spectral bounds and control the entropy term via per-component bounds. Uniform control of the mixture log-density $\log q_{\boldsymbol{\Lambda}}$ additionally requires avoiding degenerate behaviors (e.g., vanishing weights with unbounded parameters), which can be done by staying on a compact set of "natural" parameters.

Indeed, although mixtures are not an exponential family, they lie in the minimal conditional exponential family (MCEF; Lin et al., 2019), which admits analogous natural representation. We consider the compact set given by equation (4), that we recall here:

$$\Theta_K := \Big\{\boldsymbol{\Lambda} = (v_1, \ldots, v_{K-1}, \boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_K) : |v_k| \leq v_{\max}, \boldsymbol{\lambda}_k \in \Theta\Big\},$$

where $v_k = \log(\pi_k/\pi_K)$. To preserve the compactness induced by Θ, we impose additional box constraints $|v_k| \leq v_{\max}$ on the log-ratios, which implies a uniform floor on the weights:

$$\pi_k \geq \frac{1}{1 + (K-1)e^{2v_{\max}}} > 0 \qquad \text{for all } k \in [K].$$

However, choosing

$$v_{\max} \geq \frac{1}{2}\log\Big(\frac{1}{\varepsilon} - (K-1)\Big) \tag{S3}$$

allows a designated subset of components to have weights as small as ε simultaneously while respecting the box constraints. As $v_{\max} \to \infty$, the set of feasible weights approaches the simplex, so the constraint becomes non-restrictive in practice. Typically, setting $v_{\max} = 6$ is already sufficient to allow all weights to be smaller than 10^{-5}.

In brief, for mixtures, we follow the same convention as above: we optimize over untruncated mixtures while integrating over $\mathcal{S}$. Under the margin condition $\varepsilon(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = o(1)$ (e.g., $\mathrm{dist}(\boldsymbol{\mu}_k, \partial\mathcal{S}) \geq M_0\sqrt{\mathrm{eig}_{\max}(\boldsymbol{\Sigma}_k)}$ with M_0 large), replacing truncated by untruncated components perturbs $\widehat{\mathcal{L}}_{\omega,N}$ and $\mathcal{L}_\omega$ by $o(1)$ uniformly, and so leaves maximizers unchanged asymptotically.

S1.3 Modal basins-responsibility cells agreement

To investigate the link between GERVE's induced clustering and modal clustering, we compare two partitions near each mode $\boldsymbol{x}^\star_i$ of a $\mathcal{C}^2$ density: the responsibility cells $A_k(\theta)$ of an isotropic Gaussian mixture $q_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{k=1}^K \pi_k q_{\boldsymbol{\mu}_k,\sigma^2}(x)$, with $\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)$, and

the modal basins $\{\mathcal{B}_i\}$ of p. We give simple, local conditions under which they agree on inner neighborhoods of the modes.

Let $\{\boldsymbol{x}_i^\star\}_{i=1}^I$ be isolated modes of p in $\mathrm{supp}(p) \subset \mathcal{S}$. For $r > 0$, write $B_i(r) := B(\boldsymbol{x}_i^\star, r)$ the ball centered on $\boldsymbol{x}_i^\star$ and with radius r, and $\tilde{r} := r/2$. Weights satisfy $\pi_k \in [\pi_{\min}, \pi_{\max}]$ with $\pi_{\min} > 0$, $\sum_k \pi_k = 1$. For all $k \in [K]$, define posterior component responsibilities $\mathsf{resp}_k(\boldsymbol{x}; \boldsymbol{\theta}) = \dfrac{\pi_k q_{\boldsymbol{\mu}_k, \sigma^2 I}(\boldsymbol{x})}{\sum_j \pi_j q_{\boldsymbol{\mu}_j, \sigma^2 I}(\boldsymbol{x})}$ and cells $\mathcal{A}_k(\boldsymbol{\theta}) = \{\boldsymbol{x} : \mathsf{resp}_k(\boldsymbol{x}; \boldsymbol{\theta}) \geq \mathsf{resp}_j(\boldsymbol{x}; \boldsymbol{\theta}), \forall j \in [K]\}$. For $i \in [I]$, the modal basin $\mathcal{B}_i$ is the set of all initial points $\boldsymbol{x}$ such that the solution $\widetilde{\boldsymbol{x}}(t)$ of the gradient-ascent flow $\dot{\boldsymbol{x}} = \nabla p(\boldsymbol{x})$ with $\widetilde{\boldsymbol{x}}(0) = \boldsymbol{x}$ satisfies $\widetilde{\boldsymbol{x}}(t) \to \boldsymbol{x}_i^\star$ as $t \to \infty$

Assumption S1.3 (Local strong log-concavity and separation)**.** *For each i there exists $r_i > 0$ and $\kappa_i > 0$ such that $-\nabla^2 \log p(\boldsymbol{x}) \succeq \kappa_i \boldsymbol{I}$ for all $\boldsymbol{x} \in B_i(r_i)$, and the closed balls $\overline{B_i(r_i)} \subset \mathrm{supp}(p)$ are pairwise disjoint.*

Lemma S1.4 (Basin containment)**.** *Under Assumption S1.3, $B_i(r_i) \subset \mathcal{B}_i$ for each i.*

Proof of Lemma S1.4. Let $\widetilde{\boldsymbol{x}}(t)$ be a trajectory of $\dot{\boldsymbol{x}} = \nabla p(\boldsymbol{x})$ with $\widetilde{\boldsymbol{x}}(0) \in B_i(r_i)$. Choose the Lyapunov function $V(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{x}_i^\star\|^2/2$.

On $\mathrm{supp}(p)$, using $\dot{\boldsymbol{x}} = \nabla p(\boldsymbol{x}) = p(\boldsymbol{x}) \nabla \log p(\boldsymbol{x})$, we have

$$\dot{V}(\boldsymbol{x}) = \langle \dot{\boldsymbol{x}}, \boldsymbol{x} - \boldsymbol{x}_i^\star \rangle = p(\boldsymbol{x}) \langle \nabla \log p(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{x}_i^\star \rangle.$$

Since $\nabla \log p(\boldsymbol{x}_i^\star) = \nabla p(\boldsymbol{x}_i^\star) = 0$, strong log-concavity (Assumption S1.3) implies, for $\boldsymbol{x} \in B_i(r_i) \setminus \{\boldsymbol{x}_i^\star\}$,

$$\langle \nabla \log p(\boldsymbol{x}), \boldsymbol{x} - \boldsymbol{x}_i^\star \rangle \leq -\kappa_i \|\boldsymbol{x} - \boldsymbol{x}_i^\star\|^2 < 0,$$

thus, with $m_i = \inf_{\boldsymbol{x} \in \overline{B_i(r_i)}} p(\boldsymbol{x}) > 0$,

$$\dot{V}(\boldsymbol{x}) \leq -2\kappa_i p(\boldsymbol{x}) V(\boldsymbol{x}) \leq -2\kappa_i m_i V(\boldsymbol{x}).$$

By Grönwall's inequality, we have

$$V(\widetilde{\boldsymbol{x}}(t)) \leq V(\widetilde{\boldsymbol{x}}(0)) \exp(-2\kappa_i m_i t), \tag{S4}$$

as long as $\widetilde{\boldsymbol{x}}(t)$ is in $B_i(r_i)$.

Finally, we prove by contradiction that $\widetilde{\boldsymbol{x}}(t) \in B_i(r_i)$ for all $t \geq 0$. Suppose the existence of $T_{\mathrm{out}} = \inf_{t>0}\{t : \widetilde{\boldsymbol{x}}(t) \notin B_i(r_i)\} < \infty$. In this case, by continuity of $\boldsymbol{x}$, we have $\|\widetilde{\boldsymbol{x}}(T_{\mathrm{out}}) - \boldsymbol{x}_i^\star\| = r_i$.

But since $V(\widetilde{\boldsymbol{x}}(0)) < r_i^2/2$, (S4) gives for $t \in [0, T_{\mathrm{out}})$,

$$V(\widetilde{\boldsymbol{x}}(t)) < \frac{r_i^2}{2} \exp(-2\kappa_i m_i t) < \frac{r_i^2}{2}.$$

This implies $\|\widetilde{\boldsymbol{x}}(t) - \boldsymbol{x}_i^\star\| < r_i$ for all $t \in [0, T_{\mathrm{out}})$, and this contradicts $\|\widetilde{\boldsymbol{x}}(T_{\mathrm{out}}) - \boldsymbol{x}_i^\star\| = r_i$.

Thus, we have $\widetilde{\boldsymbol{x}}(t) \in B_i(r_i)$ for all $t \geq 0$ and (S4) ensures $V(\widetilde{\boldsymbol{x}}(t)) \to 0$ as $t \to \infty$, hence $\widetilde{\boldsymbol{x}}(t) \to \boldsymbol{x}_i^\star$ and we conclude that $\widetilde{\boldsymbol{x}}(0) \in \mathcal{B}_i$. □

Assumption S1.5 (One-per-mode configuration)**.** *There is a configuration* $\boldsymbol{\theta} = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_I)$ *with exactly one center* $\boldsymbol{\mu}_i \in B_i(\tilde{r}_i)$ *for each* i*, and no other centers in* $\bigcup_{i=1}^{I} B_i(r_i)$*. Therefore, by convention, we relabel components so that center* i *lies in* $B_i(\tilde{r}_i)$ *and write* $\mathcal{A}_i(\boldsymbol{\theta})$ *for the* i*th cell.*

Assumption S1.6 (Local margin)**.** *There exists* $\Delta > 0$ *such that for all* i *and all* $\boldsymbol{x} \in B_i(\tilde{r}_i)$*,*

$$\|\boldsymbol{x} - \boldsymbol{\mu}_j\| \geq \|\boldsymbol{x} - \boldsymbol{\mu}_i\| + \Delta \qquad \forall j \neq i.$$

Remark 1. If $\boldsymbol{\mu}_i \in B_i(\tilde{r}_i)$ and $\operatorname{dist}\big(B_i(\tilde{r}_i), \{\boldsymbol{\mu}_j\}_{j\neq i}\big) > 0$, then $\Delta := \inf_{\boldsymbol{x}\in B_i(\tilde{r}_i)} \min_{j\neq i}(\|\boldsymbol{x} - \boldsymbol{\mu}_j\| - \|\boldsymbol{x} - \boldsymbol{\mu}_i\|) > 0$.

Assumption S1.7 (Scale separation)**.** *There exists* $\varepsilon \in (0, 1)$ *such that*

$$\sigma \leq \frac{\Delta}{\sqrt{2\log \Psi(\varepsilon)}}, \qquad \textit{where } \Psi(\varepsilon) := \frac{(I-1)\pi_{\max}}{\varepsilon \pi_{\min}} > 1.$$

Theorem S1.8 (Responsibility on balls)**.** *Under Assumptions S1.5-S1.7, for each* i*,*

$$B_i(\tilde{r}_i) \subset \mathcal{A}_i(\boldsymbol{\theta}), \qquad \textit{and} \quad r_i(\boldsymbol{x}; \boldsymbol{\theta}) \geq (1+\varepsilon)^{-1}, \quad \forall \boldsymbol{x} \in B_i(\tilde{r}_i).$$

Proof of Theorem S1.8. Fix i and let $\boldsymbol{x} \in B_i(\tilde{r}_i)$. By Assumption S1.6, $\|\boldsymbol{x} - \boldsymbol{\mu}_j\| - \|\boldsymbol{x} - \boldsymbol{\mu}_i\| \geq \Delta$ and $\|\boldsymbol{x} - \boldsymbol{\mu}_j\| + \|\boldsymbol{x} - \boldsymbol{\mu}_i\| \geq 2\|\boldsymbol{x} - \boldsymbol{\mu}_i\| + \Delta$. Combining with Assumption S1.7, this gives

$$\frac{q_{\boldsymbol{\mu}_j, \sigma^2 \boldsymbol{I}}(\boldsymbol{x})}{q_{\boldsymbol{\mu}_i, \sigma^2 \boldsymbol{I}}(\boldsymbol{x})} = \exp\Big(-\frac{(\|\boldsymbol{x} - \boldsymbol{\mu}_j\| - \|\boldsymbol{x} - \boldsymbol{\mu}_i\|)(\|\boldsymbol{x} - \boldsymbol{\mu}_j\| + \|\boldsymbol{x} - \boldsymbol{\mu}_i\|)}{2\sigma^2}\Big) \leq \exp\Big(-\frac{\Delta^2}{2\sigma^2}\Big) \leq \frac{1}{\Psi(\varepsilon)}.$$

Summing with weights gives $\sum_{j\neq i} \frac{\pi_j q_{\boldsymbol{\mu}_j}}{\pi_i q_{\boldsymbol{\mu}_i}} \leq \frac{(I-1)\pi_{\max}}{\pi_{\min}\Psi(\varepsilon)} = \varepsilon$, hence $\mathsf{resp}_i(\boldsymbol{x}; \boldsymbol{\theta}) \geq (1+\varepsilon)^{-1} > 1/2$, which means $\boldsymbol{x} \in \mathcal{A}_i(\boldsymbol{\theta})$. □

Remark 2. Assumption S1.7 is in fact a sufficient condition for $\mathsf{resp}_i(\boldsymbol{x}; \boldsymbol{\theta}) \geq (1+\varepsilon)^{-1}$, which means competing components contribute a total probability mass $\leq \varepsilon$ relative to the dominant component i.

Corollary S1.9 (Local partition agreement)**.** *Under Assumptions S1.3-S1.7, for each* i*,*

$$\mathcal{A}_i(\boldsymbol{\theta}) \cap B_i(\tilde{r}_i) = \mathcal{B}_i \cap B_i(\tilde{r}_i) = B_i(\tilde{r}_i).$$

Proof of Corollary S1.9. Theorem S1.8 gives $B_i(\tilde{r}_i) \subset \mathcal{A}_i(\boldsymbol{\theta})$. Lemma S1.4 gives $B_i(\tilde{r}_i) \subset \mathcal{B}_i$. □

Agreement is guaranteed only on local neighborhoods of the modes $\bigcup_{i=1}^{I} B_i(\tilde{r}_i)$. In general, the responsibility cells $\{\mathcal{A}_k\}$ and the modal basins $\{\mathcal{B}_i\}$ globally disagree. Figure S1 showcases a counterexample, highlighting the differences between modal clustering and high-responsibility assignments in a Gaussian mixture model. Remarkably, a mode may not even be in its corresponding component responsibility region, notably when components significantly overlap.

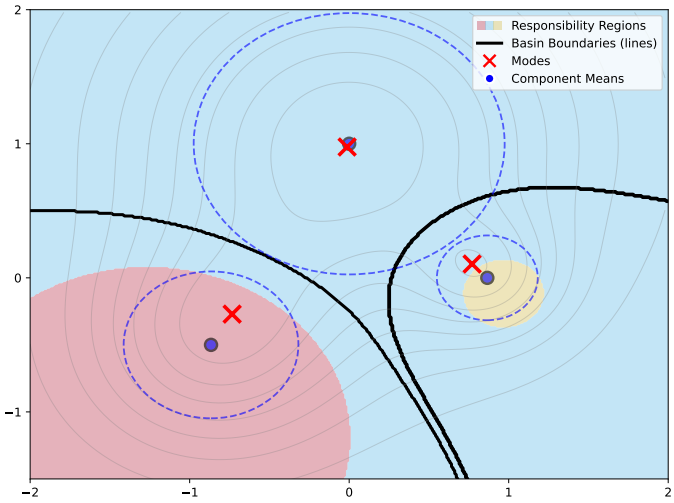

Figure S1: Responsibility regions (in color resp. blue, pink, yellow) and (approximate) modal basins for a Gaussian mixture with 3 components, means at $\boldsymbol{\mu}_1 = \big(-\cos(\pi/6), -0.5\big)$, $\boldsymbol{\mu}_2 = (0, 1)$, $\boldsymbol{\mu}_3 = \big(\cos(\pi/6), 0\big)$, mixture weights $\boldsymbol{\pi} = (0.16, 0.80, 0.04)$ and isotropic covariances $\boldsymbol{\Sigma}_k = \sigma_k^2 \boldsymbol{I}$ with variances $(\sigma_1^2, \sigma_2^2, \sigma_3^2) = (0.30, 0.95, 0.10)$. Covariances are represented as circles of radii σ_k.

S1.4 GERVE's natural-gradient convergence

Under the Robbins–Monro and bounded iterate conditions, we can show that GERVE's stochastic natural-gradient ascent (SNGA) is consistent. Convergence is obtained via a direct application of the ODE method for stochastic approximation (Borkar, 2008, Chp. 2.1).

Assumption S1.10. *Consider the update*

$$\boldsymbol{\vartheta}_{t+1} = \boldsymbol{\vartheta}_t + \rho_t \boldsymbol{\psi}_t, \tag{S5}$$

where $\boldsymbol{\psi}_t$ is an unbiased stochastic estimate of the natural gradient $\widetilde{\nabla}\widehat{\mathcal{L}}_{\omega,N}(\boldsymbol{\vartheta}_t)$ (equivalently, the gradient in the expectation-parameter space for exponential families or mixtures thereof). Let $\mathcal{F}_t := \sigma(\boldsymbol{\vartheta}_1, \boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_t)$.

Assume:

*(C1) (**Robbins–Monro**) $\sum_t \rho_t = \infty$, $\sum_t \rho_t^2 < \infty$.*

*(C2) (**Mean field**) $\widetilde{\nabla}\widehat{\mathcal{L}}_{\omega,N}$ is locally Lipschitz and of at most linear growth.*

*(C3) (**Noise**) $\mathbb{E}[\boldsymbol{\psi}_t \mid \mathcal{F}_{t-1}] = \widetilde{\nabla}\widehat{\mathcal{L}}_{\omega,N}(\boldsymbol{\vartheta}_{t-1})$ and $\mathbb{E}\|\boldsymbol{\psi}_t\|^2 \leq C(1 + \|\boldsymbol{\vartheta}_{t-1}\|^2)$.*

*(C4) (**Stability**) The iterates $(\boldsymbol{\vartheta}_t)_{t\geq 1}$ are a.s. bounded.*

Theorem S1.11 (Convergence of SNGA at fixed ω)**.** *Let* $\mathcal{S}_\omega = \{\boldsymbol{\vartheta} : \widetilde{\nabla}\widehat{\mathcal{L}}_{\omega,N}(\boldsymbol{\vartheta}) = 0\}$. *Under Assumption S1.10, we have, as* $t \to \infty$*:*

$$\operatorname{dist}(\boldsymbol{\vartheta}_t, \mathcal{S}_\omega) \to 0 \quad a.s.$$

Equivalently, every limit point of $(\boldsymbol{\vartheta}_t)$ *is almost surely stationary.*

Remark 3 (Natural-gradient regularity). For exponential families and mixtures thereof (treated as MCEF models), the "natural gradient = expectation gradient" identity (eq. (5)) ensures that $\widetilde{\nabla}\widehat{\mathcal{L}}_{\omega,N}$ inherits Lipschitz continuity from $\widehat{\mathcal{L}}_{\omega,N}$ on compact Θ, without explicit inversion of the Fisher information matrix.

Remark 4 (Boundedness of iterates). In practice, we can project the iterates in a compact parameter space (Θ for single Gaussians, Θ_K for Gaussian mixtures), ensuring (C4). Typically, without lower and upper bounds on $\boldsymbol{\Sigma}$, $\widehat{\mathcal{L}}_{\omega,N}$ can be unbounded above.

Under annealing, conditions on the schedule are sufficient to guarantee convergence. The following corollary stems from standard two-timescale results for stochastic approximation (Kushner and Yin, 2003, Chp. 8.6; Borkar, 2008, Chp. 6.1).

Corollary S1.12 (Convergence of SNGA under annealing)**.** *Suppose that Under Assumption S1.10 for each fixed* $\omega > 0$*, and that* $(\boldsymbol{\vartheta}, \omega) \mapsto \widetilde{\nabla}\widehat{\mathcal{L}}_{\omega,N}(\boldsymbol{\vartheta})$ *is jointly continuous on bounded sets.*

1. Stagewise schedule. *Let* $\omega_1 > \omega_2 > \cdots \to 0$*. If at stage* j *the run with fixed* ω_j *is long enough that* $\operatorname{dist}(\boldsymbol{\vartheta}, \mathcal{S}_{\omega_j}) \le \delta_j$ *with* $\delta_j \to 0$ *(and the next stage is warm-started at that iterate), then* $\operatorname{dist}(\boldsymbol{\vartheta}, \mathcal{S}_{\omega_j}) \to 0$ *as* $j \to \infty$*.*
2. Continuous schedule. *If* $\omega_t \to 0$ *satisfies* $|\omega_{t+1} - \omega_t| = o(\rho_t)$*, then* $\operatorname{dist}(\boldsymbol{\vartheta}_t, \mathcal{S}_{\omega_t}) \to 0$ *a.s.*

S1.5 Uncertainty quantification for mode estimation

Here, we develop the theoretical guarantees of the mode-level uncertainty quantification (UQ) procedure proposed in Section 5. In this section, we mathematically translate all the notions (including Assumption 5.2), defined in Section 5.2. These will be also used in the proofs in S2.4. Then, we investigate the consistency stability scores (Prop. S1.13) and propose strategies to handle nonconvex objectives. All the results for this section and Section 5 of the main part are proven in Section S2.4.

S1.5.1 Notations and separation assumption

We start by mathematically translating all the notions defined in Section 5.2. These will be used throughout this section and in the proofs in S2.4.

Data, measures, and convergence. $\boldsymbol{X}_{1:N}$ are i.i.d. from the population law P on the sample space $\mathcal{S}$. The empirical and bootstrap empirical measures are

$$P_N = \frac{1}{N}\sum_{i=1}^{N}\delta_{\boldsymbol{X}_i}, \qquad P_N^* = \frac{1}{N}\sum_{i=1}^{N}\delta_{\boldsymbol{X}_i^*}, \quad \boldsymbol{X}_1^*,\ldots,\boldsymbol{X}_N^* \overset{\text{i.i.d.}}{\sim} P_N \text{ given } \boldsymbol{X}_{1:N}.$$

We write $P^*(\cdot) = P(\cdot|\boldsymbol{X}_{1:N})$ for conditional (bootstrap) probability. All bootstrap limits are understood *conditionally on* $\boldsymbol{X}_{1:N}$*, in* P*-probability.*

Parameter space and objective. Fix $\omega_0 > 0$ and work on the compact parameter set Θ_K (eq. (4)). For a finite signed measure G on $\mathcal{S}$, define the criterion

$$\mathcal{L}_{\omega_0}(\boldsymbol{\Lambda}; G) := \int q_{\boldsymbol{\Lambda}}(\boldsymbol{x})dG(\boldsymbol{x}) + \omega_0\mathcal{H}_{\mathcal{S}}(q_{\boldsymbol{\Lambda}}),$$

and the population/empirical versions $\mathcal{L}_{\omega_0}(\boldsymbol{\Lambda}) := \mathcal{L}_{\omega_0}(\boldsymbol{\Lambda}; P)$ and $\widehat{\mathcal{L}}_{\omega_0,N}(\boldsymbol{\Lambda}) := \mathcal{L}_{\omega_0}(\boldsymbol{\Lambda}; P_N)$.

Estimators and bootstrap refits. Let

$$\widehat{\boldsymbol{\Lambda}}_{\omega_0,N} \in \operatorname*{arg\,max}_{\boldsymbol{\Lambda}\in\Theta_K} \widehat{\mathcal{L}}_{\omega_0,N}(\boldsymbol{\Lambda}), \qquad \boldsymbol{\Lambda}^\star_{\omega_0} \in \operatorname*{arg\,max}_{\boldsymbol{\Lambda}\in\Theta_K} \mathcal{L}_{\omega_0}(\boldsymbol{\Lambda}),$$

and define the bootstrap criterion and maximizer

$$\widehat{\mathcal{L}}^*_{\omega_0,N}(\boldsymbol{\Lambda}) := \mathcal{L}_{\omega_0}(\boldsymbol{\Lambda}; P_N^*), \qquad \widehat{\boldsymbol{\Lambda}}^*_{\omega_0,N} \in \operatorname*{arg\,max}_{\boldsymbol{\Lambda}\in\Theta_K} \widehat{\mathcal{L}}^*_{\omega_0,N}(\boldsymbol{\Lambda}).$$

We also allow a (possibly inexact) estimator $\widetilde{\boldsymbol{\Lambda}}_{\omega_0,N}$ of $\widehat{\boldsymbol{\Lambda}}_{\omega_0,N}$ (e.g., the output of a finite-iteration GERVE fit).

Quadratic expansion matrices. Write

$$\boldsymbol{H}^\star_{\omega_0} := -\nabla^2_{\boldsymbol{\Lambda}}\mathcal{L}_{\omega_0}(\boldsymbol{\Lambda}^\star_{\omega_0}), \qquad \boldsymbol{V}_{\omega_0} := \operatorname{Var}\left(\nabla_{\boldsymbol{\Lambda}} q_{\boldsymbol{\Lambda}}(X)\right)\big|_{\boldsymbol{\Lambda}=\boldsymbol{\Lambda}^\star_{\omega_0}}, \qquad \boldsymbol{W}_{\omega_0} := (\boldsymbol{H}^\star_{\omega_0})^{-1}\boldsymbol{V}_{\omega_0}(\boldsymbol{H}^\star_{\omega_0})^{-1}.$$

Mode means, matching, and projections. Here, we consider fixed targets $\boldsymbol{u}_1,\ldots,\boldsymbol{u}_{K_0} \in \mathbb{R}^d$ ($K_0 \le K$). For $\boldsymbol{\Lambda} \in \Theta_K$, let $\boldsymbol{\mu}_k(\boldsymbol{\Lambda}) \in \mathbb{R}^d$ be the k-th component mean and

$$\boldsymbol{m}(\boldsymbol{\Lambda}) := \left(\boldsymbol{\mu}_1(\boldsymbol{\Lambda})^T,\ldots,\boldsymbol{\mu}_K(\boldsymbol{\Lambda})^T\right)^T \in \mathbb{R}^{Kd}.$$

Define the matching map $\mathcal{M}(\boldsymbol{\Lambda}) \in \mathbb{R}^{dK_0}$ by selecting an injection $\pi : \{1,\ldots,K_0\} \hookrightarrow \{1,\ldots,K\}$ minimizing $\sum_{j=1}^{K_0}\|\boldsymbol{\mu}_{\pi(j)}(\boldsymbol{\Lambda}) - \boldsymbol{u}_j\|$, with a fixed lexicographic tie-break:

$$\mathcal{M}(\boldsymbol{\Lambda}) = \left(\boldsymbol{\mu}_{\pi(1)}(\boldsymbol{\Lambda})^T,\ldots,\boldsymbol{\mu}_{\pi(K_0)}(\boldsymbol{\Lambda})^T\right)^T.$$

Let $\boldsymbol{E}_j \in \mathbb{R}^{d\times dK_0}$ extract the j-th d-block and set $\mathcal{M}_j(\boldsymbol{\Lambda}) := \boldsymbol{E}_j\mathcal{M}(\boldsymbol{\Lambda}) \in \mathbb{R}^d$.

