Symmetry Guarantees Statistic Recovery in Variational Inference

Daniel Marks
Department of Computer Science
University of Oxford
daniel.marks@cs.ox.ac.uk

Dario Paccagnan
Department of Computing
Imperial College London
d.paccagnan@imperial.ac.uk

Mark van der Wilk
Department of Computer Science
University of Oxford
mark.vdwilk@cs.ox.ac.uk

Abstract

Variational inference (VI) is a central tool in modern machine learning, used to approximate an intractable target density by optimising over a tractable family of distributions. As the variational family cannot typically represent the target exactly, guarantees on the quality of the resulting approximation are crucial for understanding which of its properties VI can faithfully capture. Recent work has identified instances in which symmetries of the target and the variational family enable the recovery of certain statistics, even under model misspecification. However, these guarantees are inherently problem-specific and offer little insight into the fundamental mechanism by which symmetry forces statistic recovery. In this paper, we overcome this limitation by developing a general theory of symmetry-induced statistic recovery in variational inference. First, we characterise when variational minimisers inherit the symmetries of the target and establish conditions under which these pin down identifiable statistics. Second, we unify existing results by showing that previously known statistic recovery guarantees in location–scale families arise as special cases of our theory. Third, we apply our framework to distributions on the sphere to obtain novel guarantees for directional statistics in von Mises–Fisher families. Together, these results provide a modular blueprint for deriving new recovery guarantees for VI in a broad range of symmetry settings.

1 Introduction

One of the central problems of modern statistics and machine learning is the accurate estimation of intractable probability distributions. It most prominently arises in the context of Bayesian statistics, where the intractability of the marginal likelihood renders exact computation of the posterior distribution infeasible in all but the simplest statistical models. This has led to the development of a range of approximate inference methods to approximate or sample from such intractable distributions.

Variational inference (VI) has emerged as a popular approximate inference technique (Jordan et al., 1999; Wainwright and Jordan, 2008; Blei et al., 2017), balancing the scalability of simpler methods such as Laplace approximations (MacKay, 1992) with the fidelity of more computationally intensive approaches such as MCMC (Robert and Casella, 2004). At its core, VI recasts approximate inference as an optimisation problem: given an intractable target distribution, it seeks the best approximation to it within a tractable variational family by minimising a measure of dissimilarity between the two. The most commonly employed objective is the reverse Kullback–Leibler (KL) divergence, for which

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

a readily computable optimisation surrogate is provided by the evidence lower bound. However, the principles of VI