Local separation and linearization. For $j \in [K_0]$ and $\boldsymbol{\Lambda} \in \Theta_K$, let $a(j, \boldsymbol{\Lambda}) \in \arg\min_k \|\boldsymbol{\mu}_k(\boldsymbol{\Lambda}) - \boldsymbol{u}_j\|$ (lexicographic tie-break) and define

$$\mathsf{gap}_j(\boldsymbol{\Lambda}) := \min_{k \neq a(j,\boldsymbol{\Lambda})} \Big(\|\boldsymbol{\mu}_k(\boldsymbol{\Lambda}) - \boldsymbol{u}_j\| - \|\boldsymbol{\mu}_{a(j,\boldsymbol{\Lambda})}(\boldsymbol{\Lambda}) - \boldsymbol{u}_j\| \Big), \quad \Delta(\boldsymbol{\Lambda}) := \min_{1 \leq j \leq K_0} \mathsf{gap}_j(\boldsymbol{\Lambda}).$$

For $\delta > 0$, set $\mathcal{G}_\delta := \{\boldsymbol{\Lambda} \in \Theta_K : \Delta(\boldsymbol{\Lambda}) \geq \delta\}$. Local separation (Assumption 5.2) holds at $\boldsymbol{\Lambda}^\star_{\omega_0}$ if there exists $\delta > 0$ such that $\Delta(\boldsymbol{\Lambda}^\star_{\omega_0}) \geq \delta$. In this case, $\mathcal{M}$ is $\mathcal{C}^1$ at $\boldsymbol{\Lambda}^\star_{\omega_0}$ (Prop. S2.17) with Jacobian $\boldsymbol{J} := \nabla_{\boldsymbol{\Lambda}} \mathcal{M}(\boldsymbol{\Lambda}^\star_{\omega_0})$, and we define

$$\boldsymbol{C}_M := \boldsymbol{J} \boldsymbol{W}_{\omega_0} \boldsymbol{J}^T, \qquad \boldsymbol{C}_j := \boldsymbol{E}_j \boldsymbol{C}_{\mathcal{M}} \boldsymbol{E}_j^T.$$

S1.5.2 Consistency of stability scores

To investigate the properties of the stability scores, we write their mathematical definition. Let $\Phi(G)$ be a deterministic measurable selection from the (possibly set-valued) argmax

$$\Phi(G) \in \arg\max_{\boldsymbol{\Lambda} \in \Theta_K} \Big\{ \int q_{\boldsymbol{\Lambda}} dG + \omega_0 \mathcal{H}_{\mathcal{S}}(q_{\boldsymbol{\Lambda}}) \Big\}.$$

Fix disjoint neighborhoods $U_1, \ldots, U_{K_0}$ of $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{K_0}$. Define $F_j(G) = \mathbb{1}\{\boldsymbol{E}_j \mathcal{M}(\Phi(G)) \in U_j\}$ and the targets

$$\pi_{j,N} := P\big(F_j(P_N) = 1\big), \qquad \pi_j := \lim_{N \to \infty} \pi_{j,N}.$$

The bootstrap stability score defined in Section 5 can be reformulated $s_j = \frac{1}{L} \sum_{\ell=1}^{L} F_j(P_N^{*(\ell)})$. In other words,

$$s_j = \frac{1}{L} \sum_{\ell=1}^{L} \mathbb{1}\{\exists \text{ a matched component in } U_j \text{ for the } \ell\text{-th bootstrap fit}\}.$$

Let $\mathbb{E}^*$ and Var^* denote the conditional expectation and variance given $\boldsymbol{X}_{1:N}$. The following proposition, proven in Section S2.4.7, establishes statistical guarantees on the stability scores.

Proposition S1.13 (Consistency of stability scores)**.** *Let us index the stability scores $s_{j,L}$ by L. Under Assumptions S2.10 and 5.2 and with U_j chosen to have a positive margin at $\boldsymbol{\Lambda}^\star_{\omega_0}$ (Lemma S2.15), define*

$$\tau_{j,N} := \mathbb{E}^*\big[F_j(P_N^*)\big] = \mathbb{E}^*[s_{j,L}] \qquad \textit{(for all } L\textit{)}.$$

Then:

(i) ***(Conditional LLN).*** *For each N, conditionally on the data,*

$$s_{j,L} \xrightarrow{a.s.} \tau_{j,N} \quad \text{as } L \to \infty, \qquad \mathrm{Var}^*(s_{j,L}) = \mathrm{Var}^*\big(F_j(P_N^*)\big)/L \to 0.$$

In particular, for any $\epsilon > 0$,

$$P^*\big(|s_{j,L} - \tau_{j,N}| > \epsilon\big) \le 2\exp\big(-2L\epsilon^2\big).$$

(ii) ***(Large-N limit).*** *As $N \to \infty$,*

$$\tau_{j,N} \xrightarrow{P} \pi_j, \qquad \pi_{j,N} := P\big(F_j(P_N) = 1\big) \to \pi_j,$$

where $\pi_j := F_j(P) \in \{0, 1\}$ (equivalently, $\pi_j = P(F_j(P) = 1)$).

Consequently, for any sequence $L_N \to \infty$ as $N \to \infty$, we have $s_{j,L_N} \xrightarrow{P} \pi_j$.

S1.5.3 Handling local optima

Although $\mathcal{L}_{\omega_0}$ is nonconvex in general, in practice we use the following guardrails to mitigate the issue:

1. *Annealing + overcompleteness.* We use an overcomplete model, and we start from a higher smoothing level and continue down to ω_0. Small components are pruned. This reliably enlarges the global basin explored by GERVE.
2. *Stability scores.* Per-mode stability scores s_j are monitored, with values near 1 indicating robust recovery. For example, as a diagnostic, we can flag $s_j < 0.5$ as potentially problematic and $s_j < 0.3$ as likely unstable.
3. *Multiple initializations.* We can run several random starts at ω_0 and keep the best solution (by objective).

S2 Proofs

S2.1 Preliminaries: Empirical process properties

In this section, we first prove that the classes of Gaussian distributions and GMM with bounded covariances are Glivenko–Cantelli (GC), and admit integrable envelopes. We recall (van der Vaart and Wellner, 1996) that a class of functions $\mathcal{F}$ on $\mathbb{R}^d$ is Glivenko–Cantelli (GC) with respect to a probability measure P if, for i.i.d. samples $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ of a random variable $\boldsymbol{X} \sim P$, the empirical average of any $f \in \mathcal{F}$, $P_N f := N^{-1}\sum_{i=1}^N f(\boldsymbol{X}_i)$, converges uniformly to its expectation $Pf = \mathbb{E}_P[f(\boldsymbol{X})]$, in probability:

$$\sup_{f\in\mathcal{F}} |P_N f - Pf| \xrightarrow{P} 0 \text{ as } N \to \infty.$$

We say that $\mathcal{F}$ admits an integrable envelope F if $F(\boldsymbol{x}) \ge |f(\boldsymbol{x})|$ for all $f \in \mathcal{F}$, with $\mathbb{E}_P[F(\boldsymbol{X})] < \infty$.

Lemma S2.1 (GC for fixed-covariance Gaussian distributions)**.** *Let* $\varphi_{\mathbf{\Sigma}_0}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; 0, \mathbf{\Sigma}_0)$ *with a given* $\mathbf{\Sigma}_0 \succ 0$*, and define the class*

$$\mathcal{F}_0 = \{f_{\boldsymbol{\mu}}(\boldsymbol{x}) = \varphi_{\mathbf{\Sigma}_0}(\boldsymbol{x} - \boldsymbol{\mu}), \boldsymbol{\mu} \in \mathbb{R}^d\}.$$

If P *is a probability measure with continuous and bounded density, then* $\mathcal{F}_0$ *is Glivenko–Cantelli and has an integrable envelope with respect to* P*.*

Proof of Lemma S2.1. Standard results on empirical processes guarantee that $\mathcal{F}_0$ is GC if it admits an envelope and is totally bounded in $L_1(P)$ (see van der Vaart and Wellner, 1996, Thm. 2.4.3). A straightforward envelope is $F(\boldsymbol{x}) := (2\pi)^{-d/2}|\mathbf{\Sigma}_0|^{-1/2} \geq f_{\boldsymbol{\mu}}(\boldsymbol{x}) \geq 0$, and $\mathbb{E}_P[F(\boldsymbol{X})] = F(\boldsymbol{x}) < \infty$. It only remains to prove the total boundedness in $L_1(P)$.

1. Denote by p the density of P. Since $\|f_{\boldsymbol{\mu}}\|_{L_1(P)} = \int f_{\boldsymbol{\mu}}(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} = \int \varphi_{\mathbf{\Sigma}_0}(\boldsymbol{x} - \boldsymbol{\mu})p(\boldsymbol{x})d\boldsymbol{x} = p * \varphi_{\mathbf{\Sigma}_0}(\boldsymbol{\mu})$, then $\|f_{\boldsymbol{\mu}}\|_{L_1(P)} \to 0$ as $\|\boldsymbol{\mu}\| \to \infty$. It follows that for any $\varepsilon > 0$, there exists some M_ϵ such that $\sup_{\|\boldsymbol{\mu}\|>M_\varepsilon} \|f_{\boldsymbol{\mu}}\|_{L_1(P)} < \varepsilon/2$.
2. On the compact ball $B(0, R) = \{\boldsymbol{x} : \|\boldsymbol{x}\| \leq R\}$, $(\boldsymbol{x}, \boldsymbol{\mu}) \mapsto f_{\boldsymbol{\mu}}(\boldsymbol{x})$ is jointly $\mathcal{C}^1$. Hence, there exists L_R with $\sup_{\|\boldsymbol{x}\|\leq R} |f_{\boldsymbol{\mu}}(\boldsymbol{x}) - f_{\boldsymbol{\nu}}(\boldsymbol{x})| \leq L_R\|\boldsymbol{\mu} - \boldsymbol{\nu}\|$. Choose R so that $2P(B(0, R)^c)\|F\|_\infty < \varepsilon/2$.
3. Cover $\{\boldsymbol{\mu} \in \mathbb{R}^d, \|\boldsymbol{\mu}\| \leq M_\varepsilon\}$ with a finite δ-net[1] $\{\boldsymbol{\mu}_j, j \in [J_\delta]\}$, where $\delta = \varepsilon/(2L_R)$. That means for all $\boldsymbol{\mu}$ such that $\|\boldsymbol{\mu}\| \leq M_\varepsilon$, there exists $j \in [J_\delta]$ such that $\|\boldsymbol{\mu}-\boldsymbol{\mu}_j\| \leq \delta$. Compute

   $$\|f_{\boldsymbol{\mu}} - f_{\boldsymbol{\mu}_j}\|_{L_1(P)} = \int_{B(0,R)} |f_{\boldsymbol{\mu}}(\boldsymbol{x}) - f_{\boldsymbol{\mu}_j}(\boldsymbol{x})|p(\boldsymbol{x})d\boldsymbol{x} + \int_{B(0,R)^c} |f_{\boldsymbol{\mu}}(\boldsymbol{x}) - f_{\boldsymbol{\mu}_j}(\boldsymbol{x})|p(\boldsymbol{x})d\boldsymbol{x}.$$

   On $B(0, R)$, $|f_{\boldsymbol{\mu}}(\boldsymbol{x}) - f_{\boldsymbol{\mu}_j}(\boldsymbol{x})| \leq L_R\|\boldsymbol{\mu} - \boldsymbol{\mu}_j\| \leq \varepsilon/2$, so the first integral is smaller than $\varepsilon/2$. On $B(0, R)^c$, the integral is smaller than $2\|F\|_\infty P(B(0, R)^c) < \varepsilon/2$. Thus, $\|f_{\boldsymbol{\mu}} - f_{\boldsymbol{\mu}_j}\|_{L_1(P)} \leq \varepsilon$.

Finally, if $\|\boldsymbol{\mu}\| > M_\varepsilon$ then by 1. $\|f_{\boldsymbol{\mu}} - 0\|_{L_1(P)} = \|f_{\boldsymbol{\mu}}\|_{L_1(P)} < \varepsilon/2$, which concludes the proof that that $\mathcal{F}_0$ is totally bounded in $L_1(P)$.

□

Lemma S2.2 (GC for bounded location-scale Gaussian distributions)**.** *Let*

$$\mathcal{F} = \{f_{\boldsymbol{\mu},\mathbf{\Sigma}}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \mathbf{\Sigma}) : \boldsymbol{\mu} \in \mathbb{R}^d, \sigma^2_{\min}\boldsymbol{I} \preceq \mathbf{\Sigma} \preceq \sigma^2_{\max}\boldsymbol{I}\}.$$

If P *is a probability measure with continuous and bounded density, then* $\mathcal{F}$ *is GC and has an integrable envelope with respect to* P*.*

Proof of Lemma S2.2. The proof is analogous to that of Lemma S2.1. An admissible envelope is $F(\boldsymbol{x}) := (2\pi\sigma^2_{\min})^{-d/2} \geq \sup_{\boldsymbol{\mu},\mathbf{\Sigma}} f_{\boldsymbol{\mu},\mathbf{\Sigma}}(\boldsymbol{x})$, and $\mathbb{E}_P[F(\boldsymbol{X})] < \infty$. Analogous arguments can be used to prove the total boundedness in $L_1(P)$:

[1] A *δ-net* for a set A in a metric space $(\mathcal{X}, d)$ is a finite subset $N_\delta \subset \mathcal{X}$ such that for every $x \in A$ there exists $y \in N_\delta$ with $d(x, y) \leq \delta$. In our case, the metric is $L_1(P)$ distance between densities.

1. For large $\|\boldsymbol{\mu}\|$, $\|f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}\|_{L_1(P)} = (p * \varphi_{\boldsymbol{\Sigma}})(\boldsymbol{\mu}) \to 0$ uniformly over $\boldsymbol{\Sigma}$ in the spectral band $[\sigma^2_{\min}, \sigma^2_{\max}]$, as $\varphi_{\boldsymbol{\Sigma}}$'s tails are controlled by $\sigma_{\max}$. Define M_ε such that in $\{\boldsymbol{\mu} : \|\boldsymbol{\mu}\| > M_\varepsilon\}$, functions are $\varepsilon/2$-close to 0 in $L_1(P)$.
2. On the compact ball $B(0,R) = \{\boldsymbol{x} : \|\boldsymbol{x}\| \le R\}$, $(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \mapsto f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x})$ is jointly $\mathcal{C}^1$. Because $\boldsymbol{\Sigma}^{-1}$ has eigenvalues in $[\sigma^{-2}_{\max}, \sigma^{-2}_{\min}]$, its derivatives are uniformly bounded, hence it is Lipschitz in $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, i.e. there exists $L_R > 0$ such that

$$\sup_{\boldsymbol{x} \in B(0,R)} |f_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\boldsymbol{x}) - f_{\boldsymbol{\mu}',\boldsymbol{\Sigma}'}(\boldsymbol{x})| \le L_R \|(\boldsymbol{\mu}, \boldsymbol{\Sigma}) - (\boldsymbol{\mu}', \boldsymbol{\Sigma}')\|.$$

3. Cover $\{\boldsymbol{\mu} : \|\boldsymbol{\mu}\| \le M_\varepsilon\} \times \{\boldsymbol{\Sigma} : \sigma^2_{\min} \boldsymbol{I} \preceq \boldsymbol{\Sigma} \preceq \sigma^2_{\max} \boldsymbol{I}\}$ with a δ-net for $\delta = \varepsilon/(2L_R)$. The $L_1(P)$-distance is small on $B(0,R)$, with the tail region $\{\boldsymbol{x} : \|\boldsymbol{x}\| > R\}$ controlled by choosing R such that $2P(B(0,R)^c)\|F\|_\infty < \varepsilon/2$.

Similarly as before, we conclude that $\mathcal{F}$ is totally bounded, hence GC. □

Lemma S2.3 (GC for finite mixtures)**.** *Fix $K \in \mathbb{N}$. Let $\mathcal{F}$ be totally bounded with respect to P and with envelope $F \in L_1(P)$. Define the class of K-component mixtures with components in $\mathcal{F}$*

$$\mathcal{F}^{(K)} = \Big\{ f(\boldsymbol{x}) = \sum_{k=1}^K \pi_k f_k(\boldsymbol{x}) : \boldsymbol{\pi} \in \Delta_K, f_k \in \mathcal{F} \Big\},$$

where Δ_K is the $K-1$-dimensional simplex. Then $\mathcal{F}^{(K)}$ is GC with envelope $KF \in L_1(P)$.

Proof of Lemma S2.3. We need to prove that KF is an integrable envelope and that $\mathcal{F}^{(K)}$ is totally bounded in $L_1(P)$. First,

$$|f(\boldsymbol{x})| = \left| \sum_{k=1}^K \pi_k f_k(\boldsymbol{x}) \right| \le \sum_{k=1}^K \pi_k |f_k(\boldsymbol{x})| \le \sum_{k=1}^K |f_k(\boldsymbol{x})| \le KF(\boldsymbol{x}),$$

hence $\mathbb{E}_P[KF] < \infty$, and KF is an envelope.

The total boundedness of $\mathcal{F}^{(K)}$ comes from the total boundedness of $\mathcal{F}$:

1. Let $\varepsilon > 0$. Since $\mathcal{F}$ is totally bounded, there exists a finite $\varepsilon/2$-net $\{g^{(1)}, \dots, g^{(J_\varepsilon)}\}$ in $L_1(P)$. For each f_k, denote by $g^{(j_k)}$ an ε-approximation. For each $g^{(j_k)}$, $\|g^{(j_k)}\|_{L_1(P)} \le \|g^{(j_k)} - f_k\|_{L_1(P)} + \|f_k\|_{L_1(P)} = \varepsilon/2 + \|F\|_{L_1(P)} = F_\varepsilon$.
2. Approximate $\pi \in \Delta_K$ using a finite grid Π_η of mesh $\eta = \varepsilon/(2KF_\varepsilon)$ and denote by $\tilde{\pi}_k$ the closest element of Π_η to π_k.
3. Compute

$$\left\| \sum_{k=1}^K \pi_k f_k - \sum_{k=1}^K \widetilde{\pi}_k g^{(j_k)} \right\|_{L_1(P)} \le \sum_{k=1}^K \pi_k \|f_k - g^{(j_k)}\|_{L_1(P)} + \sum_{k=1}^K F_\varepsilon |\pi_k - \widetilde{\pi}_k|.$$

The two terms are bounded by $\varepsilon/2$, so the sum is bounded by ε.

Thus $\mathcal{F}^{(K)}$ is totally bounded, hence GC. □

Corollary S2.4 (GC for bounded-covariance Gaussian mixtures)**.** *Let $\mathcal{G}$ be the class of Gaussian distributions with covariance eigenvalues in $[\sigma_{\min}^2, \sigma_{\max}^2]$ and unconstrained means. For fixed K, the class of K-component Gaussian mixtures with those covariance bounds is GC under P and has an envelope in $L_1(P)$.*

Proof of Corollary S2.4. In the proof of Lemma S2.2, we proved the sufficient conditions to apply Lemma S2.3 with $\mathcal{F} = \mathcal{G}$, which proves in turn the result above. □

Finally, we present the P-Donsker property. It upgrades GC to a uniform CLT, which yields rates for empirical averages over $\mathcal{F}$ and underpins our asymptotic normality results. Consider a probability measure P and some i.i.d. samples $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ of a random variable $\boldsymbol{X} \sim P$. A class of functions $\mathcal{F}$ on $\mathbb{R}^d$ is P-Donsker if the empirical process G_N defined as $G_N f \coloneqq \sqrt{N}(P_N - P)f$ for $f \in \mathcal{F}$, viewed as a random element in $\ell_\infty(\mathcal{F})$, converges in distribution to a tight mean-zero Gaussian process G_P with covariance

$$\operatorname{Cov}\big(G_P f, G_P g\big) = P\left[(f - Pf)(g - Pg)\right], \qquad f, g \in \mathcal{F}.$$

Equivalently, $\mathcal{F}$ is P-Donsker if a functional CLT holds uniformly over $f \in \mathcal{F}$. A sufficient condition is that $\mathcal{F} \subset L_2(P)$ admits an $L_2(P)$ envelope and has a finite bracketing entropy integral, for example

$$\int_0^\delta \sqrt{\log N_{[]}\big(\varepsilon, \mathcal{F}, L_2(P)\big)}\, d\varepsilon < \infty,$$

where $\log N_{[]}\big(\varepsilon, \mathcal{F}, L_2(P)\big)$ denotes the ε-bracketing entropy for $\mathcal{F}$ in $L_2(P)$, see van der Vaart and Wellner (1996), Chapter 2.5.

Lemma S2.5 (GC and Donsker properties for Gaussian mixtures and their gradients)**.** *For fixed K and compact Θ_K, the classes $\mathcal{Q} = \{q_{\boldsymbol{\Lambda}} : \boldsymbol{\Lambda} \in \Theta_K\}$ and $\mathcal{Q}_\nabla = \{\nabla_{\boldsymbol{\Lambda}} q_{\boldsymbol{\Lambda}} : \boldsymbol{\Lambda} \in \Theta_K\}$ are P-Donsker (hence Glivenko–Cantelli) with envelopes in $L_2(P)$.*

Proof of Lemma S2.5. First, we prove that $\mathcal{Q}$ and $\mathcal{Q}_\nabla$ admit envelopes in $L_2(P)$, then we use the fact that they are Lipschitz to exhibit their Donsker property.

Step 1. L_2 envelopes. On set Θ_K, by continuity of Gaussian densities and compactness of Θ_K, there exist constants $C_0, C_1 < \infty$ (depending only on Θ_K) such that, for all x and $\boldsymbol{\Lambda} \in \Theta_K$,

$$0 \le q_{\boldsymbol{\Lambda}}(\boldsymbol{x}) \le C_0, \qquad \|\nabla_{\boldsymbol{\Lambda}} q_{\boldsymbol{\Lambda}}(\boldsymbol{x})\| \le C_1(1 + \|\boldsymbol{x}\| + \boldsymbol{x}\|^2).$$

Hence an envelope for $\mathcal{Q}$ is $Q(\boldsymbol{x}) \equiv C_0 \in L_2(P)$. Since P is supported on a bounded set $\mathcal{S} \subset \mathbb{R}^d$, there is $R < \infty$ with $\|\boldsymbol{X}\| \le R$ a.s., and thus an envelope for $\mathcal{Q}_\nabla$ is $Q_\nabla(\boldsymbol{X}) \le C_1(1 + R + R^2)$ a.s.. In particular $Q_\nabla \in L_2(P)$, therefore $\mathcal{Q}_\nabla = \{\nabla_{\boldsymbol{\Lambda}} q_{\boldsymbol{\Lambda}} : \boldsymbol{\Lambda} \in \Theta_K\}$ admits an $L_2(P)$ envelope.

Step 2. Lipschitz into $L_2(P)$. Fix $\boldsymbol{\Lambda}, \widetilde{\boldsymbol{\Lambda}} \in \Theta_K$ and set $\boldsymbol{\Lambda}_t = \widetilde{\boldsymbol{\Lambda}} + t(\boldsymbol{\Lambda} - \widetilde{\boldsymbol{\Lambda}})$, $t \in [0,1]$. By the mean value theorem in Banach spaces,

$$q_{\boldsymbol{\Lambda}} - q_{\widetilde{\boldsymbol{\Lambda}}} = \int_0^1 \langle \nabla_{\boldsymbol{\Lambda}} q_{\boldsymbol{\Lambda}_t}, \boldsymbol{\Lambda} - \widetilde{\boldsymbol{\Lambda}} \rangle dt,$$

so

$$\|q_{\boldsymbol{\Lambda}} - q_{\widetilde{\boldsymbol{\Lambda}}}\|_{L_2(P)} \leq \|\boldsymbol{\Lambda} - \widetilde{\boldsymbol{\Lambda}}\| \sup_{\boldsymbol{\Lambda}' \in \Theta_K} \|\nabla_{\boldsymbol{\Lambda}} q_{\boldsymbol{\Lambda}'}\|_{L_2(P)} \leq \|Q_\nabla\|_{L_2(P)} \|\boldsymbol{\Lambda} - \widetilde{\boldsymbol{\Lambda}}\|.$$

By the same argument applied to $\nabla_{\boldsymbol{\Lambda}} q_{\boldsymbol{\Lambda}}$ and using that all second derivatives $\nabla^2_{\boldsymbol{\Lambda}} q_{\boldsymbol{\Lambda}}(\boldsymbol{x})$ are uniformly bounded on $\mathcal{S} \times \Theta_K$, since Gaussian mixture densities have continuous second derivatives and Θ_K ensures bounded means and eigenvalues bounded away from 0 and ∞; there exists $C' < \infty$ with

$$\|\nabla_{\boldsymbol{\Lambda}} q_{\boldsymbol{\Lambda}} - \nabla_{\boldsymbol{\Lambda}} q_{\widetilde{\boldsymbol{\Lambda}}}\|_{L_2(P)} \leq C' \|\boldsymbol{\Lambda} - \widetilde{\boldsymbol{\Lambda}}\|.$$

Thus both maps $\boldsymbol{\Lambda} \mapsto q_{\boldsymbol{\Lambda}}$ and $\boldsymbol{\Lambda} \mapsto \nabla_{\boldsymbol{\Lambda}} q_{\boldsymbol{\Lambda}}$ are Lipschitz into $L_2(P)$ with L_2 envelopes.

Step 3. Donsker. By the Euclidean parametric-Lipschitz bracketing bound (van der Vaart and Wellner, 1996, Thm. 2.7.11), the $L_2(P)$ bracketing numbers of $\mathcal{Q}$ and $\mathcal{Q}_\nabla$ grow at most polynomially on compact Θ_K. Therefore, the bracketing integral is finite, and the bracketing Donsker theorem (van der Vaart and Wellner, 1996, Chp. 2.5) yields that $\mathcal{Q}$ and $\mathcal{Q}_\nabla$ are P-Donsker. □

S2.2 Proof of Theorem 2.1

We prove Theorem 2.1, establishing the consistency of the optimal variational solution $q^\star_\omega := \arg\max_{\mathcal{Q}} \mathcal{L}_\omega(q)$, where $\mathcal{Q}$ is the family of truncated Gaussian mixtures of $K \geq I$ components with upper-bounded covariances, as $\omega \to 0$. We proceed in three steps. First, we identify the asymptotics of the Gibbs measure g_ω via a Laplace expansion near each mode of f (Prop. S2.6 and S2.7). Then, we construct a truncated mixture $\widetilde{q}^{\mathcal{S}}_\omega$ on $\mathcal{S}$ such that $\mathrm{KL}(\widetilde{q}^{\mathcal{S}}_\omega || g_\omega) \to 0$ (Prop. S2.8). This implies $\mathrm{KL}(q^\star_\omega || g_\omega) \to 0$, so finally, combining with the Laplace expansions of g_ω, we deduce the claims of Theorem 2.1.

Setup

Let $f \in \mathcal{C}^3(\mathcal{S})$ be bounded with (finitely many) nondegenerate global modes $\{\boldsymbol{x}^\star_i\}_{i=1}^I$, and let $f^\star := \sup_{\boldsymbol{x} \in \mathcal{S}} f(\boldsymbol{x}) = f(\boldsymbol{x}^\star_i)$. Nondegeneracy means $\boldsymbol{H}_i = -\nabla^2 f(\boldsymbol{x}^\star_i) \succ 0$, for all $i \in [I]$. These modes are isolated and in the interior of $\mathcal{S}$, so there exist disjoint open, bounded neighborhoods $U_1, \ldots, U_I \subset \mathcal{S}$ of the $\boldsymbol{x}^\star_i$ such that each U_i contains no other critical point of f. Without loss of generality, assume that f is strictly positive on $\bigcup_{i=1}^I U_i$. Define $U_0 := \mathcal{S} \setminus \bigcup_{i=1}^I U_i$. The restricted objective over a family $\mathcal{Q}$ is

$$\mathcal{L}_\omega(q) = \mathbb{E}_q[f(\boldsymbol{X})] + \omega \mathcal{H}_{\mathcal{S}}(q), \qquad q \in \mathcal{Q}.$$

Here, the family $\mathcal{Q}$ is the $\mathcal{S}$-truncated Gaussian mixture family with $K \geq I$ components, and all eigenvalues of covariance matrices in $(0, \sigma^2_{\max}]$. Recall that for any Gaussian mixture $q_{\boldsymbol{\Lambda}}$ on $\mathbb{R}^d$, we write its $\mathcal{S}$-truncation

$$q^{\mathcal{S}}_{\boldsymbol{\Lambda}}(\boldsymbol{x}) := \frac{q_{\boldsymbol{\Lambda}}(\boldsymbol{x})\mathbb{1}_{\mathcal{S}}(\boldsymbol{x})}{\int_{\mathcal{S}} q_{\boldsymbol{\Lambda}}(\boldsymbol{y})d\boldsymbol{y}}.$$

Let $q^\star_\omega$ denote any choice of maximizer of $\mathcal{L}_\omega(q)$ in $\mathcal{Q}$, for a fixed $\omega > 0$. To simplify the notation, we denote the conditional distributions of $q^\star_\omega$ and g_ω on the regions U_i, $i = 0, \dots, I$, respectively by $\varphi_{i,\omega}$ and $\psi_{i,\omega}$, defined as follows:

$$\varphi_{i,\omega}(\boldsymbol{x}) := q^{\star,U_i}_\omega(\boldsymbol{x}) = \frac{q^\star_\omega(\boldsymbol{x})\mathbb{1}_{U_i}(\boldsymbol{x})}{\alpha_{i,\omega}}, \qquad \psi_{i,\omega}(\boldsymbol{x}) := g^{U_i}_\omega(\boldsymbol{x}) = \frac{g_\omega(\boldsymbol{x})\mathbb{1}_{U_i}(\boldsymbol{x})}{\gamma_{i,\omega}}, \tag{S6}$$

where

$$\alpha_{i,\omega} := \int_{U_i} q^\star_\omega(\boldsymbol{x})d\boldsymbol{x}, \qquad \gamma_{i,\omega} := \int_{U_i} g_\omega(\boldsymbol{x})d\boldsymbol{x}. \tag{S7}$$

Step 1: Laplace expansion and local Gibbs measure properties

We start by applying Laplace's method around each isolated mode $\boldsymbol{x}^\star_i$ of f to the unrestricted Gibbs optimizer g_ω. The two following propositions characterize the shape of the Gibbs measure under annealing.

Proposition S2.6 (Laplace expansion near a mode)**.** *Fix $i \in [I]$ and consider an open neighborhood U_i of mode $\boldsymbol{x}_i$. Then, as $\omega \to 0$,*

$$\int_{U_i} \exp\{f(\boldsymbol{x})/\omega\}d\boldsymbol{x} = e^{f^\star/\omega}(2\pi\omega)^{d/2}(\det \boldsymbol{H}_i)^{-1/2}\big(1 + o(1)\big),$$

and for any bounded continuous function h,

$$\int_{U_i} h(\boldsymbol{x})g_\omega(d\boldsymbol{x}) = h(\boldsymbol{x}^\star_i)c_i(\omega) + \frac{\omega}{2}\operatorname{Tr}\left(\boldsymbol{H}_i^{-1}\nabla^2 h(\boldsymbol{x}^\star_i)\right)c_i(\omega) + o(\omega),$$

where

$$c_i(\omega) = c_i + o(1), \qquad c_i := \frac{(\det \boldsymbol{H}_i)^{-1/2}}{\sum_{j=1}^I (\det \boldsymbol{H}_j)^{-1/2}}.$$

Moreover, $g_\omega \to \sum_{i=1}^I c_i \delta_{\boldsymbol{x}^\star_i}$ weakly, and there exists $C, \eta > 0$ such that $g_\omega(U_0) \leq Ce^{-\eta/\omega}$.

Proof of Proposition S2.6. The proof uses the classical Laplace method (Hwang, 1980; also see Le Minh et al., 2025, Proof of Thm. 2): in U_i, write the Taylor expansion of $f(\boldsymbol{x})$ at $\boldsymbol{x}^\star_i$,

$$f(\boldsymbol{x}) = f(\boldsymbol{x}^\star_i) - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^\star_i)^T \boldsymbol{H}_i(\boldsymbol{x} - \boldsymbol{x}^\star_i) + R_i(\boldsymbol{x}),$$

with $R_i(\boldsymbol{x}) \le L_i\|\boldsymbol{x}-\boldsymbol{x}_i^\star\|^3$, where $L_i>0$ depends on U_i. Then

$$\int_{U_i}\exp(f(\boldsymbol{x})/\omega)d\boldsymbol{x} = \exp(f(\boldsymbol{x}_i^\star)/\omega)\int_{U_i}\exp\Big\{-\frac{1}{2\omega}(\boldsymbol{x}-\boldsymbol{x}_i^\star)^T\boldsymbol{H}_i(\boldsymbol{x}-\boldsymbol{x}_i^\star)+\frac{1}{\omega}R_i(\boldsymbol{x})\Big\}d\boldsymbol{x},$$

and standard dominated convergence in Gaussian coordinates yields the first display. Summing over i and normalizing gives the weights $c_i(\omega)\to c_i$ and the weak limit. The second display follows from the same expansion applied to ψ (up to second order) and Gaussian moment calculations. The last statement can be established since the modes are isolated: there exists $\eta>0$ such that

$$\sup_{\boldsymbol{x}\in U_0} f(\boldsymbol{x}) \le f^\star-\eta,$$

so there exists $C>0$ such that

$$g_\omega(U_0)\le Ce^{-\eta/\omega}.$$

□

Proposition S2.7. *For $i\in[I]$ and $\psi_{i,\omega}$ defined in equation (S6), there exists $\eta>0$ such that*

$$\mathbb{E}_{\psi_{i,\omega}}[\boldsymbol{X}]=\boldsymbol{x}_i^\star+O(e^{-\eta/\omega}),\qquad \mathrm{Cov}_{\psi_{i,\omega}}(\boldsymbol{X})=\omega\boldsymbol{H}_i^{-1}+O(e^{-2\eta/\omega}).$$

Proof of Proposition S2.7. (i) Expectation. Consider the coordinate functions $h_j(\boldsymbol{x})=x_j$, for all $j\in[d]$. These functions are $\mathcal{C}^2$ with $h_j(\boldsymbol{x}_i^\star)=x_{i,j}^\star$ and $\nabla^2 h_j(\boldsymbol{x}_i^*)\equiv 0$. Apply Proposition S2.6 to $h=h_j$:

$$\int_{U_i}x_j g_\omega(\boldsymbol{x})d\boldsymbol{x}=x_{i,j}^\star c_i(\omega)+o(\omega).$$

Next, Proposition S2.6 also implies $\gamma_{i,\omega}=g_\omega(U_i)=c_i(\omega)+o(e^{-\eta/\omega})$. So

$$\mathbb{E}_{\psi_{i,\omega}}[X_j]=\frac{\int_{U_i}x_j g_\omega(\boldsymbol{x})d\boldsymbol{x}}{\gamma_{i,\omega}}=\frac{x_{i,j}^\star c_i(\omega)+o(\omega)}{c_i(\omega)+o(e^{-\eta/\omega})}.$$

Since $c_i(\omega)\to c_i>0$, the denominator is bounded away from 0 and we can divide: for all $j\in[d]$,

$$\mathbb{E}_{\psi_{i,\omega}}[X_j]=x_{i,j}^\star+O(e^{-\eta/\omega}).$$

Therefore,

$$\mathbb{E}_{\psi_{i,\omega}}[\boldsymbol{X}]=\boldsymbol{x}_i^\star+O(e^{-\eta/\omega}).$$

(ii) Covariance. Next, consider the centered second moment functions $h_{jk}(\boldsymbol{x}):=(x_j-x_{i,j}^\star)(x_k-x_{i,k}^\star)$, for all $(j,k)\in[d]^2$. These functions are $\mathcal{C}^2$ with $h_{jk}(\boldsymbol{x}_i^\star)=0$ and

$$\big(\nabla^2 h_{jk}(\boldsymbol{x}_i^\star)\big)_{\ell m}=\frac{\partial^2}{\partial x_\ell\partial x_m}\big((x_j-x_{i,j}^\star)(x_k-x_{i,k}^\star)\big)\,|_{\boldsymbol{x}=\boldsymbol{x}_i^\star}=\delta_{j\ell}\delta_{km}+\delta_{jm}\delta_{k\ell},$$

where $\delta_{..}$ is the Kronecker symbol, i.e. $\nabla^2 h_{jk}(\boldsymbol{x}_i^\star)$ is the matrix with entries (j,k) and (k,j) set to one and all other entries set to zero. Hence, by symmetry of $\boldsymbol{H}_i^{-1}$,

$$\mathrm{Tr}\left(\boldsymbol{H}_i^{-1}\nabla^2 h_{jk}(\boldsymbol{x}_i^\star)\right) = \left(\boldsymbol{H}_i^{-1}\right)_{jk} + \left(\boldsymbol{H}_i^{-1}\right)_{kj} = 2\left(\boldsymbol{H}_i^{-1}\right)_{jk}.$$

Apply Proposition S2.6 to h_{jk}:

$$\int_{U_i} h_{jk}(\boldsymbol{x}) g_\omega(\boldsymbol{x}) d\boldsymbol{x} = \omega\left(\boldsymbol{H}_i^{-1}\right)_{jk} c_i(\omega) + o(\omega).$$

Again, we can divide by $\gamma_{i,\omega} = c_i(\omega) + o(e^{-\eta/\omega})$:

$$\mathbb{E}_{\psi_{i,\omega}}[(X_j - x_{i,j}^\star)(X_k - x_{i,k}^\star)^T] = \frac{\int_{U_i} h_{jk}(\boldsymbol{x}) g_\omega(\boldsymbol{x}) d\boldsymbol{x}}{\gamma_{i,\omega}} = \omega\left(\boldsymbol{H}_i^{-1}\right)_{jk} + O(e^{-\eta/\omega}).$$

Therefore,

$$\mathbb{E}_{\psi_{i,\omega}}[(\boldsymbol{X} - \boldsymbol{x}_i^\star)(\boldsymbol{X} - \boldsymbol{x}_i^\star)^T] = \omega\boldsymbol{H}_i^{-1} + O(e^{-\eta/\omega}).$$

Finally,

$$\begin{aligned}\mathrm{Cov}_{\psi_{i,\omega}}(\boldsymbol{X}) &= \mathbb{E}_{\psi_{i,\omega}}[(\boldsymbol{X} - \mathbb{E}_{\psi_{i,\omega}}[\boldsymbol{X}])(\boldsymbol{X} - \mathbb{E}_{\psi_{i,\omega}}[\boldsymbol{X}])^T] \\ &= \mathbb{E}_{\psi_{i,\omega}}[(\boldsymbol{X} - \boldsymbol{x}_i^\star)(\boldsymbol{X} - \boldsymbol{x}_i^\star)^T] - (\mathbb{E}_{\psi_{i,\omega}}[\boldsymbol{X}] - \boldsymbol{x}_i^\star)(\mathbb{E}_{\psi_{i,\omega}}[\boldsymbol{X}] - \boldsymbol{x}_i^\star)^T,\end{aligned}$$

where we have already shown that $\mathbb{E}_{\psi_{i,\omega}}[\boldsymbol{X}] - \boldsymbol{x}_i^\star = O(e^{-\eta/\omega})$, so

$$\mathrm{Cov}_{\psi_{i,\omega}}(\boldsymbol{X}) = \omega\boldsymbol{H}_i^{-1} + O(e^{-2\eta/\omega}).$$

□

Step 2: Competitor mixture and KL convergence

For $\omega > 0$, define the Gaussian distributions, for $i \in [I]$:

$$\widetilde{q}_{i,\omega}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{x}_i^\star, \omega\boldsymbol{H}_i^{-1}).$$

These Gaussian dsitributions approximate locally the Gibbs measure at each mode. We define the associated mixture with Laplace weights

$$\widetilde{q}_\omega(\boldsymbol{x}) = \sum_{i=1}^{I} c_i \widetilde{q}_{i,\omega}(\boldsymbol{x}),$$

where $c_i \propto (\det \boldsymbol{H}_i)^{-1/2}$ and $\sum_i c_i = 1$. We use the truncated mixture $\widetilde{q}_\omega^{\mathcal{S}} := \widetilde{q}_\omega \mathbb{1}_{\mathcal{S}} / \int_{\mathcal{S}} \widetilde{q}_\omega$ as a competitor for $q_\omega^\star$ in $\mathcal{Q}$. In this step, we prove that $\widetilde{q}_\omega^{\mathcal{S}}$ converges in KL to g_ω.

Proposition S2.8. *Let* $\widetilde{q}_\omega^{\mathcal{S}} := \widetilde{q}_\omega \mathbb{1}_{\mathcal{S}} / \int_{\mathcal{S}} \widetilde{q}_\omega$ *be the truncation of* $\widetilde{q}_\omega$ *on* $\mathcal{S}$*. Then,* $\widetilde{q}_\omega^{\mathcal{S}} \in \mathcal{Q}$ *and, as* $\omega \to 0$*,*

$$\mathrm{KL}(\widetilde{q}_\omega^{\mathcal{S}} \| g_\omega) \longrightarrow 0.$$

To prove Proposition S2.8, we derive an intermediate lemma.

Lemma S2.9. *Let* $\mathrm{KL}_{\mathcal{S}}(\widetilde{q}_\omega \| g_\omega) := \int_{\mathcal{S}} \widetilde{q}_\omega \log \widetilde{q}_\omega / g_\omega$*. As* $\omega \to 0$*,*

$$\mathrm{KL}_{\mathcal{S}}(\widetilde{q}_\omega \| g_\omega) \longrightarrow 0.$$

Proof of Lemma S2.9. We first compute the KL divergence for a single Gaussian $\widetilde{q}_{i,\omega}$ and show that it converges to a finite constant. We then use the disjoint-mode structure to prove that the mixture $\widetilde{q}_\omega$ is asymptotically indistinguishable from g_ω in KL.

Step 1. Log-ratio for a single component. Fix i. On U_i we write

$$f(\boldsymbol{x}) = f(\boldsymbol{x}_i^\star) - \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_i^\star)^T \boldsymbol{H}_i (\boldsymbol{x} - \boldsymbol{x}_i^\star) + \Delta_i(\boldsymbol{x}), \qquad |\Delta_i(\boldsymbol{x})| \le C \|\boldsymbol{x} - \boldsymbol{x}_i^\star\|^3.$$

By definition,

$$g_\omega(\boldsymbol{x}) = \frac{\exp\{f(\boldsymbol{x})/\omega\}}{Z_\omega}, \qquad \widetilde{q}_{i,\omega}(\boldsymbol{x}) = (2\pi\omega)^{-d/2} (\det \boldsymbol{H}_i)^{1/2} \exp\Big(-\frac{1}{2\omega}(\boldsymbol{x} - \boldsymbol{x}_i^\star)^T \boldsymbol{H}_i (\boldsymbol{x} - \boldsymbol{x}_i^\star)\Big).$$

From Proposition S2.6,

$$Z_\omega = e^{f(\boldsymbol{x}_i^\star)/\omega} (2\pi\omega)^{d/2} \sum_{j=1}^{I} (\det \boldsymbol{H}_j)^{-1/2} \big(1 + o(1)\big),$$

so, for $\boldsymbol{x} \in U_i$,

$$\begin{aligned}
\log g_\omega(\boldsymbol{x}) &= \frac{f(\boldsymbol{x})}{\omega} - \log Z_\omega \\
&= \frac{f(\boldsymbol{x}_i^\star)}{\omega} - \frac{1}{2\omega}(\boldsymbol{x} - \boldsymbol{x}_i^\star)^T \boldsymbol{H}_i (\boldsymbol{x} - \boldsymbol{x}_i^\star) + \frac{\Delta_i(\boldsymbol{x})}{\omega} \\
&\quad - \frac{f(\boldsymbol{x}_i^\star)}{\omega} - \frac{d}{2}\log(2\pi\omega) - \log\Big(\sum_{j=1}^{I} (\det \boldsymbol{H}_j)^{-1/2}\Big) + o(1) \\
&= -\frac{1}{2\omega}(\boldsymbol{x} - \boldsymbol{x}_i^\star)^T \boldsymbol{H}_i (\boldsymbol{x} - \boldsymbol{x}_i^\star) + \frac{\Delta_i(\boldsymbol{x})}{\omega} - \frac{d}{2}\log(2\pi\omega) - \log\Big(\sum_{j=1}^{I} (\det \boldsymbol{H}_j)^{-1/2}\Big) + o(1).
\end{aligned}$$

On the other hand,

$$\log \widetilde{q}_{i,\omega}(\boldsymbol{x}) = -\frac{1}{2\omega}(\boldsymbol{x} - \boldsymbol{x}_i^\star)^T \boldsymbol{H}_i (\boldsymbol{x} - \boldsymbol{x}_i^\star) - \frac{d}{2}\log(2\pi\omega) + \frac{1}{2}\log\det \boldsymbol{H}_i.$$

Subtracting gives

$$\log \frac{\widetilde{q}_{i,\omega}(\boldsymbol{x})}{g_\omega(\boldsymbol{x})} = -\frac{\Delta_i(\boldsymbol{x})}{\omega} + \frac{1}{2}\log\det \boldsymbol{H}_i + \log\Big(\sum_{j=1}^{I}(\det \boldsymbol{H}_j)^{-1/2}\Big) + o(1)$$
$$= -\frac{\Delta_i(\boldsymbol{x})}{\omega} - \log c_i + o(1),$$

where

$$c_i = \frac{(\det \boldsymbol{H}_i)^{-1/2}}{\sum_{j=1}^{I}(\det \boldsymbol{H}_j)^{-1/2}}.$$

For $\boldsymbol{X} \sim \widetilde{q}_{i,\omega}$, write $\boldsymbol{X} = \boldsymbol{x}_i^\star + \sqrt{\omega}\boldsymbol{Y}$ with $\boldsymbol{Y} \sim \mathcal{N}(0, \boldsymbol{H}_i^{-1})$. The remainder bound implies

$$|\Delta_i(\boldsymbol{X})| \le C\|\boldsymbol{X} - \boldsymbol{x}_i^\star\|^3 = C\omega^{3/2}\|\boldsymbol{Y}\|^3,$$

so

$$\frac{\Delta_i(\boldsymbol{X})}{\omega} = O\big(\sqrt{\omega}\|\boldsymbol{Y}\|^3\big).$$

Since $\boldsymbol{Y}$ has finite moments of all orders, $\mathbb{E}[\Delta_i(\boldsymbol{X})/\omega] \to 0$ as $\omega \to 0$. Dominated convergence on U_i yields, as $\omega \to 0$,

$$\mathrm{KL}_{\mathcal{S}}(\widetilde{q}_{i,\omega}\|g_\omega) = \int_{\mathcal{S}} \widetilde{q}_{i,\omega}(\boldsymbol{x}) \log \frac{\widetilde{q}_{i,\omega}(\boldsymbol{x})}{g_\omega(\boldsymbol{x})} d\boldsymbol{x} \longrightarrow -\log c_i.$$

Step 2. KL for the mixture $\widetilde{q}_\omega$ near the modes. Define

$$\widetilde{q}_\omega(\boldsymbol{x}) = \sum_{i=1}^{I} c_i \widetilde{q}_{i,\omega}(\boldsymbol{x}).$$

We show that the limiting constant $-\log c_i$ cancels mode by mode inside the mixture, so that $\mathrm{KL}_{\mathcal{S}}(\widetilde{q}_\omega\|g_\omega) \to 0$.

Fix $R > 0$ and for each i define the shrinking ball $B_i(R,\omega) := \{\boldsymbol{x} : \|\boldsymbol{x} - \boldsymbol{x}_i^\star\| \le R\sqrt{\omega}\} \subset U_i$ for all small ω. Two modes $\boldsymbol{x}_i^\star \ne \boldsymbol{x}_j^\star$ are separated, so there exists $\delta > 0$ such that $\|\boldsymbol{x}_i^\star - \boldsymbol{x}_j^\star\| \ge \delta$. But for $\boldsymbol{x} \in B_i(R,\omega)$, by the triangle inequality,

$$\|\boldsymbol{x} - \boldsymbol{x}_j^\star\| \ge \delta - R\sqrt{\omega} > \delta/2,$$

for sufficiently small ω. Thus, since all covariances are $O(\omega)$, there exist a constant $a > 0$ such that for $j \ne i$ and $\boldsymbol{x} \in B_i(R,\omega)$,

$$\frac{\widetilde{q}_{j,\omega}(\boldsymbol{x})}{\widetilde{q}_{i,\omega}(\boldsymbol{x})} \le e^{-a/\omega}.$$

Therefore, on $B_i(R,\omega)$,

$$\widetilde{q}_\omega(\boldsymbol{x}) = c_i \widetilde{q}_{i,\omega}(\boldsymbol{x})\Big(1 + O\big(e^{-a/\omega}\big)\Big),$$

so

$$\begin{aligned}\log \frac{\widetilde{q}_\omega(\boldsymbol{x})}{g_\omega(\boldsymbol{x})} &= \log\Big(\frac{c_i \widetilde{q}_{i,\omega}(\boldsymbol{x})}{g_\omega(\boldsymbol{x})}\Big) + \log\Big(1 + O\big(e^{-a/\omega}\big)\Big) \\ &= \log c_i + \log \frac{\widetilde{q}_{i,\omega}(\boldsymbol{x})}{g_\omega(\boldsymbol{x})} + o(1)\end{aligned}$$

with an $o(1)$ term uniform in $\boldsymbol{x} \in B_i(R,\omega)$ as $\omega \to 0$.

Substituting the expansion from Step 1,

$$\log \frac{\widetilde{q}_{i,\omega}(\boldsymbol{x})}{g_\omega(\boldsymbol{x})} = -\frac{\Delta_i(\boldsymbol{x})}{\omega} - \log c_i + o(1),$$

we obtain, on $B_i(R,\omega)$,

$$\log \frac{\widetilde{q}_\omega(\boldsymbol{x})}{g_\omega(\boldsymbol{x})} = -\frac{\Delta_i(\boldsymbol{x})}{\omega} + o(1). \tag{S8}$$

Let $\boldsymbol{X} \sim \widetilde{q}_{i,\omega}$. As above, $\boldsymbol{X} = \boldsymbol{x}_i^\star + \sqrt{\omega}\boldsymbol{Y}$ with $\boldsymbol{Y} \sim \mathcal{N}(0, \boldsymbol{H}_i^{-1})$, so

$$\left|\frac{\Delta_i(\boldsymbol{X})}{\omega}\right| \le C\sqrt{\omega}\|\boldsymbol{Y}\|^3, \qquad \mathbb{E}\left[\left|\frac{\Delta_i(\boldsymbol{X})}{\omega}\right|\right] \to 0.$$

Using (S8) and dominated convergence on $B_i(R,\omega)$ yields

$$\int_{B_i(R,\omega)} \widetilde{q}_{i,\omega}(\boldsymbol{x}) \log \frac{\widetilde{q}_\omega(\boldsymbol{x})}{g_\omega(\boldsymbol{x})} d\boldsymbol{x} \longrightarrow 0 \qquad \text{as } \omega \to 0,$$

for each fixed R.

Step 3. KL for the mixture $\widetilde{q}_\omega$ in the tails. By Gaussian tail bounds,

$$\widetilde{q}_{i,\omega}\big(B_i(R,\omega)^c\big) \le C_1 e^{-d_1 R^2}$$

for some constants $C_1, d_1 > 0$, uniformly in ω. On $B_i(R,\omega)^c \cap U_i$, both $\widetilde{q}_\omega$ and g_ω are exponentially small in R^2 and $|\log(\widetilde{q}_\omega/g_\omega)|$ grows at most polynomially in $\|\boldsymbol{x} - \boldsymbol{x}_i^\star\|$, so

$$\int_{B_i(R,\omega)^c \cap U_i} \widetilde{q}_{i,\omega}(\boldsymbol{x}) \Big|\log \frac{\widetilde{q}_\omega(\boldsymbol{x})}{g_\omega(\boldsymbol{x})}\Big| d\boldsymbol{x} \le C_2 e^{-d_2 R^2}$$

for suitable constants $C_2, d_2 > 0$.

On U_0, Proposition S2.6 gives $g_\omega(U_0) \le Ce^{-\eta/\omega}$, and the same Gaussian tail arguments yield $\widetilde{q}_{i,\omega}(U_0) \le C_1' e^{-a_i/\omega}$, so the contribution of U_0 to the KL integral is bounded by $C_2' e^{-a_i'/\omega}$.

Step 4. Conclusion. Putting everything together and summing over i with weights c_i,

$$\begin{aligned}\mathrm{KL}_{\mathcal{S}}(\widetilde{q}_\omega \| g_\omega) &= \int \widetilde{q}_\omega(\boldsymbol{x}) \log \frac{\widetilde{q}_\omega(\boldsymbol{x})}{g_\omega(\boldsymbol{x})} d\boldsymbol{x} \\ &= \sum_{i=1}^{I} c_i \int \widetilde{q}_{i,\omega}(\boldsymbol{x}) \log \frac{\widetilde{q}_\omega(\boldsymbol{x})}{g_\omega(\boldsymbol{x})} d\boldsymbol{x} \\ &= \sum_{i=1}^{I} c_i \left[\int_{B_i(R,\omega)} \widetilde{q}_{i,\omega}(\boldsymbol{x}) \log \frac{\widetilde{q}_\omega(\boldsymbol{x})}{g_\omega(\boldsymbol{x})} d\boldsymbol{x} + \int_{B_i(R,\omega)^c \cup U_0} \widetilde{q}_{i,\omega}(\boldsymbol{x}) \log \frac{\widetilde{q}_\omega(\boldsymbol{x})}{g_\omega(\boldsymbol{x})} d\boldsymbol{x} \right].\end{aligned}$$

For each fixed R, the first term in brackets tends to 0 as $\omega \to 0$, by the local analysis above. The second term is bounded in absolute value by $Ce^{-cR^2} + C'e^{-a/\omega}$, independently of ω.

Given $\varepsilon > 0$, pick R large enough that the tail bound is $< \varepsilon/2$, then let $\omega \to 0$ to make the local terms $< \varepsilon/2$. Thus, $\mathrm{KL}_{\mathcal{S}}(\widetilde{q}_\omega \| g_\omega) \to 0$. □

Proof of Proposition S2.8. Note that bounding the Gaussian tails in $\widetilde{q}_\omega$ gives $\beta_\omega := \int_{\mathbb{R}^d \setminus \mathcal{S}} \widetilde{q}_\omega(\boldsymbol{x}) d\boldsymbol{x} = O(e^{-c/\omega})$, for some $c > 0$. From Lemma S2.9, $\mathrm{KL}_{\mathcal{S}}(\widetilde{q}_\omega || g_\omega) = o(1)$. Since $\widetilde{q}_\omega^{\mathcal{S}} = \widetilde{q}_\omega/(1-\beta)$ on $\mathcal{S}$, we have

$$\begin{aligned}\mathrm{KL}(\widetilde{q}_\omega^{\mathcal{S}} || g_\omega) - \mathrm{KL}_{\mathcal{S}}(\widetilde{q}_\omega || g_\omega) &= \left(\frac{1}{1-\beta_\omega} - 1 \right) \mathrm{KL}_{\mathcal{S}}(\widetilde{q}_\omega || g_\omega) - \log(1-\beta_\omega) \\ &= O(\beta_\omega) \mathrm{KL}_{\mathcal{S}}(\widetilde{q}_\omega || g_\omega) + O(\beta_\omega) \\ &= O(e^{-c/\omega}).\end{aligned}$$

Therefore, we can conclude that $\left| \mathrm{KL}(\widetilde{q}_\omega^{\mathcal{S}} || g_\omega) - \mathrm{KL}_{\mathcal{S}}(\widetilde{q}_\omega || g_\omega) \right| \to 0$, which proves the proposition. □

Step 3: Conclusion, proof of Theorem 2.1

We are now going to prove all the statements of Theorem 2.1. Recall the conditionals of g_ω and $q_\omega^\star$ on U_i, $i = 0, \ldots, I$:

$$\varphi_{i,\omega}(\boldsymbol{x}) = \frac{q_\omega^\star(\boldsymbol{x}) \mathbb{1}_{U_i}(\boldsymbol{x})}{\alpha_{i,\omega}}, \qquad \psi_{i,\omega}(\boldsymbol{x}) = \frac{g_\omega(\boldsymbol{x}) \mathbb{1}_{U_i}(\boldsymbol{x})}{\gamma_{i,\omega}},$$

where

$$\alpha_{i,\omega} = \int_{U_i} q_\omega^\star(\boldsymbol{x}) d\boldsymbol{x}, \qquad \gamma_{i,\omega} = \int_{U_i} g_\omega(\boldsymbol{x}) d\boldsymbol{x},$$

and the competitor mixture $\widetilde{q}_\omega^{\mathcal{S}} = \widetilde{q}_\omega \mathbb{1}_{\mathcal{S}} / \int_{\mathcal{S}} \widetilde{q}_\omega(\boldsymbol{x}) d\boldsymbol{x}$, with

$$\widetilde{q}_\omega(\boldsymbol{x}) = \sum_{i=1}^{I} c_i \mathcal{N}(\boldsymbol{x}; \boldsymbol{x}_i^\star, \omega \boldsymbol{H}_i^{-1}).$$

Proof of Theorem 2.1. The principle of the proof is as follows: because the competitor $\widetilde{q}_\omega^{\mathcal{S}}$ of Proposition S2.8 converges in KL to g_ω, the same holds for $q_\omega^\star$, and therefore $q_\omega^\star$ inherits the asymptotics of the Gibbs measure g_ω. Steps 1-2 prove Theorem 2.1(i), and Step 3 proceeds for Theorem 2.1(ii).

Step 1. Convergence of $q_\omega^\star$ in KL. For any density q on $\mathcal{S}$, we have the variational identity,

$$\mathcal{L}_\omega(q) = \omega \log Z_\omega - \omega \mathrm{KL}(q \| g_\omega).$$

The competitor mixture $\widetilde{q}_\omega^{\mathcal{S}} \in \mathcal{Q}$. Since $q_\omega^\star$ maximizes $\mathcal{L}_\omega$ over $\mathcal{Q}$, we have, as $\omega \to 0$,

$$\mathrm{KL}(q_\omega^\star || g_\omega) \leq \mathrm{KL}(\widetilde{q}_\omega^{\mathcal{S}} || g_\omega) \to 0,$$

where the limit follows from Proposition S2.8. Thus, the optimal truncated mixture satisfies

$$\mathrm{KL}(q_\omega^\star \| g_\omega) \to 0. \tag{S9}$$

Step 2. Asymptotic mass in mode neighborhoods. For $i = 0, \ldots, I$, consider $\alpha_{i,\omega}$ and $\gamma_{i,\omega}$ of (S7). Let $T : \mathcal{S} \to \{0, 1, \ldots, I\}$ send $x \in U_j$ to j. Let $\boldsymbol{\alpha}_\omega$ and $\boldsymbol{\gamma}_\omega$ be the induced discrete laws. Theorem 4.1 of Kullback and Leibler, 1951 gives

$$\mathrm{KL}(\boldsymbol{\alpha}_\omega \| \boldsymbol{\gamma}_\omega) \leq \mathrm{KL}(q_\omega^\star \| g_\omega) \to 0.$$

From Proposition S2.6, $\gamma_{i,\omega} \to c_i$ for $i \in [I]$, and $\gamma_{0,\omega} \to 0$. Hence (S9) implies

$$\alpha_{i,\omega} \to c_i, \qquad i \in [I], \qquad \alpha_{0,\omega} \to 0.$$

Step 3. Local conditional densities. For $i \in [I]$, consider the conditional densities $\varphi_{i,\omega}$ and $\psi_{i,\omega}$ from (S6). The chain rule for the KL divergence over the partition $\{U_0, U_1, \ldots, U_I\}$ gives

$$\mathrm{KL}(q_\omega^\star \| g_\omega) = \mathrm{KL}(\boldsymbol{\alpha}_\omega \| \boldsymbol{\gamma}_\omega) + \sum_{i=0}^{I} \alpha_{i,\omega} \mathrm{KL}(\varphi_{i,\omega} \| \psi_{i,\omega}).$$

Using (S9) and $\alpha_{i,\omega} \to c_i > 0$, we obtain for each $i \in [I]$,

$$\mathrm{KL}(\varphi_{i,\omega} \| \psi_{i,\omega}) \to 0.$$

Furthermore, since all densities live on the bounded set U_i, Pinsker's inequality implies convergence in total variation. From Proposition S2.7, the conditional law $\psi_{i,\omega}$ has mean $\boldsymbol{x}_i^\star + o(1)$ and covariance $o(1)$. Hence, the KL convergence forces the same limits for the mixture's local conditional moments:

$$\mathbb{E}_{\varphi_{i,\omega}}[\boldsymbol{X}] = \boldsymbol{x}_i^\star + o(1), \qquad \mathrm{Cov}_{\varphi_{i,\omega}}(\boldsymbol{X}) = o(1).$$

□

S2.3 Proof of Theorem 3.1

In this section, we prove Theorem 3.1, establishing fixed-temperature asymptotics for the empirical maximizers when the sample size grows to infinity.

For convenience, we recall the assumptions of Theorem 3.1.

Assumption S2.10. *Define the following set of assumptions:*

(B1) *(**Data**)* $X_1, \dots, X_N \overset{i.i.d.}{\sim} p$ *supported on a bounded* $\mathcal{S} \subset \mathbb{R}^d$*, with* p *bounded and continuous on* $\mathcal{S}$*.*

(B2) *(**Unicity of the population maximizer**)* $\mathcal{L}_{\omega_0}$ *admits a unique maximizer* $\boldsymbol{\Lambda}^\star_{\omega_0} \in \operatorname{int}(\Theta_K)$*.*

(B3) *(**Nondegeneracy of the population maximizer**) At the maximizer* $\boldsymbol{\Lambda}^\star_{\omega_0}$*,* $\boldsymbol{H}^\star_{\omega_0} := \boldsymbol{H}_{\omega_0}(\boldsymbol{\Lambda}^\star_{\omega_0})$ *is positive definite.*

Each of the following subsections proves one of the three statements in Theorem 3.1:

(i) Under (B1), the population objective $\mathcal{L}_{\omega_0}$ attains its maximum on Θ_K.

(ii) Under (B1)-(B2), any empirical maximizer $\widehat{\boldsymbol{\Lambda}}_{\omega_0,N} \in \arg\max \widehat{\mathcal{L}}_{\omega_0,N}$ satisfies

$$\widehat{\boldsymbol{\Lambda}}_{\omega_0,N} \xrightarrow{P} \boldsymbol{\Lambda}^\star_{\omega_0}.$$

(iii) Under (B1)-(B2)-(B3), any empirical maximizer $\widehat{\boldsymbol{\Lambda}}_{\omega_0,N} \in \arg\max \widehat{\mathcal{L}}_{\omega_0,N}$ satisfies

$$\sqrt{N}\big(\widehat{\boldsymbol{\Lambda}}_{\omega_0,N} - \boldsymbol{\Lambda}^\star_{\omega_0}\big) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \boldsymbol{W}_{\omega_0}),$$

where $\boldsymbol{W}_{\omega_0} := \big(\boldsymbol{H}^\star_{\omega_0}\big)^{-1} \boldsymbol{V}_{\omega_0} \big(\boldsymbol{H}^\star_{\omega_0}\big)^{-1}$, and $\boldsymbol{V}_{\omega_0} := \operatorname{Var}\big(\nabla_{\boldsymbol{\Lambda}} q_{\boldsymbol{\Lambda}}(\boldsymbol{X})\big)\big|_{\boldsymbol{\Lambda}=\boldsymbol{\Lambda}^\star_{\omega_0}}$.

S2.3.1 Existence of a population maximizer

Proof of (i). On the compact Θ_K, $(\boldsymbol{\Lambda}, \boldsymbol{x}) \mapsto q_{\boldsymbol{\Lambda}}(\boldsymbol{x})$ is continuous and is uniformly bounded on the bounded set $\mathcal{S}$. For each $x \in \mathcal{S}$, $\boldsymbol{\Lambda} \mapsto q_{\boldsymbol{\Lambda}}(\boldsymbol{x})$ is continuous. By dominated convergence (bounded envelope on $\mathcal{S}$), $\boldsymbol{\Lambda} \mapsto \mathbb{E}[q_{\boldsymbol{\Lambda}}(\boldsymbol{X})]$ is continuous. The entropy term $\boldsymbol{\Lambda} \mapsto \mathcal{H}_{\mathcal{S}}(q_{\boldsymbol{\Lambda}})$ is continuous on Θ_K as well. Thus, by Weierstrass, $\mathcal{L}_{\omega_*}$ is continuous on compact Θ_K and attains its maximum. □

S2.3.2 Consistency of the empirical maximizers

The proof of this result consists of two steps: the derivation of a uniform law of large numbers, due to the Glivenko–Cantelli (GC) property of the Gaussian mixture family in Θ_K (see App. S2.1, Cor. S2.4), and the application of the Argmax Theorem. For more details on the GC property, we refer to Section S2.1.

Proof of (ii). The class $\{q_{\boldsymbol{\Lambda}} : \boldsymbol{\Lambda} \in \Theta_K\}$ is Glivenko–Cantelli with an integrable envelope on $\mathcal{S}$, so as $N \to \infty$:

$$\sup_{\boldsymbol{\Lambda}\in\Theta_K} \left| \frac{1}{N} \sum_{i=1}^{N} q_{\boldsymbol{\Lambda}}(\boldsymbol{X}_i) - \mathbb{E}_p[q_{\boldsymbol{\Lambda}}(\boldsymbol{X})] \right| \xrightarrow{P} 0.$$

Adding the deterministic term $\omega_* \mathcal{H}_{\mathcal{S}}(q_{\boldsymbol{\Lambda}})$ yields

$$\sup_{\boldsymbol{\Lambda}\in\Theta_K} \left| \widehat{\mathcal{L}}_{\omega_*,N}(\boldsymbol{\Lambda}) - \mathcal{L}_{\omega_*}(\boldsymbol{\Lambda}) \right| \xrightarrow{P} 0.$$

Since $\mathcal{L}_{\omega_*}$ has a unique maximizer $\boldsymbol{\Lambda}^\star_{\omega_*}$ (Assumption (B2)), the Argmax Theorem (van der Vaart, 1998, Thm. 5.7) yields $\widehat{\boldsymbol{\Lambda}}_{\omega_*,N} \xrightarrow{P} \boldsymbol{\Lambda}^\star_{\omega_*}$, as $N \to \infty$. □

S2.3.3 Asymptotic normality of the empirical maximizers

The asymptotic normality result is derived from a Taylor expansion and the Donsker property on the Gaussian mixture family but also on the gradients. Lemma S2.5 gives these properties.

Proof of (iii). By Lemma S2.5, $\mathcal{Q}$ is GC, so $\sup_{\boldsymbol{\Lambda}\in\Theta_K} \left| \widehat{\mathcal{L}}_{\omega_*,N}(\boldsymbol{\Lambda}) - \mathcal{L}_{\omega_*}(\boldsymbol{\Lambda}) \right| \xrightarrow{P} 0$, hence $\widehat{\boldsymbol{\Lambda}}_{\omega_*,N} \xrightarrow{P} \boldsymbol{\Lambda}^\star_{\omega_*}$ by the standard M-estimator consistency theorem (van der Vaart, 1998, Thm. 5.7). Write

$$\nabla_{\boldsymbol{\Lambda}} \widehat{\mathcal{L}}_{\omega_*,N}(\boldsymbol{\Lambda}) = \frac{1}{N} \sum_{i=1}^{N} \nabla_{\boldsymbol{\Lambda}} q_{\boldsymbol{\Lambda}}(X_i) + \omega_* \nabla_{\boldsymbol{\Lambda}} \mathcal{H}_{\mathcal{S}}(q_{\boldsymbol{\Lambda}}),$$
$$\nabla_{\boldsymbol{\Lambda}} \mathcal{L}_{\omega_*}(\boldsymbol{\Lambda}) = \mathbb{E}[\nabla_{\boldsymbol{\Lambda}} q_{\boldsymbol{\Lambda}}(X)] + \omega_* \nabla_{\boldsymbol{\Lambda}} \mathcal{H}_{\mathcal{S}}(q_{\boldsymbol{\Lambda}}).$$

Thus $\nabla_{\boldsymbol{\Lambda}} \widehat{\mathcal{L}}_{\omega_*,N}(\boldsymbol{\Lambda}^\star_{\omega_*}) - \nabla_{\boldsymbol{\Lambda}} \mathcal{L}_{\omega_*}(\boldsymbol{\Lambda}^\star_{\omega_*}) = (P_N - P) \nabla_{\boldsymbol{\Lambda}} q_{\boldsymbol{\Lambda}}(X)\big|_{\boldsymbol{\Lambda}=\boldsymbol{\Lambda}^\star_{\omega_*}}$ has mean zero and variance $\boldsymbol{V}_{\omega_*}/N$.

Furthermore, a Taylor expansion around $\boldsymbol{\Lambda}^\star_{\omega_*}$ gives

$$0 = \nabla_{\boldsymbol{\Lambda}} \widehat{\mathcal{L}}_{\omega_*,N}(\widehat{\boldsymbol{\Lambda}}_{\omega_*,N}) = \nabla_{\boldsymbol{\Lambda}} \widehat{\mathcal{L}}_{\omega_*,N}(\boldsymbol{\Lambda}^\star_{\omega_*}) - \boldsymbol{H}^\star_{\omega_*}(\widehat{\boldsymbol{\Lambda}}_{\omega_*,N} - \boldsymbol{\Lambda}^\star_{\omega_*}) + r_N,$$

where, by the integral form of Taylor's theorem,

$$r_N = \frac{1}{2}(\widehat{\boldsymbol{\Lambda}}_{\omega_*,N} - \boldsymbol{\Lambda}^\star_{\omega_*})^T \Big[\nabla^3_{\boldsymbol{\Lambda}} \widehat{\mathcal{L}}_{\omega_*,N}(\boldsymbol{\Lambda}^\star_{\omega_*} + \tau(\widehat{\boldsymbol{\Lambda}}_{\omega_*,N} - \boldsymbol{\Lambda}^\star_{\omega_*})) \Big] (\widehat{\boldsymbol{\Lambda}}_{\omega_*,N} - \boldsymbol{\Lambda}^\star_{\omega_*})$$

for some $\tau \in (0,1)$. Since Θ_K is compact and $\mathcal{S}$ (the support of P) is bounded, the parametric map $\boldsymbol{\Lambda} \mapsto q_{\boldsymbol{\Lambda}}(\boldsymbol{x})$ is C^∞ and all third derivatives are uniformly bounded on $\mathcal{S} \times \Theta_K$. Hence, $\nabla^3_{\boldsymbol{\Lambda}} \widehat{\mathcal{L}}_{\omega_*,N}$ is uniformly bounded, giving $|r_N| \leq C \|\widehat{\boldsymbol{\Lambda}}_{\omega_*,N} - \boldsymbol{\Lambda}^\star_{\omega_*}\|^2 = o_P(\|\widehat{\boldsymbol{\Lambda}}_{\omega_*,N} - \boldsymbol{\Lambda}^\star_{\omega_*}\|)$.

Therefore, by the CLT for the P-Donsker gradient class (Lemma S2.5), $\sqrt{N}\nabla_{\boldsymbol{\Lambda}}\widehat{\mathcal{L}}_{\omega_*,N}(\boldsymbol{\Lambda}^\star_{\omega_*}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \boldsymbol{V}_{\omega_*})$. Since $\boldsymbol{H}^\star_{\omega_*}$ is positive definite, rearranging yields the usual Z-estimator linearization (van der Vaart, 1998, Thm. 5.41):

$$\sqrt{N}(\widehat{\boldsymbol{\Lambda}}_{\omega_*,N} - \boldsymbol{\Lambda}^\star_{\omega_*}) = (\boldsymbol{H}^\star_{\omega_*})^{-1}\sqrt{N}\nabla_{\boldsymbol{\Lambda}}\widehat{\mathcal{L}}_{\omega_*,N}(\boldsymbol{\Lambda}^\star_{\omega_*}) + o_P(1),$$

and therefore $\sqrt{N}(\widehat{\boldsymbol{\Lambda}}_{\omega_*,N} - \boldsymbol{\Lambda}^\star_{\omega_*}) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, (\boldsymbol{H}^\star_{\omega_*})^{-1}\boldsymbol{V}_{\omega_*}(\boldsymbol{H}^\star_{\omega_*})^{-1}\big)$, as $N \to \infty$. □

S2.4 Proofs for Sections 5 and S1.5

S2.4.1 Preliminaries: Bounded-Lipschitz metric

Let $(\mathcal{S}, d)$ be a metric space. For $f : \mathcal{S} \to \mathbb{R}$ define

$$\|f\|_\infty = \sup_{\boldsymbol{x}\in\mathcal{S}} |f(\boldsymbol{x})|, \qquad \mathrm{Lip}(f) = \sup_{\boldsymbol{x}\neq\boldsymbol{y}} \frac{|f(\boldsymbol{x}) - f(\boldsymbol{y})|}{d(\boldsymbol{x}, \boldsymbol{y})}.$$

For two measures ν, ν' on $\mathcal{S}$, the bounded-Lipschitz (BL) metric is

$$d_{\mathrm{BL}}(\nu, \nu') := \sup_{\|f\|_{\mathrm{BL}}\leq 1} \Big| \int f d\nu - \int f d\nu' \Big|, \qquad \|f\|_{\mathrm{BL}} := \|f\|_\infty + \mathrm{Lip}(f).$$

Convergence in d_{BL} implies weak convergence (Dudley, 2002), denoted by $\xrightarrow{w}$. We use it to state bootstrap results compactly.

S2.4.2 Technical lemmas

We state these lemmas, with proofs in Section S2.4.3.

Lemma S2.11 (Regularity for Gaussian mixtures)**.** *For a bounded space $\mathcal{S}$ and on the compact Θ_K, there exists a neighborhood $\mathcal{V} \subset \Theta_K$ of $\boldsymbol{\Lambda}^\star_{\omega_0}$ such that*

(i) $\boldsymbol{\Lambda} \mapsto q_{\boldsymbol{\Lambda}}(\boldsymbol{x})$ is $\mathcal{C}^2$ on $\mathcal{V}$ for every $\boldsymbol{x} \in \mathcal{S}$;

(ii) $\sup_{\boldsymbol{\Lambda}\in\mathcal{V}} \|\nabla_{\boldsymbol{\Lambda}} q_{\boldsymbol{\Lambda}}(\boldsymbol{x})\| \leq F(\boldsymbol{x})$ for a bounded-Lipschitz function F on $\mathcal{S}$.

Lemma S2.12 (Empirical-process CLT and bootstrap)**.** *Let $\mathcal{F} := \big\{\boldsymbol{x} \mapsto \nabla_{\boldsymbol{\Lambda}} q_{\boldsymbol{\Lambda}}(\boldsymbol{x}) \in \mathbb{R}^p : \boldsymbol{\Lambda} \in \Theta_K\big\}$. We have, as $N \to \infty$:*

$$\sqrt{N}(P_N - P) \xrightarrow{w} G_P \quad \text{in } \ell^\infty(\mathcal{F}), \qquad \sqrt{N}(P^*_N - P_N) \xrightarrow{w} G_P \quad \text{in } \ell^\infty(\mathcal{F}),$$

conditionally on $\boldsymbol{X}_{1:N}$, in P-probability, and G_P denote the P-Brownian bridge indexed by $\mathcal{F}$.

Lemma S2.13 (Hadamard differentiability of the argmax map)**.** *Let $\mathcal{V} \subset \Theta_K$ be the set in which Lemma S2.12 holds, and*

$$\mathcal{D}_0(\mathcal{S}) := \{h : h \text{ is a finite signed measure on } \mathcal{S}, h(\mathcal{S}) = 0\},$$

equipped with $\|h\|_{\mathcal{F}} := \sup_{\boldsymbol{\Lambda}\in\mathcal{V}} \int \nabla_{\boldsymbol{\Lambda}} q_{\boldsymbol{\Lambda}} dh$. Define $S : \mathcal{D}_0(\mathcal{S}) \times \mathcal{V} \to \mathbb{R}^{\dim(\Theta_K)}$ by

$$S(G, \boldsymbol{\Lambda}) = \int \nabla_{\boldsymbol{\Lambda}} q_{\boldsymbol{\Lambda}}(\boldsymbol{x}) dG(\boldsymbol{x}) + \omega_0 \nabla_{\boldsymbol{\Lambda}} \mathcal{H}_{\mathcal{S}}(q_{\boldsymbol{\Lambda}}).$$

Under Assumption S2.10, we have the following

(i) There exists a neighborhood $\mathcal{U}$ of P such that for every $G \in \mathcal{U}$, the equation over $\boldsymbol{\Lambda}$

$$S(G, \boldsymbol{\Lambda}) = 0$$

admits a unique solution $\boldsymbol{\Lambda}(G) \in \mathcal{V}$.

(ii) The map $\Phi : G \mapsto \boldsymbol{\Lambda}(G)$ is Hadamard differentiable at $G = P$, tangentially to $\mathcal{D}_0(\mathcal{S})$, with

$$D\Phi_P(h) = (\boldsymbol{H}^*_{\omega_0})^{-1} \int \nabla_{\boldsymbol{\Lambda}} q_{\boldsymbol{\Lambda}^*_{\omega_0}}(\boldsymbol{x}) dh(\boldsymbol{x}).$$

Lemma S2.14 (Measurability of the matching functional)**.** *$\boldsymbol{\Lambda} \mapsto \mathcal{M}(\boldsymbol{\Lambda})$ is Borel-measurable on Θ_K. In fact, $\mathcal{M}$ is continuous on a finite Borel partition of Θ_K.*

Lemma S2.15 (Continuity of the stability indicator F_j)**.** *Fix $j \in [K_0]$ and let $U_j = B(\boldsymbol{u}_j, r_j)$ be an open ball. Assume (by shrinking r_j if needed) that $\boldsymbol{E}_j \mathcal{M}(\boldsymbol{\Lambda}^\star_{\omega_0})$ lies at positive distance from ∂U_j. Then, under Assumption S2.10, for any deterministic measurable selector Φ from the argmax, $F_j(G) = \mathbb{1}\{\boldsymbol{E}_j \mathcal{M}(\Phi(G)) \in U_j\}$ is continuous at $G = P$ with respect to weak convergence: if $G_n \xrightarrow{w} P$, then $F_j(G_n) \to F_j(P)$.*

S2.4.3 Proofs of lemmas

Proof of Lemma S2.12. By Lemma S2.5, the class $\mathcal{F}$ is P-Donsker and admits a measurable envelope in $L_2(P)$. Hence, it is also Glivenko–Cantelli and

$$\sqrt{N}(P_N - P) \xrightarrow{w} G_P \quad \text{in } \ell^\infty(\mathcal{F}),$$

where G_P is a P-Brownian bridge indexed by $\mathcal{F}$. Furthermore, the nonparametric bootstrap empirical process satisfies the conditional functional CLT (van der Vaart and Wellner, 1996, Thm. 3.6.3):

$$\sqrt{N}(P^*_N - P_N) \xrightarrow{w} G_P \quad \text{in } \ell^\infty(\mathcal{F}),$$

conditionally on $\boldsymbol{X}_{1:N}$, in P-probability. □

Proof of Lemma S2.13. (i) By Lemma S2.11, for all $\boldsymbol{x} \in \mathcal{S}$, the map $\boldsymbol{\Lambda} \mapsto \nabla_{\boldsymbol{\Lambda}} q_{\boldsymbol{\Lambda}}(\boldsymbol{x})$ is $\mathcal{C}^1$ on $\mathcal{V}$, and $\sup_{\boldsymbol{\Lambda}\in\mathcal{V}} \|\nabla_{\boldsymbol{\Lambda}} q_{\boldsymbol{\Lambda}}(\boldsymbol{x})\| \le F(\boldsymbol{x})$ for some bounded function F. Therefore, for any measure $G \in \mathcal{D}_0(\mathcal{S})$ (with $\|G\|_{\mathcal{F}} < \infty$), the map $\boldsymbol{\Lambda} \mapsto \int \nabla_{\boldsymbol{\Lambda}} q_{\boldsymbol{\Lambda}} dG$ is $\mathcal{C}^1$ on $\mathcal{V}$.

This implies that $S(G, \boldsymbol{\Lambda})$ is $\mathcal{C}^1$ in $\boldsymbol{\Lambda}$, and

$$S(P, \boldsymbol{\Lambda}^\star_{\omega_0}) = 0, \qquad \partial_{\boldsymbol{\Lambda}} S(P, \boldsymbol{\Lambda}^\star_{\omega_0}) = \nabla^2_{\boldsymbol{\Lambda}} \mathcal{L}_{\omega_0}(\boldsymbol{\Lambda}^\star_{\omega_0}) = -\boldsymbol{H}^*_{\omega_0} \prec 0,$$

so by Assumption S2.10, $\partial_{\boldsymbol{\Lambda}} S(P, \boldsymbol{\Lambda}^\star_{\omega_0})$ is invertible. □

Proof of Lemma S2.14. Let Π be the finite set of injections $\pi : \{1, \dots, K_0\} \hookrightarrow \{1, \dots, K\}$ and define

$$\mathsf{cost}_\pi(\boldsymbol{\Lambda}) := \sum_{j=1}^{K_0} \|\boldsymbol{\mu}_{\pi(j)}(\boldsymbol{\Lambda}) - \boldsymbol{u}_j\|.$$

Each cost_π is continuous in $\boldsymbol{\Lambda}$ (composition of continuous maps). For $\pi \in \Pi$, set the open "uniqueness" region

$$\mathcal{U}_\pi := \Big\{ \boldsymbol{\Lambda} \in \Theta_K : \mathsf{cost}_\pi(\boldsymbol{\Lambda}) < \mathsf{cost}_\rho(\boldsymbol{\Lambda}) \forall \rho \neq \pi \Big\}.$$

On $\mathcal{U}_\pi$ the minimizer is uniquely π, hence

$$\mathcal{M}(\boldsymbol{\Lambda}) = \big(\boldsymbol{\mu}_{\pi(1)}(\boldsymbol{\Lambda})^T, \dots, \boldsymbol{\mu}_{\pi(K_0)}(\boldsymbol{\Lambda})^T\big)^T,$$

which is continuous there because each $\boldsymbol{\mu}_k$ is continuous.

Equip Π with the lexicographic order $\prec_{\text{lex}}$: for $\pi, \rho \in \Pi$, write $\pi \prec_{\text{lex}} \rho$ if there exists $j^\star$ with $\pi(j) = \rho(j)$ for all $j < j^\star$ and $\pi(j^\star) < \rho(j^\star)$. For each $\pi \in \Pi$, define the tie region

$$\mathcal{T}_\pi := \Big\{ \boldsymbol{\Lambda} \in \Theta_K : \mathsf{cost}_\pi(\boldsymbol{\Lambda}) \le \mathsf{cost}_\eta(\boldsymbol{\Lambda}) \forall \eta \in \Pi \text{ and } \mathsf{cost}_\pi(\boldsymbol{\Lambda}) < \mathsf{cost}_\rho(\boldsymbol{\Lambda}) \forall \rho \in \Pi \text{ with } \rho \prec_{\mathsf{lex}} \pi \Big\}.$$

Here, each set $\{\mathsf{cost}_\pi \le \mathsf{cost}_\eta\}$ is closed and each $\{\mathsf{cost}_\pi < \mathsf{cost}_\rho\}$ is open, so $\mathcal{T}_\pi$ is Borel. On $\mathcal{T}_\pi$, the tie-break selects π, hence the same formula for $\mathcal{M}$ holds and $\mathcal{M}$ is continuous there.

Finally,

$$\Theta_K = \Big(\bigsqcup_{\pi\in\Pi} \mathcal{U}_\pi \Big) \sqcup \Big(\bigsqcup_{\pi\in\Pi} \mathcal{T}_\pi \Big).$$

is a finite Borel partition on whose pieces $\mathcal{M}$ is continuous. Therefore $\mathcal{M}$ is Borel measurable on Θ_K. □

Proof of Lemma S2.15. By Lemma S2.13, Φ is continuous at P, so $G_n \xrightarrow{w} P$ implies $\Phi(G_n) \to \Phi(P) = \boldsymbol{\Lambda}^\star_{\omega_0}$. By Proposition S2.17, $\mathcal{M}$ is $\mathcal{C}^1$ (hence continuous) at $\boldsymbol{\Lambda}^\star_{\omega_0}$. Furthermore, $\boldsymbol{E}_j$ is linear and the ball-membership map is continuous away from the boundary, e.g. via

$$\psi_j(\boldsymbol{y}) := r_j - \|\boldsymbol{y} - \boldsymbol{u}_j\| \quad \text{(continuous)}, \qquad \mathbb{1}\{\boldsymbol{y} \in U_j\} = \mathbb{1}\{\psi_j(\boldsymbol{y}) > 0\}.$$

Therefore

$$\psi_j\big(\boldsymbol{E}_j\mathcal{M}(\Phi(G_n))\big) \longrightarrow \psi_j\big(\boldsymbol{E}_j\mathcal{M}(\boldsymbol{\Lambda}^\star_{\omega_0})\big).$$

By the positive-margin assumption, the limit point is not 0, so by continuity of the indicator map $x \mapsto \mathbb{1}\{x > 0\}$, we have $F_j(G_n) \to F_j(P)$. □

S2.4.4 Proof of Theorem 5.1

Theorem 5.1 is a direct consequence of Theorem 3.1(iii) and the following result (Thm. S2.16), which we prove below.

Theorem S2.16 (Bootstrap validity for parameters)**.** *Under Assumption S2.10,*

$$d_{\mathrm{BL}}\left(\mathrm{Law}^*\left(\sqrt{N}(\widehat{\boldsymbol{\Lambda}}^*_{\omega_0,N} - \widehat{\boldsymbol{\Lambda}}_{\omega_0,N})\right),\ \mathrm{Law}\left(\sqrt{N}(\widehat{\boldsymbol{\Lambda}}_{\omega_0,N} - \boldsymbol{\Lambda}^\star_{\omega_0})\right)\right) \xrightarrow{P} 0,$$

where $\mathrm{Law}(\cdot)$ *denotes law under* P *and* $\mathrm{Law}^*(\cdot)$ *law under* P^**.*

Proof of Theorem S2.16. Consider the argmax functional

$$\Phi(G) := \operatorname*{arg\,max}_{\boldsymbol{\Lambda}\in\Theta_K}\Big\{\int q_{\boldsymbol{\Lambda}}\,dG + \omega_0\mathcal{H}_{\mathcal{S}}(q_{\boldsymbol{\Lambda}})\Big\}.$$

By Lemma S2.13, Φ is Hadamard differentiable at $G = P$ with derivative $D\Phi_P$. By Lemma S2.5, the gradient class $\mathcal{G}$ is P-Donsker with an $L_2(P)$ envelope. Hence, by Lemma S2.12,

$$\sqrt{N}(P^*_N - P_N) \xrightarrow{w}_* \mathbb{G}_P \quad \text{in } \ell^\infty(\mathcal{G}) \text{ in } P\text{-probability.}$$

Therefore, by the bootstrap functional delta method (van der Vaart and Wellner, 1996, Thm. 3.9.11),

$$\sqrt{N}(\widehat{\boldsymbol{\Lambda}}^*_{\omega_0,N} - \widehat{\boldsymbol{\Lambda}}_{\omega_0,N}) = D\Phi_P\big(\sqrt{N}(P^*_N - P_N)\big) + o_{P^*}(1) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \boldsymbol{W}^\star_{\omega_0}),$$

conditionally on the data, in P-probability. This implies

$$d_{\mathrm{BL}}\left(\mathrm{Law}^*\left(\sqrt{N}(\widehat{\boldsymbol{\Lambda}}^*_{\omega_0,N} - \widehat{\boldsymbol{\Lambda}}_{\omega_0,N})\right),\ \mathrm{Law}\left(\sqrt{N}(\widehat{\boldsymbol{\Lambda}}_{\omega_0,N} - \boldsymbol{\Lambda}^\star_{\omega_0})\right)\right) \xrightarrow{P} 0.$$

□

S2.4.5 Proof of Proposition 5.4

Proof of Proposition 5.4. By a mean-value expansion of the score,

$$\nabla_{\boldsymbol{\Lambda}}\widehat{\mathcal{L}}_{\omega_0,N}(\widetilde{\boldsymbol{\Lambda}}_{\omega_0,N}) - \nabla_{\boldsymbol{\Lambda}}\widehat{\mathcal{L}}_{\omega_0,N}(\widehat{\boldsymbol{\Lambda}}_{\omega_0,N}) = \widehat{\boldsymbol{H}}_{\omega_0,N}(\bar{\boldsymbol{\Lambda}})(\widetilde{\boldsymbol{\Lambda}}_{\omega_0,N} - \widehat{\boldsymbol{\Lambda}}_{\omega_0,N}),$$

with $\nabla_{\boldsymbol{\Lambda}}\widehat{\mathcal{L}}_{\omega_0,N}(\widehat{\boldsymbol{\Lambda}}_{\omega_0,N}) = 0$ and there exists $c > 0$ such that $\mathrm{eig}_{\min}(\widehat{\boldsymbol{H}}_{\omega_0,N}(\bar{\boldsymbol{\Lambda}})) \geq c > 0$ with probability tending to 1 by Assumption S2.10 (B3). Hence

$$\|\widetilde{\boldsymbol{\Lambda}}_{\omega_0,N} - \widehat{\boldsymbol{\Lambda}}_{\omega_0,N}\| \leq c^{-1}\big\|\nabla_{\boldsymbol{\Lambda}}\widehat{\mathcal{L}}_{\omega_0,N}(\widetilde{\boldsymbol{\Lambda}}_{\omega_0,N})\big\| = o_P(N^{-1/2}).$$

Therefore $\sqrt{N}(\widetilde{\boldsymbol{\Lambda}}_{\omega_0,N} - \widehat{\boldsymbol{\Lambda}}_{\omega_0,N}) = o_P(1)$, and Slutsky yields the claimed Gaussian limit. For the bootstrap, the same argument holds conditionally with starred (*) quantities, giving the stated conclusions by the bootstrap delta method. □

S2.4.6 Proof of Theorem 5.3

Theorem 5.3 is obtained by combining Proposition S2.17, Theorem S2.18, and Corollary S2.20, proven below.

Mode-matching map smoothness

Proposition S2.17 (Mode-matching map smoothness)**.** *Under Assumption 5.2, there exists a neighborhood $\mathcal{V} \subset \Theta_K \cap \mathcal{G}_{\delta/2}$ of $\boldsymbol{\Lambda}^\star_{\omega_0}$ such that $\mathcal{M}$ is $\mathcal{C}^1$ on $\mathcal{V}$. Its Jacobian at $\boldsymbol{\Lambda}^\star_{\omega_0}$ is $\boldsymbol{J} = \nabla_{\boldsymbol{\Lambda}}\mathcal{M}(\boldsymbol{\Lambda}^\star_{\omega_0})$.*

Proof of Proposition S2.17. By Assumption 5.2, there exists $\delta > 0$ such that the nearest-component indices are unique at $\boldsymbol{\Lambda}^\star_{\omega_0}$. By continuity of $\boldsymbol{\Lambda} \mapsto \|\boldsymbol{\mu}_k(\boldsymbol{\Lambda}) - \boldsymbol{u}_j\|$, these indices are locally constant, so there exists a neighborhood $\mathcal{V} \subset \mathcal{G}_{\delta/2}$ of $\boldsymbol{\Lambda}^\star_{\omega_0}$ on which the matching is fixed. Write $\pi^\star(j)$ for the index matched to $\boldsymbol{u}_j$ at $\boldsymbol{\Lambda}^\star_{\omega_0}$. We have, for all $\boldsymbol{\Lambda} \in \mathcal{V}$,

$$\mathcal{M}(\boldsymbol{\Lambda}) = \big(\boldsymbol{\mu}_{\pi^\star(1)}(\boldsymbol{\Lambda})^T, \ldots, \boldsymbol{\mu}_{\pi^\star(K_0)}(\boldsymbol{\Lambda})^T\big)^T.$$

Since $\boldsymbol{m}(\boldsymbol{\Lambda}) = (\boldsymbol{\mu}_1(\boldsymbol{\Lambda})^T, \ldots, \boldsymbol{\mu}_K(\boldsymbol{\Lambda})^T)^T$ is $\mathcal{C}^1$ in a neighborhood of $\boldsymbol{\Lambda}^\star_{\omega_0}$, the composition above shows that $\mathcal{M}$ is $\mathcal{C}^1$ on $\mathcal{V}$. In particular, at $\boldsymbol{\Lambda}^\star_{\omega_0}$, its Jacobian is

$$\boldsymbol{J} = \nabla_{\boldsymbol{\Lambda}}\mathcal{M}(\boldsymbol{\Lambda}^\star_{\omega_0}) = \begin{bmatrix} \nabla_{\boldsymbol{\Lambda}}\boldsymbol{\mu}_{\pi^\star(1)}(\boldsymbol{\Lambda}^\star_{\omega_0}) \\ \vdots \\ \nabla_{\boldsymbol{\Lambda}}\boldsymbol{\mu}_{\pi^\star(K_0)}(\boldsymbol{\Lambda}^\star_{\omega_0}) \end{bmatrix}.$$

□

CLT for matched modes

Theorem S2.18 (CLT and bootstrap for $\mathcal{M}$)**.** *Under Assumptions S2.10 and 5.2, as $N \to \infty$:*

$$\sqrt{N}\big(\mathcal{M}(\widehat{\boldsymbol{\Lambda}}_{\omega_0,N}) - \mathcal{M}(\boldsymbol{\Lambda}^\star_{\omega_0})\big) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \boldsymbol{C}_{\mathcal{M}}), \quad \sqrt{N}\big(\mathcal{M}(\widehat{\boldsymbol{\Lambda}}^*_{\omega_0,N}) - \mathcal{M}(\widehat{\boldsymbol{\Lambda}}_{\omega_0,N})\big) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \boldsymbol{C}_{\mathcal{M}}),$$

conditionally on $\boldsymbol{X}_{1:N}$, in P-probability, with $\boldsymbol{C}_{\mathcal{M}} = \boldsymbol{J}\boldsymbol{W}_{\omega_0}\boldsymbol{J}^T$.

Proof of Theorem S2.18. By Theorem 3.1(iii), $\sqrt{N}(\widehat{\boldsymbol{\Lambda}}_{\omega_0,N} - \boldsymbol{\Lambda}^\star_{\omega_0}) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \boldsymbol{W}_{\omega_0})$. Furthermore, by Proposition S2.17, $\mathcal{M}$ is $\mathcal{C}^1$ at $\boldsymbol{\Lambda}^\star_{\omega_0}$ with Jacobian $\boldsymbol{J}$, so the functional delta method (van der Vaart, 1998, Thm. 20.8) yields

$$\sqrt{N}\big(\mathcal{M}(\widehat{\boldsymbol{\Lambda}}_{\omega_0,N}) - \mathcal{M}(\boldsymbol{\Lambda}^\star_{\omega_0})\big) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, \boldsymbol{J}\boldsymbol{W}_{\omega_0}\boldsymbol{J}^T\big),$$

which proves the first claim.

Finally, by the bootstrap functional delta method (van der Vaart and Wellner, 1996, Thm 3.9.11),

$$\sqrt{N}\big(\mathcal{M}(\widehat{\boldsymbol{\Lambda}}^*_{\omega_0,N}) - \mathcal{M}(\widehat{\boldsymbol{\Lambda}}_{\omega_0,N})\big) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \boldsymbol{J}\boldsymbol{W}_{\omega_0}\boldsymbol{J}^T)$$

conditionally on $\boldsymbol{X}_{1:N}$, in P-probability, which gives the second claim. □

Corollary S2.19 (Setwise convergence)**.** *As a consequence of Theorem S2.18, for any Borel set $B \subset \mathbb{R}^{dK_0}$ whose boundary has zero probability under $\mathcal{N}(0, \boldsymbol{C}_{\mathcal{M}})$, as $N \to \infty$:*

$$P^*\big(\mathcal{M}(\widehat{\boldsymbol{\Lambda}}^*_{\omega_0,N}) \in B\big) - P\big(\mathcal{M}(\widehat{\boldsymbol{\Lambda}}_{\omega_0,N}) \in B\big) \xrightarrow{P} 0.$$

In particular, percentile (or studentized) bootstrap confidence regions for each matched mode coordinate are asymptotically valid.

Proof of Corollary S2.19. This is directly implied by Theorem S2.18 and the Portmanteau theorem (van der Vaart, 1998, Lem. 2.2). □

Per-mode projections

Corollary S2.20 (Per-mode projections)**.** *As a consequence of Theorem S2.18, for the j-th matched mode $\mathcal{M}_j := \boldsymbol{E}_j\mathcal{M}$ with $\boldsymbol{E}_j$ extracting the j-th block, as $N \to \infty$:*

$$\sqrt{N}\big(\mathcal{M}_j(\widehat{\boldsymbol{\Lambda}}_{\omega_0,N}) - \mathcal{M}_j(\boldsymbol{\Lambda}^\star_{\omega_0})\big) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \boldsymbol{C}_j), \quad \sqrt{N}\big(\mathcal{M}_j(\widehat{\boldsymbol{\Lambda}}^*_{\omega_0,N}) - \mathcal{M}_j(\widehat{\boldsymbol{\Lambda}}_{\omega_0,N})\big) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \boldsymbol{C}_j),$$

and percentile bootstrap ellipses for $\mathcal{M}_j$ are asymptotically valid.

Proof of Corollary S2.20. The asymptotic normality results are directly implied by Theorem S2.18 and the linearity of the extractor $\boldsymbol{E}_j$. Then, the ellipses are asymptotically valid as a consequence of the setwise convergence given by Corollary S2.19. □

S2.4.7 Proof of Proposition S1.13

Proof of Proposition S1.13. By Lemma S2.15, F_j is continuous at P with respect to weak convergence.

(i) Conditional LLN. Given the data, $\{F_j(P_N^{*(\ell)})\}_{\ell=1}^L$ are i.i.d. Bernoulli with mean $\mathbb{E}^*[F_j(P_N^*)] = \tau_{j,N}$ and finite variance. Hence, by the strong law, $s_j \to \tau_{j,N}$ almost surely

under the bootstrap (a.s.*), and $\mathrm{Var}^*(s_j) = \mathrm{Var}^*(F_j(P_N^*))/L \to 0$. Hoeffding's inequality yields the tail bound.

(ii) Large-N limits. Since $P_N \xrightarrow{w} P$ and F_j is continuous at P, we have $F_j(P_N) \to F_j(P)$ in probability, so $\pi_{j,N} \to \pi_j := F_j(P)$. For the bootstrap mean, standard nonparametric bootstrap consistency at continuity points gives $F_j(P_N^*) \xrightarrow{P} F_j(P)$ in P-probability. Because $F_j \in [0,1]$ is bounded, conditional bounded convergence implies $\mathbb{E}^*[F_j(P_N^*)] \xrightarrow{P} F_j(P) = \pi_j$.

Finally, for any $L = L_N \to \infty$ as $N \to \infty$,

$$|s_j - \pi_j| \le |s_j - \tau_{j,N}| + |\tau_{j,N} - \pi_j| \xrightarrow{P} 0.$$

By (i) and (ii), we obtain the last claim $s_j \xrightarrow{P} \pi_j$. □

S2.5 Proofs for Section S1.2

S2.5.1 Proof of Proposition S1.1

Proof of Proposition S1.1. Write $q^{\mathcal{S}} = q/\psi$, $\varepsilon = 1 - \psi$. On the one hand, we have

$$\left|\mathbb{E}_{q^{\mathcal{S}}}[p] - \mathbb{E}_q[p]\right| = \left|\frac{1}{\psi} - 1\right| \int_{\mathcal{S}} pq \le M\varepsilon.$$

On the other hand, we have $\mathcal{H}_{\mathcal{S}}(q^{\mathcal{S}}) = \frac{1}{\psi}\mathcal{H}_{\mathcal{S}}(q) + \log\psi$. Hence,

$$\mathcal{H}_{\mathcal{S}}(q^{\mathcal{S}}) - \mathcal{H}_{\mathcal{S}}(q) = \left(\frac{1}{\psi} - 1\right)\mathcal{H}_{\mathcal{S}}(q) + \log\psi = \frac{\varepsilon}{\psi}\mathcal{H}_{\mathcal{S}}(q) + \log(1-\varepsilon).$$

On a compact parameter set and bounded $\mathcal{S}$, $C_q = \sup_{\boldsymbol{\lambda}} \sup_{\boldsymbol{x}\in\mathcal{S}} \left|\log q_{\boldsymbol{\lambda}}(\boldsymbol{x})\right| < \infty$ is well-defined and $|\mathcal{H}_{\mathcal{S}}(q)| \le C_q \int_{\mathcal{S}} q = C_q\psi$. Therefore,

$$\left|\mathcal{H}_{\mathcal{S}}(q^{\mathcal{S}}) - \mathcal{H}_{\mathcal{S}}(q)\right| \le C_q\varepsilon + |\log(1-\varepsilon)|.$$

Combining the two inequalities, we obtain the pointwise bound. The uniform bound follows by taking suprema over Θ. □

S2.5.2 Proof of Theorem S1.2

Before proving Theorem S1.2, we derive three supporting lemmas.

Lemma S2.21 (Argmax stability)**.** *Let Θ be compact and let $F, G : \Theta \to \mathbb{R}$ be continuous functions. Assume F has a maximizer $\boldsymbol{\lambda}^\star$ and that for some $\rho > 0$ the value gap*

$$m(\rho) := F(\boldsymbol{\lambda}^\star) - \sup_{\|\boldsymbol{\lambda}-\boldsymbol{\lambda}^\star\|\ge\rho} F(\boldsymbol{\lambda}) > 0.$$

If $\|G - F\|_\infty := \sup_{\boldsymbol{\lambda}\in\Theta} |G(\boldsymbol{\lambda}) - F(\boldsymbol{\lambda})| \le m(\rho)/3$*, then every maximizer of* G *lies in* $B(\boldsymbol{\lambda}^\star, \rho)$*.*

If, in addition, $\boldsymbol{\lambda}^\star$ *is the unique interior maximizer of* F*, then there exists* $\rho_0 > 0$ *such that the same bound holds for every* $0 < \rho \le \rho_0$*, then we have* $\arg\max_{\boldsymbol{\lambda}\in\Theta} G(\boldsymbol{\lambda}) = \{\boldsymbol{\lambda}^\star\}$*.*

Proof of Lemma S2.21. Fix $\rho > 0$ with $m(\rho) > 0$ and set $\delta := \|G - F\|_\infty$. For any $\boldsymbol{\lambda}$ with $\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^\star\| \ge \rho$,

$$G(\boldsymbol{\lambda}) \le F(\boldsymbol{\lambda}) + \delta \le F(\boldsymbol{\lambda}^\star) - m(\rho) + \delta,$$

while $G(\boldsymbol{\lambda}^\star) \ge F(\boldsymbol{\lambda}^\star) - \delta$. If $\delta < m(\rho)/2$ (hence, under $\delta \le m(\rho)/3$), then $G(\boldsymbol{\lambda}^\star) > G(\boldsymbol{\lambda})$ for all $\|\boldsymbol{\lambda} - \boldsymbol{\lambda}^\star\| \ge \rho$, so every maximizer of G lies in $B(\boldsymbol{\lambda}^\star, \rho)$.

If $\boldsymbol{\lambda}^\star$ is the unique interior maximizer of F, choose $\rho_0 > 0$ so that $B(\boldsymbol{\lambda}^\star, \rho_0)$ contains no other maximizer of F. For any $0 < \rho \le \rho_0$, the above argument ensures that every maximizer of G lies in $B(\boldsymbol{\lambda}^\star, \rho)$. Since $\boldsymbol{\lambda}^\star$ is the only maximizer of F in this ball, the strict inequality above then implies $\arg\max_{\boldsymbol{\lambda}\in\Theta} G = \{\boldsymbol{\lambda}^\star\}$. □

Lemma S2.22 (Uniform tail)**.** *Fix* $\omega > 0$*. Assume the uniform margin*

$$\operatorname{dist}\big(\mu(\boldsymbol{\lambda}), \partial\mathcal{S}\big) \ge M_0 \sqrt{\operatorname{eig}_{\max}(\Sigma(\boldsymbol{\lambda}))}, \qquad \forall \boldsymbol{\lambda} \in \Theta.$$

Then the uniform tail bound

$$\varepsilon_{\max} := \sup_{\boldsymbol{\lambda}\in\Theta} \Big(1 - \int_{\mathcal{S}} q_{\boldsymbol{\lambda}}(\boldsymbol{x}) d\boldsymbol{x}\Big) \le C e^{-M_0^2/2}$$

holds for some constant $C > 0$*.*

Proof of Lemma S2.22. Let $\boldsymbol{\lambda} \in \Theta$, $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and $r := M_0 \sqrt{\operatorname{eig}_{\max}(\boldsymbol{\Sigma})}$. The margin condition $\operatorname{dist}(\boldsymbol{\mu}(\boldsymbol{\lambda}), \partial\mathcal{S}) \ge r$ implies $\overline{B(\boldsymbol{\mu}, r)} \subset \mathcal{S}$, so $\varepsilon(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathbb{P}(\boldsymbol{X} \notin \mathcal{S}) \le \mathbb{P}(\|\boldsymbol{X} - \boldsymbol{\mu}\| \ge r)$. Since $\boldsymbol{\Sigma} \preceq \operatorname{eig}_{\max}(\boldsymbol{\Sigma})\boldsymbol{I}$, for $\boldsymbol{Z} \sim \mathcal{N}(0, \boldsymbol{I})$, there exists $C > 0$ such that

$$\mathbb{P}(\|\boldsymbol{X} - \boldsymbol{\mu}\| \ge r) \le \mathbb{P}(\|\boldsymbol{Z}\| \ge r/\sqrt{\operatorname{eig}_{\max}(\boldsymbol{\Sigma})}) = \mathbb{P}(\|\boldsymbol{Z}\| \ge M_0) \le C e^{-M_0^2/2},$$

by chi-squared tail bounds for $\|\boldsymbol{Z}\|^2$. Thus, $\varepsilon_{\max} = \sup_{\boldsymbol{\lambda}\in\Theta} \varepsilon(\boldsymbol{\lambda}) \le C e^{-M_0^2/2}$. □

Proof of Theorem S1.2. By Lemma S2.22, $\varepsilon_{\max} \le C e^{-M_0^2/2}$. Proposition S1.1 then yields the uniform bound

$$\sup_{\boldsymbol{\lambda}\in\Theta} \big|G_\omega(\boldsymbol{\lambda}) - F_\omega(\boldsymbol{\lambda})\big| \le \delta_\omega(\varepsilon_{\max}).$$

Fix $\rho > 0$ and set $m(\rho) := F_\omega(\boldsymbol{\lambda}_\omega^\star) - \sup_{\|\boldsymbol{\lambda} - \boldsymbol{\lambda}_\omega^\star\| \ge \rho} F_\omega(\boldsymbol{\lambda}) > 0$. Choose M_0, hence $\varepsilon_{\max}$, large enough that $\delta_\omega(\varepsilon_{\max}) < m(\rho)/3$. By Lemma S2.21, every maximizer of G_ω lies in $B(\boldsymbol{\lambda}_\omega^\star, \rho)$. If $\boldsymbol{\lambda}_\omega^\star$ is a unique interior maximizer of F_ω, the argmaxes coincide. □

S3 Algorithms and variants

This section details the update rules for GERVE for different variational families and provides pseudo-codes for implementation. For each case, we:

- Recall the parameterization and its link to natural gradients.
- Present the specific update rules and their mini-batch version.
- Give a pseudo-code for practical use.
- Provide derivation notes for readers interested in the intermediate steps.

S3.1 Natural-gradient update rule

For distributions in an exponential family or MCEF, we can use (5) (Lin et al., 2019; Khan and Rue, 2023) so that the natural-gradient update rule writes:

$$\boldsymbol{\lambda}_{t+1} = \boldsymbol{\lambda}_t + \rho_t \nabla_{\boldsymbol{M}} \widehat{\mathcal{L}}_{\omega_t, N}(\boldsymbol{\lambda})|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}_t},$$

which will be the starting point for all the following derivations.

S3.2 Fixed-covariance Gaussian case

Parameterization. We consider $q_{\boldsymbol{\lambda}}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, s^{-1}\boldsymbol{I})$ with fixed precision $s > 0$.

- Natural parameter: $\boldsymbol{\lambda} = s\boldsymbol{\mu}$;
- Sufficient statistic: $\boldsymbol{x}$;
- Expectation parameter: $\boldsymbol{\mu} = \mathbb{E}_{q_{\boldsymbol{\lambda}}}[\boldsymbol{X}]$.

Only $\boldsymbol{\mu}$ is optimized.

GERVE update rule. Considering $\nabla_{\boldsymbol{\mu}} \mathcal{H}_{\mathcal{S}}(q_{\boldsymbol{\lambda}}) \approx \nabla_{\boldsymbol{\mu}} \mathcal{H}(q_{\boldsymbol{\lambda}}) = 0$, the fixed-covariance Gaussian GERVE update rule is:

$$\boxed{\boldsymbol{\mu}_{t+1} = \boldsymbol{\mu}_t + \frac{\rho_t}{N} \sum_{i=1}^{N} (\boldsymbol{X}_i - \boldsymbol{\mu}_t) q_{\boldsymbol{\lambda}_t}(\boldsymbol{X}_i).}$$

This is equivalent to gradient ascent on the kernel density estimate

$$p_h(\boldsymbol{\mu}) = \frac{1}{N} \sum_{i=1}^{N} \varphi_h(\boldsymbol{X}_i - \boldsymbol{\mu}),$$

where φ_h is the Gaussian kernel with bandwidth $h = s^{-1}$. This reduces to the classic Gaussian mean-shift update when the step sizes are chosen adaptively $\rho_t = p_h(\boldsymbol{\mu}_t)^{-1}$.

Mini-batch implementation. For a batch $\boldsymbol{X}_{1:B}^{(t)}$ sampled uniformly with replacement from the dataset, i.e. $J_1^{(t)}, \ldots, J_B^{(t)} \overset{\text{i.i.d.}}{\sim} \text{Unif}([N])$, $\boldsymbol{X}_{1:B}^{(t)} := (\boldsymbol{X}_{J_1^{(t)}}, \ldots, \boldsymbol{X}_{J_B^{(t)}})$, define:

$$\widehat{\boldsymbol{g}}_B^{(t)} = \frac{s}{B} \sum_{i=1}^{B} (\boldsymbol{X}_i^{(t)} - \boldsymbol{\mu}_t) q_{\boldsymbol{\lambda}_t}(\boldsymbol{X}_i^{(t)}).$$

Conditioned on $\boldsymbol{X}_{1:N}$, $\widehat{\boldsymbol{g}}_B^{(t)}$ is an unbiased estimator of $\nabla_{\boldsymbol{\mu}} \widehat{\mathcal{L}}_{\omega_t,N}(\boldsymbol{\lambda})|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}_t}$.

The mini-batch version of the update rule is:

$$\boxed{\boldsymbol{\mu}_{t+1} = \boldsymbol{\mu}_t + \rho_t s^{-1} \widehat{\boldsymbol{g}}_B^{(t)}.}$$

Algorithm. Algorithm 3 presents the pseudo-code corresponding to this implementation of GERVE.

Algorithm 3: Fixed-covariance Gaussian GERVE (with mini-batches)

```
Given samples X_{1:N}.
Set T, B, s, λ_1, ρ_{1:T}.
for t = 1:T do
    Sample J_{1:B}^{(t)} ~ Unif([N]) i.i.d., set X_i^{(t)} ← X_{J_i^{(t)}} for i = 1:B.
    Compute ĝ_B^{(t)} from X_{1:B}^{(t)}, λ_t.
    Update μ_{t+1} = μ_t + ρ_t s^{-1} ĝ_B^{(t)}.
end
return λ_{T+1} (= μ_{T+1}).
```

Derivation details. From the natural-gradient rule for exponential families with $\boldsymbol{\lambda} = s\boldsymbol{\mu}$, we have:

$$s\boldsymbol{\mu}_{t+1} = s\boldsymbol{\mu}_t + \rho_t \nabla_{\boldsymbol{\mu}} \widehat{\mathcal{L}}_{\omega_t,N}(\boldsymbol{\lambda})|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}_t}.$$

The gradient of the empirical objective is:

$$\nabla_{\boldsymbol{\mu}} \widehat{\mathcal{L}}_{\omega_t,N}(\boldsymbol{\lambda}) = \frac{s}{N} \sum_{i=1}^{N} (\boldsymbol{X}_i - \boldsymbol{\mu}) q_{\boldsymbol{\lambda}}(\boldsymbol{X}_i) + \omega \nabla_{\boldsymbol{\mu}} \mathcal{H}_{\mathcal{S}}(q_{\boldsymbol{\lambda}}),$$

which yields the update rule above.

S3.3 Gaussian mixture case

Parameterization. For K components:

$$q_{\boldsymbol{\Lambda}}(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k q_{\boldsymbol{\lambda}_k}(\boldsymbol{x}),$$

where the weights $\pi_1, \ldots, \pi_K$ sum to 1, and each $q_{\boldsymbol{\lambda}_k}(\boldsymbol{x})$ is a Gaussian density with natural parameters $\boldsymbol{\lambda}_k$.

- MCEF natural parameters:

$$\boldsymbol{\Lambda} = (v_1, \ldots, v_{K-1}, \boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_K),$$

  where $v_k := \log(\pi_k/\pi_K)$;
- MCEF expectation parameters:

$$\boldsymbol{M} = (\pi_1, \ldots, \pi_{K-1}, \boldsymbol{M}_1, \ldots, \boldsymbol{M}_K),$$

  where $\boldsymbol{M}_k := (\boldsymbol{m}_k^{(1)}, \boldsymbol{m}_k^{(2)}) = (\pi_k \boldsymbol{\mu}_k, \pi_k(\boldsymbol{S}_k^{-1} + \boldsymbol{\mu}_k \boldsymbol{\mu}_k^T))$.

GERVE update rules. The Gaussian mixture GERVE update rules are:

$$\begin{aligned}
\boldsymbol{S}_{k,t+1} &= \boldsymbol{S}_{k,t} - \frac{\rho_t}{N} \boldsymbol{S}_{k,t} \sum_{i=1}^{N} ((\boldsymbol{X}_i - \boldsymbol{\mu}_{k,t})(\boldsymbol{X}_i - \boldsymbol{\mu}_{k,t})^T \boldsymbol{S}_{k,t} - \boldsymbol{I}) q_{\boldsymbol{\lambda}_{k,t}}(\boldsymbol{X}_i) \\
&\quad - \rho_t \omega_t \nabla_{\boldsymbol{S}_k^{-1}} \mathcal{H}_{\mathcal{S}}(q_{\boldsymbol{\Lambda}})|_{\boldsymbol{\Lambda}=\boldsymbol{\Lambda}_t}, \\
\boldsymbol{\mu}_{k,t+1} &= \boldsymbol{\mu}_{k,t} + \frac{\rho_t}{N} \boldsymbol{S}_{k,t+1}^{-1} \boldsymbol{S}_{k,t} \sum_{i=1}^{N} (\boldsymbol{X}_i - \boldsymbol{\mu}_{k,t}) q_{\boldsymbol{\lambda}_{k,t}}(\boldsymbol{X}_i) + \rho_t \omega_t \nabla_{\boldsymbol{\mu}_k} \mathcal{H}_{\mathcal{S}}(q_{\boldsymbol{\Lambda}})|_{\boldsymbol{\Lambda}=\boldsymbol{\Lambda}_t}, \\
v_{k,t+1} &= v_{k,t} + \frac{\rho_t}{N} \sum_{i=1}^{N} (q_{\boldsymbol{\lambda}_{k,t}}(\boldsymbol{X}_i) - q_{\boldsymbol{\lambda}_{K,t}}(\boldsymbol{X}_i)) - \rho_t \omega_t \nabla_{\pi_k} \mathcal{H}_{\mathcal{S}}(q_{\boldsymbol{\Lambda}})|_{\boldsymbol{\Lambda}=\boldsymbol{\Lambda}_t}.
\end{aligned}$$

Mini-batch and Monte-Carlo implementation. For a batch $\boldsymbol{X}_{1:B}^{(t)}$ sampled uniformly with replacement from the dataset, i.e. $J_1^{(t)}, \ldots, J_B^{(t)} \overset{\text{i.i.d.}}{\sim} \text{Unif}([N]), \boldsymbol{X}_{1:B}^{(t)} := (\boldsymbol{X}_{J_1^{(t)}}, \ldots, \boldsymbol{X}_{J_B^{(t)}})$, define:

$$\begin{aligned}
\widehat{f}_{k,B}^{(t)} &= \frac{1}{B} \sum_{i=1}^{B} (q_{\boldsymbol{\lambda}_{k,t}}(\boldsymbol{X}_i^{(t)}) - q_{\boldsymbol{\lambda}_{K,t}}(\boldsymbol{X}_i^{(t)})), \\
\widehat{\boldsymbol{g}}_{k,B}^{(t)} &= \frac{\pi_{k,t}}{B} \boldsymbol{S}_{k,t} \sum_{i=1}^{B} (\boldsymbol{X}_i^{(t)} - \boldsymbol{\mu}_{k,t}) q_{\boldsymbol{\lambda}_{k,t}}(\boldsymbol{X}_i^{(t)}), \\
\widehat{\boldsymbol{H}}_{k,B}^{(t)} &= \frac{\pi_{k,t}}{2B} \boldsymbol{S}_{k,t} \sum_{i=1}^{B} ((\boldsymbol{X}_i^{(t)} - \boldsymbol{\mu}_{k,t})(\boldsymbol{X}_i^{(t)} - \boldsymbol{\mu}_{k,t})^T \boldsymbol{S}_{k,t} - \boldsymbol{I}) q_{\boldsymbol{\lambda}_{k,t}}(\boldsymbol{X}_i^{(t)}).
\end{aligned}$$

Conditioned on $\boldsymbol{X}_{1:N}$, $\widehat{\boldsymbol{g}}_{k,B}^{(t)}$, $\widehat{\boldsymbol{H}}_{k,B}^{(t)}$, and $\widehat{f}_{k,B}^{(t)}$ are unbiased estimators of their full-data counterparts.

Entropy derivatives must also be estimated. For some size B_e, draw additional Monte-Carlo samples $\boldsymbol{Z}_{1:B_e}^{(k,t)} \sim q_{\boldsymbol{\lambda}_{k,t}}$ to compute $\widehat{\boldsymbol{\eta}}_{k,B_e}^{(t)}$, $\widehat{\boldsymbol{\gamma}}_{k,B_e}^{(t)}$, and $\widehat{\varphi}_{k,B_e}^{(t)}$ (see details below).

We obtain the practical updates:

$$\boxed{\begin{aligned}
\boldsymbol{S}_{k,t+1} &= \boldsymbol{S}_{k,t} - \frac{2\rho_t}{\pi_{k,t}}(\widehat{\boldsymbol{H}}_{k,B}^{(t)} + \omega_t \widehat{\boldsymbol{\eta}}_{k,B_e}^{(t)}), \\
\boldsymbol{\mu}_{k,t+1} &= \boldsymbol{\mu}_{k,t} + \rho_t \boldsymbol{S}_{k,t+1}^{-1}(\widehat{\boldsymbol{g}}_{k,B}^{(t)} + \omega_t \widehat{\boldsymbol{\gamma}}_{k,B_e}^{(t)}), \\
v_{k,t+1} &= v_{k,t} + \rho_t(\widehat{f}_{k,B}^{(t)} + \omega_t \widehat{\varphi}_{k,B_e}^{(t)}).
\end{aligned}}$$

Algorithm. Algorithm 4 depicts our proposed implementation of Gaussian mixture GERVE.

Algorithm 4: Gaussian mixture GERVE (with mini-batches)

```
Given samples X_{1:N}.
Set T, B, B_e, K, Λ_1, ρ_{1:T}, ω_{1:T}.
for t = 1:T do
    Sample J_{1:B}^{(t)} ~i.i.d.~ Unif([N]), set X_i^{(t)} ← X_{J_i^{(t)}} for i = 1:B.
    for k = 1:K do
        Compute ĝ_{k,B}^{(t)}, Ĥ_{k,B}^{(t)} from X_{1:B}^{(t)}, Λ_t.
        Sample Z_{1:B_e}^{(k,t)} ~i.i.d.~ N(μ_{k,t}, S_{k,t}^{-1}).
        Compute γ̂_{k,B_e}^{(t)}, η̂_{k,B_e}^{(t)} from Z_{1:B_e}^{(k,t)}, Λ_t.
        Update S_{k,t+1}, μ_{k,t+1}.
    end
    for k = 1:(K−1) do
        Compute f̂_{k,B}^{(t)}, φ̂_{k,B_e}^{(t)} from X_{1:B}^{(t)}, Z_{1:B_e}^{(k,t)}, Z_{1:B_e}^{(K,t)}, Λ_t.
        Update v_{k,t+1}.
    end
end
return Λ_{T+1}.
```

Derivation details. The natural-gradient update rules are:

$$\begin{aligned}
v_{k,t+1} &= v_{k,t} + \rho_t \nabla_{\pi_k} \widehat{\mathcal{L}}_{\omega_t,N}(\boldsymbol{\Lambda})|_{\boldsymbol{\Lambda}=\boldsymbol{\Lambda}_t}, \\
\boldsymbol{S}_{k,t+1}\boldsymbol{\mu}_{k,t+1} &= \boldsymbol{S}_{k,t}\boldsymbol{\mu}_{k,t} + \rho_t \nabla_{\boldsymbol{m}_k^{(1)}} \widehat{\mathcal{L}}_{\omega_t,N}(\boldsymbol{\Lambda})|_{\boldsymbol{\Lambda}=\boldsymbol{\Lambda}_t}, \\
-\frac{1}{2}\boldsymbol{S}_{k,t+1} &= -\frac{1}{2}\boldsymbol{S}_{k,t} + \rho_t \nabla_{\boldsymbol{m}_k^{(2)}} \widehat{\mathcal{L}}_{\omega_t,N}(\boldsymbol{\Lambda})|_{\boldsymbol{\Lambda}=\boldsymbol{\Lambda}_t}.
\end{aligned}$$

We rearrange them to obtain:

$$\begin{aligned}
\boldsymbol{\mu}_{k,t+1} &= \boldsymbol{\mu}_{k,t} + \rho_t \boldsymbol{S}_{k,t+1}^{-1}(\nabla_{\boldsymbol{m}_k^{(1)}} \widehat{\mathcal{L}}_{\omega_t,N}(\boldsymbol{\Lambda})|_{\boldsymbol{\Lambda}=\boldsymbol{\Lambda}_t} + 2(\nabla_{\boldsymbol{m}_k^{(2)}} \widehat{\mathcal{L}}_{\omega_t,N}(\boldsymbol{\Lambda})|_{\boldsymbol{\Lambda}=\boldsymbol{\Lambda}_t})\boldsymbol{\mu}_{k,t}),\\
\boldsymbol{S}_{k,t+1} &= \boldsymbol{S}_{k,t} - 2\rho_t \nabla_{\boldsymbol{m}_k^{(2)}} \widehat{\mathcal{L}}_{\omega_t,N}(\boldsymbol{\Lambda})|_{\boldsymbol{\Lambda}=\boldsymbol{\Lambda}_t},\\
v_{k,t+1} &= v_{k,t} + \rho_t \nabla_{\pi_k} \widehat{\mathcal{L}}_{\omega_t,N}(\boldsymbol{\Lambda})|_{\boldsymbol{\Lambda}=\boldsymbol{\Lambda}_t}.
\end{aligned}$$

We apply the chain rule to express the gradients w.r.t. the expectation parameters as gradients w.r.t. $\boldsymbol{\mu}_k$ and $\boldsymbol{S}_k^{-1}$:

$$\begin{aligned}
\nabla_{\boldsymbol{m}_k^{(1)}} \widehat{\mathcal{L}}_{\omega_t,N}(\boldsymbol{\Lambda}) &= \frac{1}{\pi_k}\left(\nabla_{\boldsymbol{\mu}_k} \widehat{\mathcal{L}}_{\omega_t,N}(\boldsymbol{\Lambda}) - 2(\nabla_{\boldsymbol{S}_k^{-1}} \widehat{\mathcal{L}}_{\omega_t,N}(\boldsymbol{\Lambda}))\boldsymbol{\mu}_k\right),\\
\nabla_{\boldsymbol{m}_k^{(2)}} \widehat{\mathcal{L}}_{\omega_t,N}(\boldsymbol{\Lambda}) &= \frac{1}{\pi_k} \nabla_{\boldsymbol{S}_k^{-1}} \widehat{\mathcal{L}}_{\omega_t,N}(\boldsymbol{\Lambda}).
\end{aligned}$$

These gradients are

$$\begin{aligned}
\nabla_{\boldsymbol{\mu}_k} \widehat{\mathcal{L}}_{\omega_t,N}(\boldsymbol{\Lambda}) &= \nabla_{\boldsymbol{\mu}_k}\left(\frac{1}{N}\sum_{i=1}^N q_{\boldsymbol{\Lambda}}(\boldsymbol{X}_i)\right) + \omega_t \nabla_{\boldsymbol{\mu}_k} \mathcal{H}_{\mathcal{S}}(q_{\boldsymbol{\Lambda}})\\
&= \frac{\pi_k}{N}\boldsymbol{S}_k \sum_{i=1}^N (\boldsymbol{X}_i - \boldsymbol{\mu}_k) q_{\boldsymbol{\lambda}_k}(\boldsymbol{X}_i) + \omega_t \nabla_{\boldsymbol{\mu}_k} \mathcal{H}_{\mathcal{S}}(q_{\boldsymbol{\Lambda}}),
\end{aligned}$$

and

$$\begin{aligned}
\nabla_{\boldsymbol{S}_k^{-1}} \widehat{\mathcal{L}}_{\omega_t,N}(\boldsymbol{\Lambda}) &= \nabla_{\boldsymbol{S}_k^{-1}}\left(\frac{1}{N}\sum_{i=1}^N q_{\boldsymbol{\Lambda}}(\boldsymbol{X}_i)\right) + \omega_t \nabla_{\boldsymbol{S}_k^{-1}} \mathcal{H}_{\mathcal{S}}(q_{\boldsymbol{\Lambda}})\\
&= \frac{\pi_k}{2N}\boldsymbol{S}_k \sum_{i=1}^N ((\boldsymbol{X}_i - \boldsymbol{\mu}_k)(\boldsymbol{X}_i - \boldsymbol{\mu}_k)^T \boldsymbol{S}_k - \boldsymbol{I}) q_{\boldsymbol{\lambda}_k}(\boldsymbol{X}_i) + \omega_t \nabla_{\boldsymbol{S}_k^{-1}} \mathcal{H}_{\mathcal{S}}(q_{\boldsymbol{\Lambda}}).
\end{aligned}$$

Given that $\pi_K = 1 - \sum_{k=1}^K \pi_k$, we also have

$$\begin{aligned}
\nabla_{\pi_k} \widehat{\mathcal{L}}_{\omega_t,N}(\boldsymbol{\Lambda}) &= \nabla_{\pi_k}\left(\frac{1}{N}\sum_{i=1}^N q_{\boldsymbol{\Lambda}}(\boldsymbol{X}_i)\right) + \omega_t \nabla_{\pi_k} \mathcal{H}_{\mathcal{S}}(q_{\boldsymbol{\Lambda}})\\
&= \frac{1}{N}\sum_{i=1}^N (q_{\boldsymbol{\lambda}_k}(\boldsymbol{X}_i) - q_{\boldsymbol{\lambda}_K}(\boldsymbol{X}_i)) + \omega_t \nabla_{\pi_k} \mathcal{H}_{\mathcal{S}}(q_{\boldsymbol{\Lambda}}).
\end{aligned}$$

Combining these expressions, we find the Gaussian mixture GERVE update rules.

Entropy derivatives (for Monte-Carlo estimation). The entropy derivatives do not admit a closed form, but they admit an integral representation. For the derivatives w.r.t. $\boldsymbol{\mu}_k$ and $\boldsymbol{S}_k^{-1}$, the following expressions can be used:

$$\nabla_{\boldsymbol{S}_k^{-1}} \mathcal{H}_{\mathcal{S}}(q_{\boldsymbol{\Lambda}}) = \frac{\pi_k}{2} \boldsymbol{S}_k \mathbb{E}_{q_{\boldsymbol{\lambda}_k}}[((\boldsymbol{X} - \boldsymbol{\mu}_k)(\boldsymbol{X} - \boldsymbol{\mu}_k)^T \boldsymbol{S}_k - \boldsymbol{I}) \log q_{\boldsymbol{\Lambda}}(\boldsymbol{X}) \mathbb{1}_{\mathcal{S}}(\boldsymbol{X})]$$

and

$$\nabla_{\boldsymbol{\mu}_k} \mathcal{H}_{\mathcal{S}}(q_{\boldsymbol{\Lambda}}) = \pi_k \boldsymbol{S}_k \mathbb{E}_{q_{\boldsymbol{\lambda}_k}}[(\boldsymbol{X} - \boldsymbol{\mu}_k) \log q_{\boldsymbol{\Lambda}}(\boldsymbol{X}) \mathbb{1}_{\mathcal{S}}(\boldsymbol{X})].$$

Finally, for the derivative w.r.t. π_k, we use:

$$\nabla_{\pi_k} \mathcal{H}_{\mathcal{S}}(q_{\boldsymbol{\Lambda}}) = \mathbb{E}_{q_{\boldsymbol{\lambda}_k}}[\log q_{\boldsymbol{\Lambda}}(\boldsymbol{X}) \mathbb{1}_{\mathcal{S}}(\boldsymbol{X})] - \mathbb{E}_{q_{\boldsymbol{\lambda}_K}}[\log q_{\boldsymbol{\Lambda}}(\boldsymbol{X}) \mathbb{1}_{\mathcal{S}}(\boldsymbol{X})].$$

S3.4 Computational considerations

Recall GERVE's complexity for Gaussian mixtures with K components (Sec. 4.4):

- Full covariances: $O(d^3BKT)$.
- Diagonal, isotropic or fixed covariances: $O(dBKT)$.

Simplified families sometimes retain most of the algorithm's qualitative behavior but improve scalability.

This is comparable to related methods:

- *Mean-shift mode-seeking.* A single run of classical mean-shift with fixed, spherical bandwidths and mini-batches costs $O(dBT)$. Running from K initializations (to target K modes) yields $O(dBKT)$, comparable to fixed-covariance GERVE. However, unlike GERVE, these independent trajectories lack a repulsion mechanism and may fail to explore the full modal structure.
- *Mean-shift clustering.* Since mean-shift clustering launches one trajectory per data point, the cost is $O(dNBT)$ for spherical bandwidths. This can be substantially more expensive than GERVE, which only launches K trajectories to find the cluster centroids.
- *EM for Gaussian mixtures (GMM-EM).* The EM algorithm (Dempster et al., 1977; McLachlan and Peel, 2000) with full covariances has cost $O\big((d^2B + d^3)KT\big)$ per run, due to matrix inversions and determinant computations. For diagonal covariances, it reduces to $O(dBKT)$, which is comparable to fixed-covariance GERVE. In contrast, GERVE uses unified natural-gradient steps rather than alternating E- and M-steps, and does not require likelihood maximization, making it applicable in likelihood-free settings.

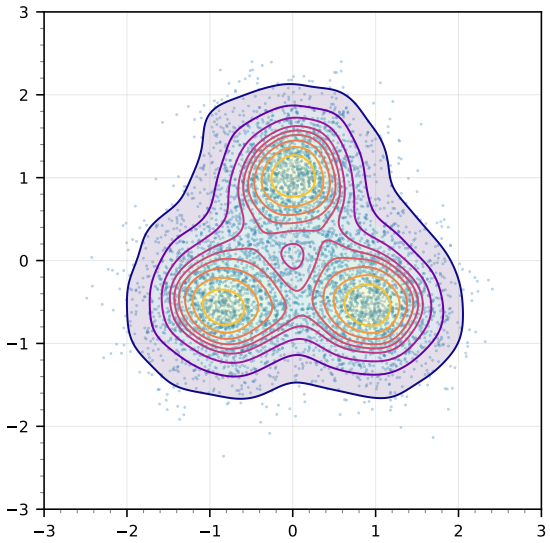

Figure S2: Sample of $N = 6000$ points from a three-component Gaussian mixture whose means are located at the nodes of an equilateral triangle. Background shading: Gaussian KDE with bandwidth h selected by Scott's rule (Scott, 1992), revealing three high-density regions.

S4 Simulations and benchmark

S4.1 Details on the clustering example of Section 6.1

Sample. The dataset consists of $N = 6000$ points from a 2D Gaussian mixture with three equally weighted components with means on the nodes of an equilateral triangle, at $(0, 1)$, $(\cos(\pi/6), -0.5)$, $(-\cos(\pi/6), -0.5)$ and isotropic covariance $\sigma^2 \boldsymbol{I}$ with $\sigma^2 = 0.25$. Figure S2 shows a sample and a kernel density estimate (KDE), revealing $J = 3$ regions of high density.

GERVE hyperparameters. The mixture parameters are initialized with $\boldsymbol{\mu}_{1,1}, \ldots, \boldsymbol{\mu}_{K,1} \sim \text{Uniform}([-2, 0]^2)$, and for all $k \in [K]$, $\boldsymbol{\Sigma}_{k,1} = 2\boldsymbol{I}$, $\pi_{k,1} = 1/K$. Table S1 gives the other hyperparameters used for GERVE in this clustering task.

Additional figures. Figure S3 shows results for $K = 3$ including an additional comparison with k-means for $K = 7$.

Table S1: Hyperparameters used for GERVE in the clustering simulations.

Hyperparameters	Values
T	40000
B	1000
K	$\{3, 7\}$
ω_t	$50/t^{1.1} + 0.004$
ρ_t	$10^{-4}(50/\omega_t)^{0.7}$

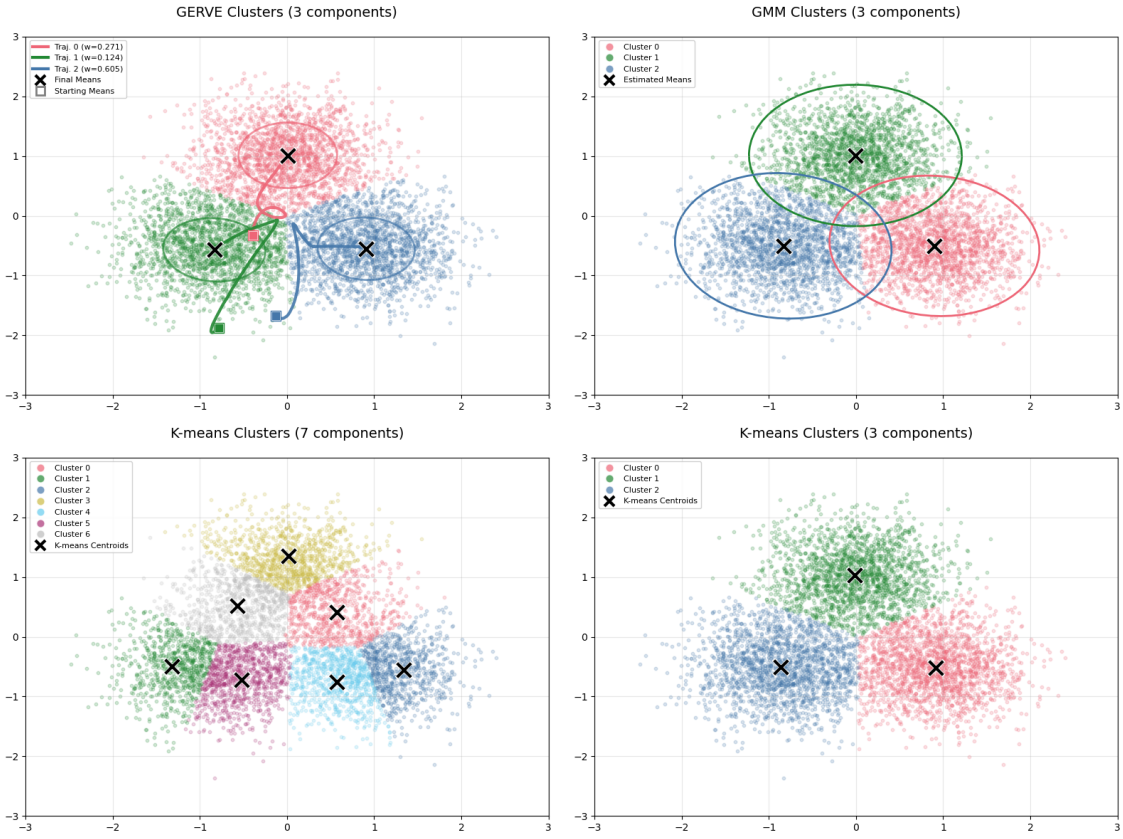

Figure S3: Top row: Clustering with $K = 3$ for GERVE (left) and GMM-EM (right). Ellipses represent the covariance matrices of the Gaussian components. Bottom row: Clustering with k-means for $K = 7$ (left) and $K = 3$ (right).

S4.2 Mode-estimation

Model. We use the same triangle configuration with means at $(0, 1)$, $(\cos(\pi/6), -0.5)$, $(-\cos(\pi/6), -0.5)$ and isotropic covariance $\sigma^2 \boldsymbol{I}$, but with smaller covariance $\sigma^2 = 0.1$. The three means are well separated, so the density global modes approximately coincide with the means: $\boldsymbol{u}_1 \approx (0, 1)$, $\boldsymbol{u}_2 \approx (\cos(\pi/6), -0.5)$, $\boldsymbol{u}_3 \approx (-\cos(\pi/6), -0.5)$. For each experiment replicate, we sample N points.

Metrics. Each method is replicated $n_{\text{rep}} = 100$ times and provides K estimated modes $\widehat{\boldsymbol{u}} = (\widehat{\boldsymbol{u}}_1, \ldots, \widehat{\boldsymbol{u}}_K)$. We assess performance in terms of mode estimation using the following metrics:

- *Mode recovery* (MR_ϵ): number of true modes $\{\boldsymbol{u}_i, i \in [I]\}$ recovered within a strict tolerance ϵ:
$$\mathsf{MR}_\epsilon(\widehat{\boldsymbol{u}}) = \sum_{i=1}^{I} \mathbb{1}\left\{ \min_{k \in [K]} \left\| \widehat{\boldsymbol{u}}_k - \boldsymbol{u}_i \right\|_2 < \epsilon \right\}.$$
We set $\epsilon = 10^{-2}$, a small value relative to $\sigma = \sqrt{0.1}$.
- *Hungarian matching sum* (HM): minimum linear assignment cost between true and estimated modes,
$$\mathsf{HM}(\widehat{\boldsymbol{u}}) = \min_{\tau} \sum_{i=1}^{I} \left\| \widehat{\boldsymbol{u}}_{\tau(i)} - \boldsymbol{u}_i \right\|_2,$$
i.e., the minimum total distance over injective maps ($K \geq I$) $\tau : \{1, \ldots, I\} \hookrightarrow \{1, \ldots, K\}$ (one distinct estimate per true mode). This can be computed via the Hungarian algorithm (Kuhn, 1955).
- *Nearest-neighbor sum* (NN): aggregate distance from each estimate to the closest true mode,
$$\mathsf{NN}(\widehat{\boldsymbol{u}}) = \sum_{k=1}^{K} \min_{i \in [I]} \left\| \widehat{\boldsymbol{u}}_k - \boldsymbol{u}_i \right\|_2.$$

Hyperparameters and grid search. For GERVE, the initial components have equal weights $1/K$ and covariances $\sigma_1^2 \boldsymbol{I}$, $\sigma_1 > 0$. The annealing schedule is $\omega_t = \omega_1 / t^\beta$ and the step sizes are $\rho_t = \rho_1 (\omega_1 / \omega_t)^\gamma$. We used $T = 4000$ iterations and mini-batches of size $B = 1000$. The mean-shift algorithm is run with a mini-batch size $B = 1000$ and uses a bandwidth h_m for its Gaussian kernel. The feature significance method uses a bandwidth h_f to identify the points of significant curvature or gradient. For each method, we performed a grid or line search to optimize hyperparameters with respect to each metric and sample size. Table S2 gives the values used in the grid search procedure.

For Gaussian mean-shift, we use a mini-batch fixed-covariance Gaussian GERVE (equivalent updates to Gaussian mean-shift with bandwidth h) with adaptive step size $\rho_t = B \Big(\sum_{b=1}^{B} q_{\boldsymbol{\lambda}_t}(\boldsymbol{X}_{1:B}^{(t)}) \Big)^{-1}$, where $\boldsymbol{X}_{1:B}^{(t)}$ is the mini-batch sample.

Table S2: Grid of hyperparameter values used for GERVE, mean-shift, and the feature significance method in the mode-finding simulations.

Hyperparameters	Values
σ_1^2	$\{0.1, 0.2\}$
ω_1	$\{10, 50\}$
β	$\{1.1, 1.3, 1.5\}$
ρ_1	$\{0.01, 0.03, 0.1\}$
γ	$\{0.2, 0.3, 0.4\}$
h_m	$\{0.001, 0.002, 0.005, 0.01, 0.02, 0.05, 0.1, 0.2\}$
h_f	$\{0.0001, 0.0003, 0.001, 0.003, 0.01, 0.03, 0.1, 0.3\}$

Additional simulation results. Figure S4 shows performance metrics for GERVE and the two baseline methods over a sweep of $K \in \{3, \ldots, 7\}$.

S5 Clustering performance on UCI datasets

We assess robustness to misspecification on real data and compare to classic clustering methods. We also provide an ablation study and report runtimes.

S5.1 Benchmark

Datasets. The three datasets can be found in the UCI Machine Learning Repository (*archive.ics.uci.edu*).

- **Iris**: $N = 150$, $d = 4$ (sepal length/width, petal length/width), $J = 3$ species (setosa, versicolor, virginica).
- **Wine**: $N = 178$, $d = 13$ chemical attributes from $J = 3$ cultivars.
- **Pendigits**: $N \approx 11000$, $d = 16$ (eight pen-tip (x, y) pairs, normalized), $J = 10$ classes (digits 0 to 9).

Clustering metrics. We report the Adjusted Rand Index (ARI, pairwise agreement), the Adjusted Mutual Information (AMI, mutual information between estimated and true classes), Purity (per-cluster majority proportion), and Macro-F1 (classwise F1, averaged).

GERVE default setup. GERVE uses diagonal-covariance mixtures with weights fixed to k-means proportions. Means are initialized at k-means centroids, covariances at $\sigma_1^2 \boldsymbol{I}$. We include a brief fixed-covariance burn-in of L_B iterations within a total budget of T before learning covariances. We use overcompletion $K = J + 5$ to enable component merging. Other hyperparameters are listed in Table S3.

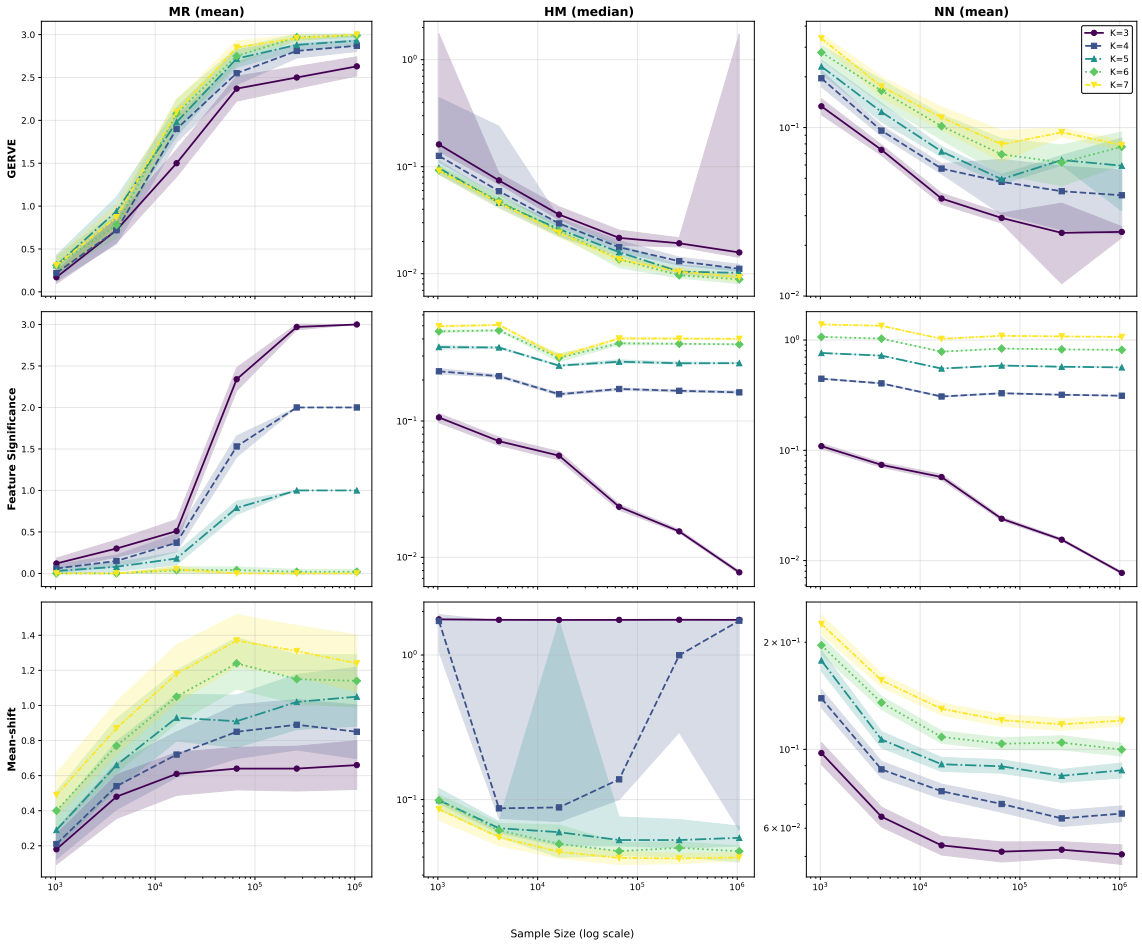

Figure S4: Mode-estimation performance of GERVE (first row), Feature Significance (second row), and Gaussian mean-shift (third row) with respect to sample size, for the triangle mixture ($I = 3$ true modes). Curves show, for varying $K \in \{3, 4, 5, 6, 7\}$, means (MR_ϵ, NN) or medians (HM) over $n_{\text{rep}} = 100$ replicates. Bands are 95% confidence intervals: means use t-intervals, medians use bootstrap percentiles (5000 samples).

Table S3: Dataset characteristics and hyperparameter values used for GERVE default setup.

	Dataset	Iris	Wine	Pendigits
Dataset characteristics	N	150	178	10992
	d	4	13	16
	J	3	3	10
GERVE default setup	T	2000	2000	2000
	K	8	8	15
	L_B	100	100	100
	σ_1^2	1	1	1
	ω_1	0.1	10^{-4}	10^{-3}
	ω_t	ω_1/t^2	ω_1/t^2	ω_1/t^2
	ρ_1	10^6	10^{12}	10^{11}
	ρ_t	ρ_1/t^2	ρ_1/t^2	ρ_1/t^2
	$\sigma^2_{\min}$	0.01	0.1	0.01
	$\sigma^2_{\max}$	1	1	1
	$D_{\boldsymbol{\mu}}$	0.1	0.1	0.2
	$D_{\boldsymbol{\Sigma}}$	1	1	1

We enforce stability by clipping parameter updates ($\|\boldsymbol{\mu}_{k,t+1} - \boldsymbol{\mu}_{k,t}\|_2 \leq D_{\boldsymbol{\mu}}$, $\|\boldsymbol{\Sigma}_{k,t+1} - \boldsymbol{\Sigma}_{k,t}\|_F \leq D_{\boldsymbol{\Sigma}}$) and constraining the parameter space via projections ($\sigma^2_{\min} I \preceq \boldsymbol{\Sigma}_k \preceq \sigma^2_{\max} I$). These constraints have little practical influence on results but improve stability in high dimensions.

Baseline methods. We compare GERVE to mean-shift (with a flat kernel), GMM-EM, and k-means. Our main label-free baselines are implemented using Python's scikit-learn package:

- **Flat mean-shift (MS-Scott):** flat kernel, bandwidth from Scott's rule (Scott, 1992). We set the maximum number of iterations to 300 (scikit-learn `MeanShift`'s default).
- **GMM-BIC:** Gaussian mixture model with diagonal covariances, the number of components selected by BIC at each run. Initial mixture parameters are set using k-means++. We stop early when the lower bound average gain is below 10^{-3} and set the maximum number of iterations to 100 (scikit-learn `GaussianMixture`'s default).
- **k-means (KM-Elbow):** number of centroids selected by the elbow method. Initial centroid locations are set using greedy k-means++. We stop early when the lower bound average gain is below 10^{-4} and set the maximum number of iterations to 300 (scikit-learn `KMeans`'s default).

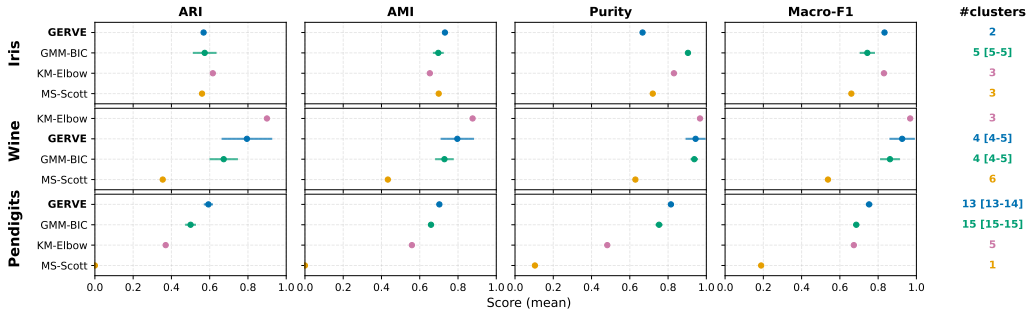

Figure S5: Clustering metrics (means ±95% CI) and number of effective clusters (median and [Q1-Q3] quantiles when applicable, or constant across all replicates) for Iris, Wine, Pendigits over $n_{\text{rep}} = 10$. Label-free methods ordered by average rank within each dataset.

For reference, we also report oracle variants that use ground truth and are not available in practice: GERVE-K ($K = J$), MS-K (bandwidth chosen by line search so that the number of clusters is closest to J), MS-ARI (bandwidth maximizing ARI), GMM-K ($K = J$), GMM-ARI (K maximizing ARI), KM-K ($K = J$), KM-ARI (K maximizing ARI).

GMM-BIC selects K by BIC at each run. The -ARI oracle variants are chosen once per dataset by maximizing mean ARI across seeds and then held fixed across runs.

Results. Figure S5 shows means and 95% confidence intervals over $n_{\text{rep}} = 10$ runs for the four metrics, and the median numbers of effective clusters and their [Q1-Q3] quantiles. GERVE is competitive with label-free baselines while adapting the number of clusters through merging. For reference, comparison against oracle baselines is provided in Figure S6.

Iris: GERVE matches baselines on ARI/AMI/Macro-F1. It merges two close species, yielding two effective clusters versus $J = 3$, with a small drop in Purity and a cleaner modal partition. *Wine:* GERVE attains strong scores across metrics. k-means edges out the top values on this near-spherical dataset, while GERVE remains robust without tuning K via BIC. *Pendigits:* No method reaches $J = 10$ in this overlapping, high-dimensional setting. GERVE produces the effective cluster count closest to J and leads on all four metrics, indicating consistent recovery of global structure even with class splits.

We emphasize ARI, AMI, and Macro-F1, since Purity can inflate with larger K. In practice, GERVE returns a smaller effective K by merging redundant components. Compared to mean-shift, it scales better in dimension and avoids bandwidth tuning through adaptive covariance learning.

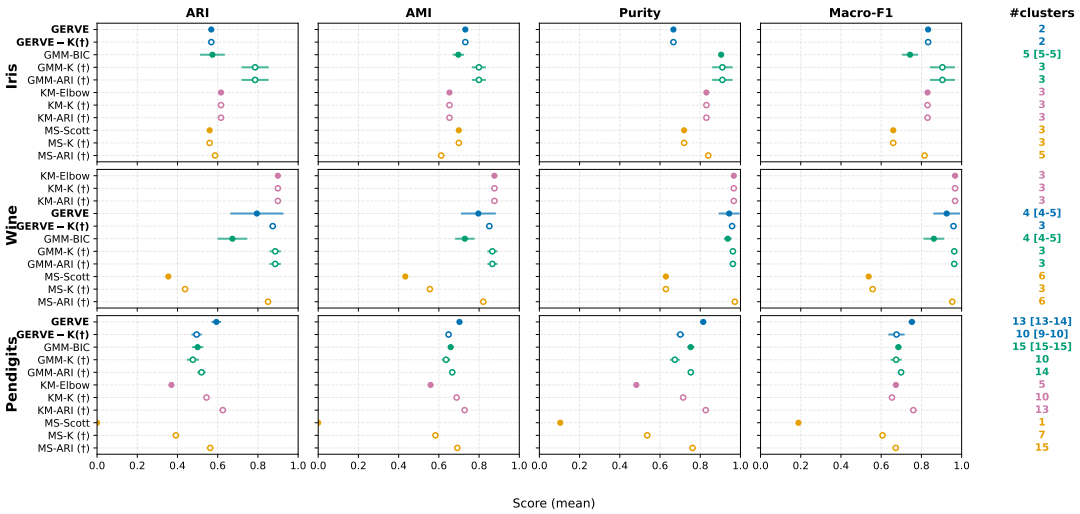

Figure S6: Clustering metrics (means ±95% CI) and effective clusters (median and [Q1-Q3] quantiles when applicable, or constant across all replicates) for Iris, Wine, Pendigits over $n_{\text{rep}} = 10$. Label-free methods (GERVE, MS-Scott, GMM-BIC, KM-Elbow) are ordered by average rank within each dataset. Oracle variants are marked with the † symbol, and directly placed below the corresponding label-free method.

S5.2 Ablation study

We now turn to an ablation study to examine how different configuration choices (weights, overcompleteness, annealing, covariance structure, burn-in) affect the behavior of GERVE.

We start with the same default GERVE dataset-specific configuration. For each dataset, we create new configurations by varying one factor at a time relative to the default: (i) Weights: equal vs. proportional to k-means group sizes; (ii) Overcompleteness: $K \in \{J, J{+}5\}$; (iii) Annealing: $\omega_t \in \{0, \omega_1/t^\beta, \beta \in \{0.5, 1, 2, 3\}\}$; (iv) Covariance: fixed, diagonal, full; (v) Burn-in: $L_B \in \{0, 50, 100, 200, 500\}$.

Figures S7-S8 report, for each configuration, the mean of each metric (with 95% confidence intervals) and the effective number of clusters (median [Q1-Q3]). We note the main takeaways for each varying factor: (i) Weights proportional to k-means help on Wine (higher ARI/AMI/Purity/Macro-F1), without significant effect elsewhere; (ii) Starting overcomplete slightly increases the effective number of clusters but helps on Pendigits, where class overlap is substantial; (iii) Faster annealing improves end-state consolidation under fixed iteration budgets, $\omega_t \equiv 0$ underperforms on overlapping classes, illustrating the importance of the exploration induced by the entropy term; (iv) Diagonal covariances are a good speed vs. robustness compromise, as full covariances add cost and variance without systematic gains; (v) A fixed-covariance short burn-in ($L_B \approx 100$) can help, but longer burn-ins steal iterations from covariance learning.

S5.3 Runtime

To illustrate how GERVE's computational cost depends on data dimension, the number of mixture components, and covariance parameterization, we report mean runtimes across four datasets (Triangle mixture, Iris, Wine, Pendigits), under a common configuration. This is not intended as a scalability benchmark, since runtime is implementation-dependent and sensitive to engineering choices.

Table S4 shows average wall-clock runtimes over ten replicates. As expected, cost increases with dimensionality and with the flexibility of the covariance model. Fixed covariances are the fastest to compute but also the least flexible, while full covariances are substantially slower, especially in higher dimensions. The diagonal setting offers a practical compromise: only slightly slower than the fixed covariance setting, yet avoiding the instability and high cost of full covariance updates. Overall, runtimes remain modest on medium-sized datasets, indicating that GERVE is practically usable in realistic clustering scenarios even when covariance learning is included.

We also plot execution time as a function of K on the Pendigits dataset. The scaling is approximately linear in K, with diagonal covariances remaining a practical compromise between fixed and full parameterizations.

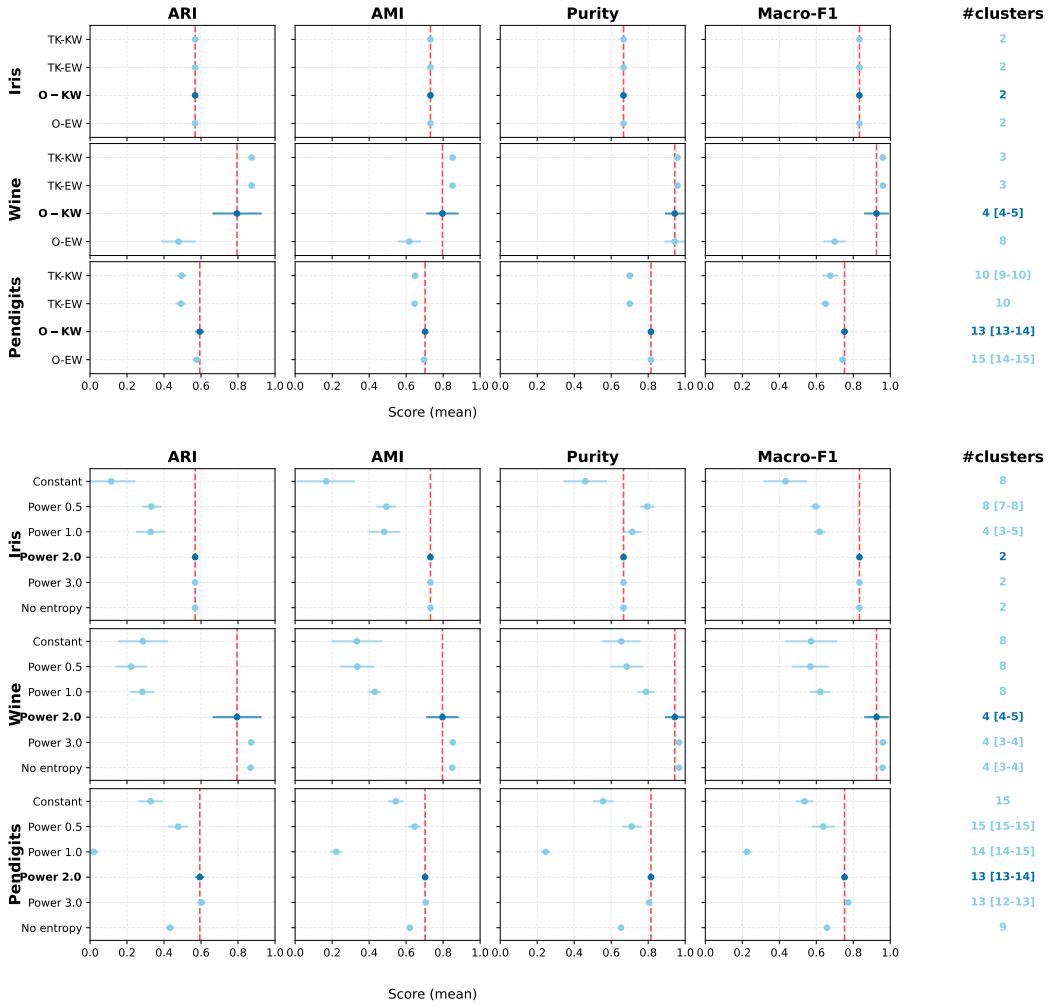

Figure S7: **Ablation – Weights and overcompleteness (top) and Annealing (bottom).** For each dataset (rows), the first four columns show ARI/AMI/Purity/Macro-F1 (means ± 95% t-CIs over $n_{\text{rep}} = 10$ runs), the rightmost column shows the effective number of clusters (median [Q1–Q3], or constant across all runs). Bold marks the default GERVE configuration. The dashed reference line in each metric panel marks the default's mean for that metric. Dataset-specific J values are listed in Table S3.
Top. Prefixes: **T-** for $K = J$ (true groups), **O-** for $K = J{+}5$ (overcomplete). Suffixes: **-KW** for k-means-proportional weights, **-EW** for equal weights.
Bottom. Schedules: *Constant* ($\omega_t \equiv \omega_1$), *Power-x* ($\omega_t = \omega_1 t^{-x}$), *No entropy* ($\omega_t \equiv 0$).

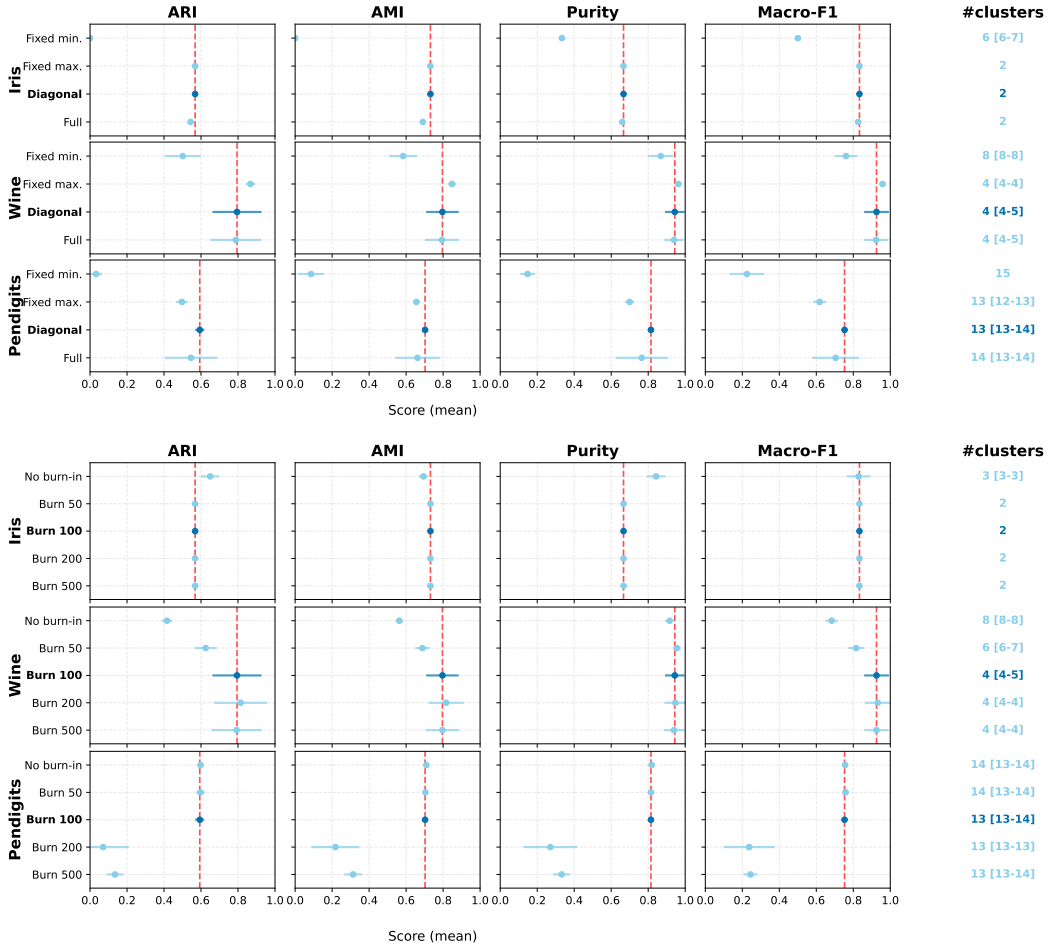

Figure S8: **Ablation – Covariance (top) and Burn-in phase (bottom).** For each dataset (rows), the first four columns show ARI/AMI/Purity/Macro-F1 (means $\pm$ 95% t-CIs over $n_{\text{rep}} = 10$ runs), the rightmost column shows the effective number of clusters (median [Q1–Q3], or constant across all runs). Bold marks the default GERVE configuration. The dashed reference line in each metric panel marks the default's mean for that metric.
Top. Covariance parameterizations: *Fixed min.* ($\boldsymbol{\Sigma}_k \equiv \sigma^2_{\min}\boldsymbol{I}$), *Fixed max.* ($\boldsymbol{\Sigma}_k \equiv \sigma^2_{\max}\boldsymbol{I}$), *Diagonal* (learned diagonal $\boldsymbol{\Sigma}_k$ with bounds $\sigma^2_{\min}\boldsymbol{I} \preceq \boldsymbol{\Sigma}_k \preceq \sigma^2_{\max}\boldsymbol{I}$), and *Full* (learned full $\boldsymbol{\Sigma}_k$ under the same bounds). Dataset-specific values for $\sigma^2_{\min}, \sigma^2_{\max}$ and the initialization σ^2_0 (for learned covariances) are listed in Table S3.
Bottom. Burn-in variants: *No burn-in* ($L = 0$), *Burn-x* (fixed-covariance phase with $L = x$ iterations). The total iteration budget T is the same across variants and listed in Table S3.

Table S4: Mean wall-clock runtime (in seconds) over 10 replicates for overcomplete GERVE ($K = J + 5$) with weight updates, $T = 500$, $B = 150$, $L = 0$, using identical hyperparameters across all datasets. Values are mean $\pm$ standard deviation. All runs on the same machine. This study illustrates cost trends by covariance parameterization, number of components, and dimensionality.

Dataset	Fixed	Diagonal	Full
Triangle ($d = 2, J = 3$)	3.5 ± 0.6	3.9 ± 0.8	4.2 ± 0.8
Iris ($d = 4, J = 3$)	3.5 ± 0.7	3.8 ± 0.7	4.6 ± 0.9
Wine ($d = 13, J = 3$)	4.2 ± 0.8	4.8 ± 0.9	7.8 ± 1.2
Pendigits ($d = 16, J = 10$)	13.4 ± 2.1	15.2 ± 2.6	24.6 ± 3.3

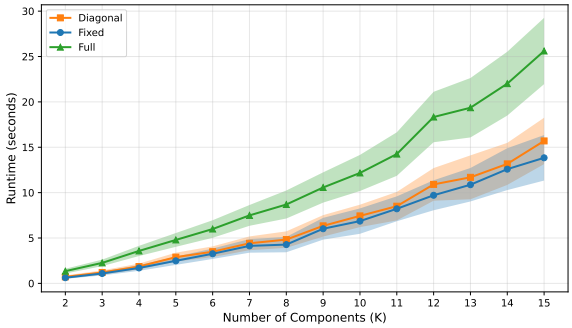

Figure S9: GERVE runtime with respect to the number of components K on the Pendigits dataset, for diagonal, fixed and full covariance matrices. Curves show mean runtime over 10 replicates, shaded bands indicate ± 1 standard deviation.

S6 Details on the UK collision case study

Data normalization and processing. We apply GERVE on a normalized space : the Greater London window $[-0.54, 0.33] \times [51.28, 51.70]$ (longitude-latitude) into a centered rectangle $[-0.7, 0.7] \times [-0.35, 0.35]$ preserving the aspect ratio and with area ≈ 1. This transformation is applied to the coordinates of the data points. To smoothen the underlying distribution p, we generate artificial data points uniformly in the region between the normalized rectangle and the $[-2, 2] \times [-2, 2]$ square, with density 1000 times smaller than the mean density of the data in the normalized rectangle.

GERVE hyperparameters. We target the top $H = 10$ hotspots. We fit an overcomplete mixture with $K = 2H = 20$.

- *Temperature.* We fix the stopping temperature at $\omega^\dagger = 1$. This choice follows the elbow of the resolved-mode curve represented in Figure S10 and described in Sec. 7.

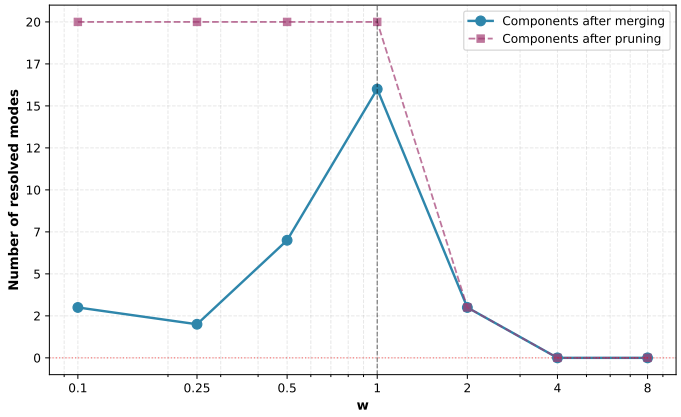

Figure S10: Diagnostic for choosing ω: resolved modes after successive pruning (dashed) and merging (solid) for values of ω (in log-scale). The elbow at $\omega^\dagger = 1$ maximizes resolved modes before decline.

The same $\omega^\dagger$ is used for the bootstrap refits.

- *Step sizes.* The step-size schedule is $\rho_t = 0.2t^{-0.1}$.
- *Initialization.* Component means use k-means++ centroids on the working coordinates. Initial covariances are $\mathbf{\Sigma}_{k,1} = \sigma_1^2 \boldsymbol{I}$ with $\sigma_1^2 = 5 \times 10^{-3}$.
- *Covariance bounds.* We constrain $\sigma^2_{\min} = 1 \times 10^{-5}$ and $\sigma^2_{\max} = 1 \times 10^{-2}$.
- *Domain.* Optimization is performed on $\mathcal{S} = [-2.05, 2.05] \times [-2.05, 2.05]$. Mapping to British National Grid is used only for reporting distances and ellipses.
- *Entropy term.* We use $B = 100$ Monte Carlo samples per iteration to estimate the entropy contributions.
- *Early stopping.* We cap iterations at $T = 10{,}000$. We stop early if the following hold at three consecutive checks, evaluated every 10 iterations:

$$\left\|\boldsymbol{\mu}_{k,t+1} - \boldsymbol{\mu}_{k,t}\right\|_2 < 10^{-2} \quad \text{and} \quad \frac{\left\|\boldsymbol{S}_{k,t+1} - \boldsymbol{S}_{k,t}\right\|_F}{\left\|\boldsymbol{S}_{k,t}\right\|_F} < 10^{-1} \quad \text{for all } k.$$

- *Pruning and merging.* A component is pruned if

$$\mathrm{eig}_{\max}(\mathbf{\Sigma}_k) > \sigma^2_{\min} + \beta\big(\sigma_1^2 - \sigma^2_{\min}\big),$$

with $\beta \approx 0.018$, so that the threshold equals $10\sigma^2_{\min}$. Components means within 0.005 (≈ 200 meters in the British National Grid EPSG:27700) are merged.

- *Bootstrap settings.* We use $L = 500$ resamples at $\omega^\dagger$. Bootstrap modes are matched to the baseline with a Hungarian assignment and adaptive gates around each baseline mode as described below.

Mode matching and per-mode ellipses. We z-score coordinates using the baseline mode centers. For mode k, we compute nn_k as the minimum distance to any other baseline mode. The gate is $\tau_k = \mathrm{clip}(\eta \mathrm{nn}_k, \tau_{\min}, \tau_{\max})$ with defaults $\eta = 0.35$, $\tau_{\min} = 0.08$, $\tau_{\max} = 0.22$. Since $\eta < 0.5$, gates stay inside midpoints between modes, which prevents cross-assignments. The clips avoid degenerate gates when modes are extremely close or very isolated. After Hungarian assignment, pairs with distance larger than τ_k are rejected. Accepted matches form the bootstrap cloud for mode k, from which we compute stability s_k and the 95% chi-square confidence ellipse.

To obtain confidence regions in meters, we project coordinates to the British National Grid (EPSG:27700). For each resolved mode k, let $\{\widehat{\boldsymbol{\mu}}^{(\ell)}\}_{\ell=1}^{L}$ denote the matched bootstrap centers. We report 95% confidence ellipses from the empirical covariance $\widehat{V}_k = \mathrm{Var}^*(\widehat{\boldsymbol{\mu}}_k^{(\ell)})$: if $\mathrm{eig}_1 \geq \mathrm{eig}_2$ are eigenvalues of $\widehat{V}_k$ and $\boldsymbol{e}_1$ is the principal eigenvector, then the semi-axes are $a = \sqrt{\chi^2_{2;1-\alpha}\mathrm{eig}_1}$ and $b = \sqrt{\chi^2_{2;1-\alpha}\mathrm{eig}_2}$, with angle $\mathrm{atan2}(\boldsymbol{e}_{1,x}, \boldsymbol{e}_{1,y})$ (in degrees) relative to Easting.

Figure S11 shows, for each baseline mode, the cloud of matched bootstrap centers, the adaptive gate used for matching, and the resulting confidence ellipse, illustrating mode stability and localization uncertainty.

2020-2024 hotspots. Table S5 reports the 17 hotspots found by GERVE and Figure S12 displays them in the complete search space.

Mean-shift details. We use scikit-learn's `MeanShift` with a flat (uniform) kernel on the same projected coordinates. The reference bandwidth h_0 is estimated as the 0.3-quantile of pairwise distances. We then sweep h on a grid from $0.02h_0$ to $0.2h_0$ and report the resulting centers. Figure S13 shows the centers across the sweep. Smaller h recovers many additional peripheral modes, while larger h merges central structure. In Figure S14, we overlay the $0.05h_0$ and $0.1h_0$ centers in the zoomed-in window and compare them with GERVE's centers and 95% ellipses. At $0.1h_0$, visual inspection against the road network some mean-shift centers fall slightly off carriageways, so we select a smaller bandwidth ($0.05h_0$) as a baseline comparator for GERVE.

2014-2019 hotspots and comparison to 2020-2024. We fit GERVE to collision data from 2014 to 2019 with identical $\omega^\dagger$ and K. Confidence ellipses are obtained after $L = 500$ bootstrap fits. Table S6 reports the 16 modes found by GERVE and Figure S15 displays them with their 95% confidence intervals. Figure S16 overlays the 2014-2019 hotspots on a map of the 2020-2024 ones in the zoomed-in window.

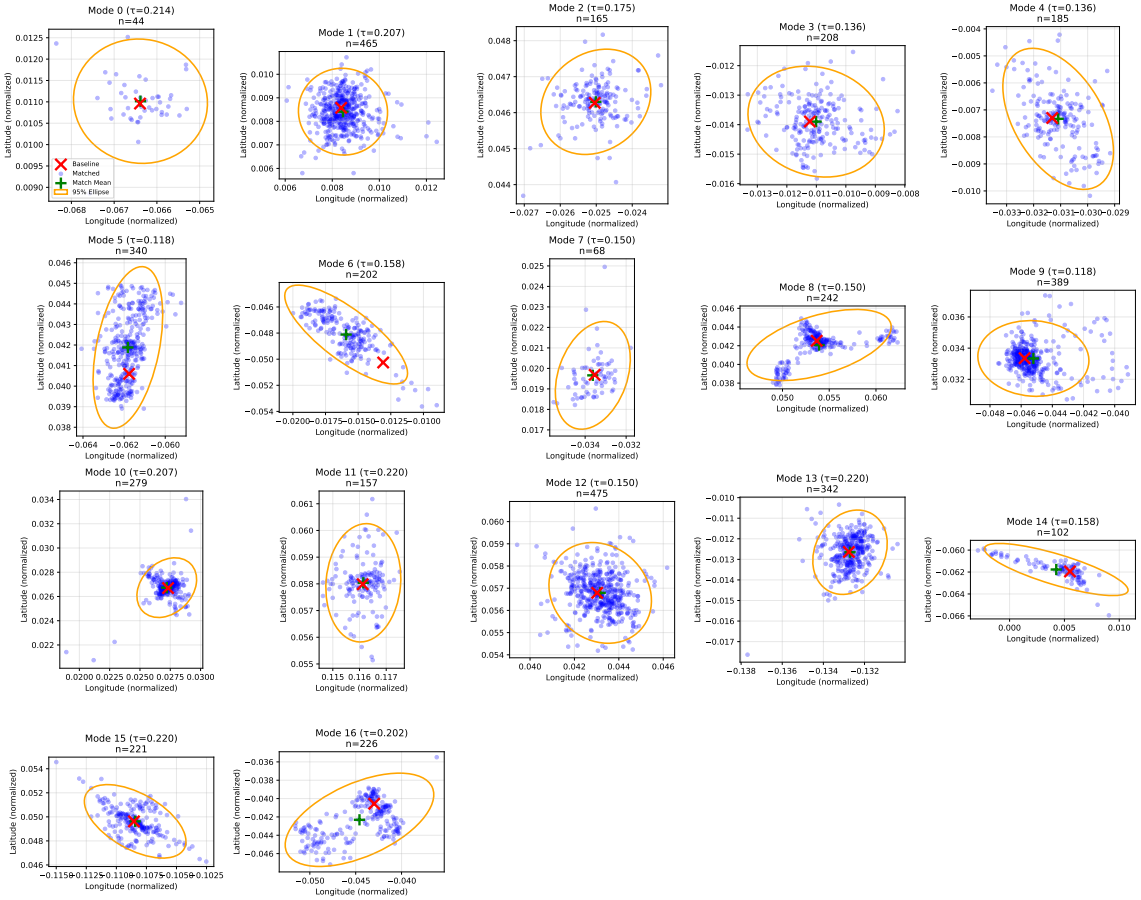

Figure S11: Distribution of the matched modes across the $L = 500$ bootstrap fits for each baseline hotspot from the 2020-2024 data, with the estimated confidence ellipses and means.

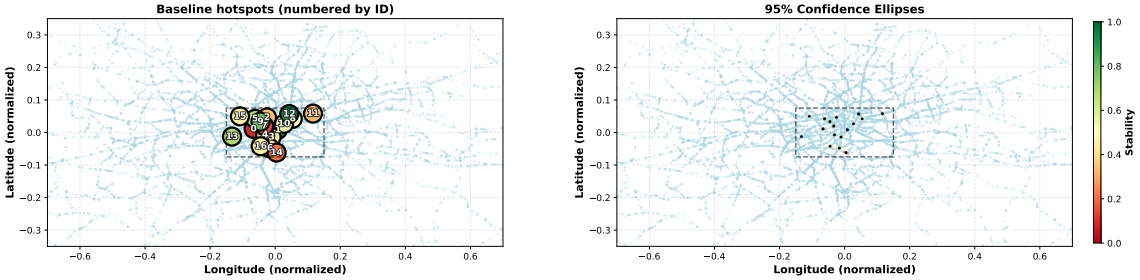

Figure S12: Baseline collision hotspots identified by GERVE for the Greater London Area between 2020 and 2024, with stability scores and 95% confidence ellipses, in normalized coordinates. The dashed rectangle is the zoomed-in window of Figure 3.

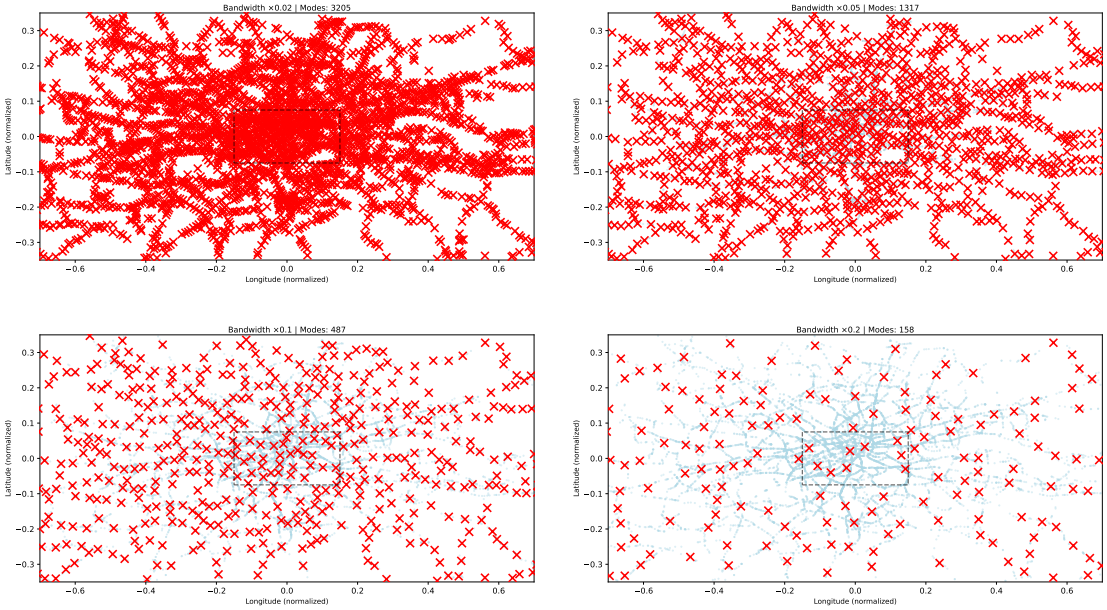

Figure S13: Mean-shift modes across different bandwidths ($h = 0.02h_0, 0.05h_0, 0.1h_0, 0.2h_0$) in the normalized window. The dashed rectangle is the zoomed-in window.

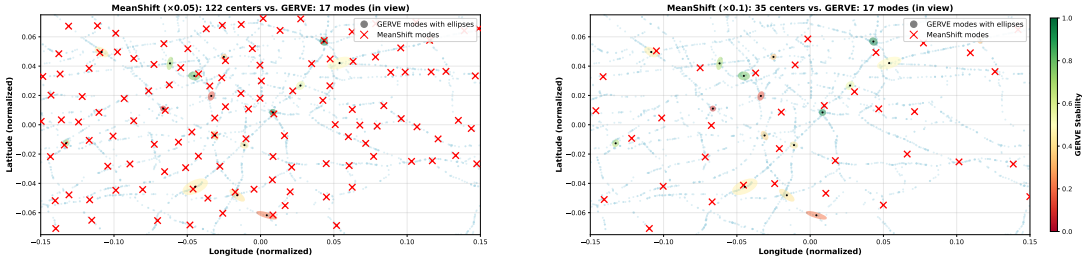

Figure S14: Comparison of mean-shift ($h = 0.05h_0, 0.1h_0$) and GERVE modes, with confidence ellipses for GERVE, in the zoomed-in window.

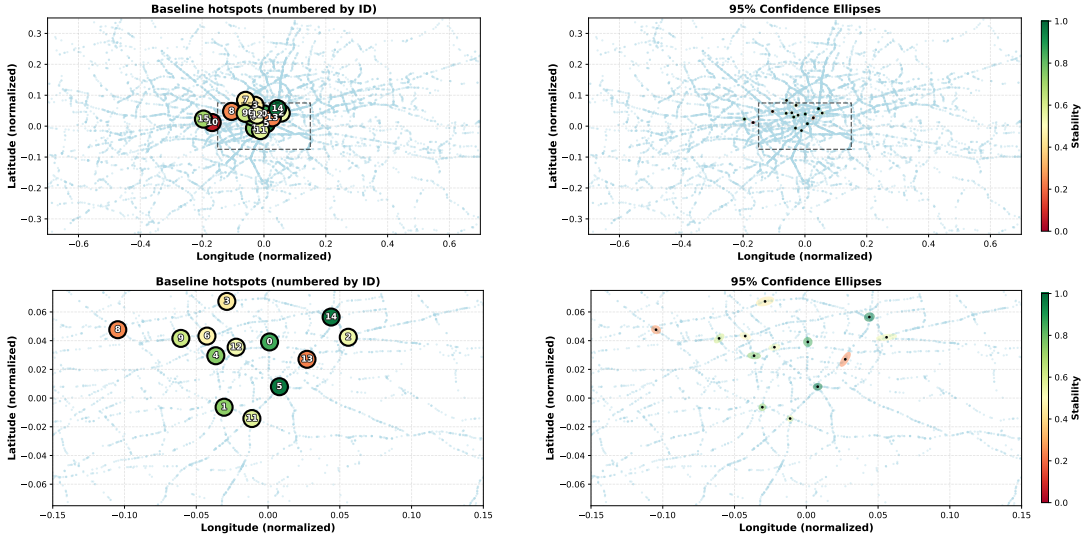

Figure S15: Baseline collision hotspots identified by GERVE for the Greater London Area between 2014 and 2019, with stability scores and 95% confidence ellipses, in normalized coordinates. Top: Shows the normalized window containing all data points. The dashed rectangle is the zoomed-in window. Bottom: Zoomed-in version for comparison to Figure 3.

Table S5: Reported hotspots for year 2024 with (un-normalized) coordinates, confidence ellipse dimensions (a, b), angles, stability, bootstrap matches, and local counts (radius 200 m). Hotspots are listed in decreasing stability order.

ID	Name	Longitude (°)	Latitude (°)	a (m)	b (m)	Angle (°)	Stability	Count
12	Shoreditch	-0.078909	51.524226	134	121	-54	0.95	32
1	Elephant & Circus	-0.099874	51.495003	120	86	80	0.93	27
9	Piccadilly Circus	-0.132294	51.510060	162	150	89	0.78	27
5	Oxford Circus	-0.142322	51.515222	265	66	84	0.68	31
13	Gunter Grove	-0.185143	51.482279	134	83	72	0.68	22
10	London Bridge	-0.088544	51.506068	156	117	60	0.56	21
8	Aldgate	-0.072412	51.515338	357	207	27	0.48	14
16	Clapham HS	-0.131914	51.464362	409	255	37	0.45	23
15	Edgware Road	-0.170483	51.519900	230	144	-41	0.44	22
3	Oval	-0.111630	51.481528	120	105	81	0.42	19
6	Brixton	-0.114597	51.460859	301	115	-49	0.40	38
4	Vauxhall	-0.123762	51.485494	177	82	-68	0.37	27
2	Holborn	-0.120090	51.517891	87	78	54	0.33	25
11	Mile End	-0.034823	51.524982	149	59	88	0.31	23
14	Herne Hill	-0.102411	51.452610	305	91	-24	0.20	10
7	Westminster	-0.125293	51.501799	177	76	79	0.14	17
0	Victoria	-0.145081	51.496576	96	68	84	0.09	20

Table S6: Reported hotspots for year 2019 with (un-normalized) coordinates, confidence ellipse dimensions (a, b), angles, stability, bootstrap matches, and local counts (radius 200 m). Hotspots are listed in decreasing stability order.

ID	Name	Longitude (°)	Latitude (°)	a (m)	b (m)	Angle (°)	Stability	Count
14	Shoreditch	-0.078437	51.524026	159	134	-78	0.96	35
5	Elephant & Circus	-0.100192	51.494709	123	112	-86	0.96	27
0	Ludgate Circus	-0.104340	51.513547	177	110	-80	0.85	20
4	Charing Cross	-0.126966	51.507739	168	114	-4	0.75	24
1	Vauxhall	-0.123471	51.486071	111	102	-66	0.74	23
15	Sheperd's Bush	-0.223191	51.503715	179	133	-3	0.70	25
9	Oxford Circus	-0.141671	51.515031	197	96	56	0.62	26
11	Oval	-0.111782	51.481310	98	72	-79	0.57	23
2	Aldgate	-0.071210	51.515511	279	100	14	0.54	15
12	Aldwych	-0.118401	51.511326	145	108	15	0.53	23
6	St Giles Circus	-0.130695	51.516011	151	109	-29	0.47	15
3	King's Cross	-0.122322	51.530607	252	155	24	0.44	22
7	Camden Town	-0.140953	51.540307	197	90	59	0.44	19
8	Edgware	-0.168222	51.518691	138	137	73	0.25	22
13	London Bridge	-0.088592	51.506237	318	85	70	0.23	14
10	Olympia	-0.206178	51.496948	156	72	7	0.07	14

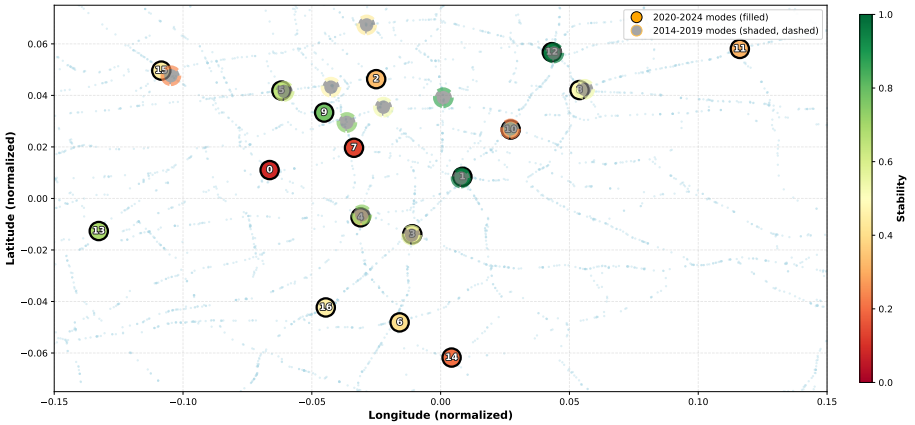

Figure S16: Comparison of 2014-2019 hotspots and 2020-2024 hotspots in the zoomed-in window.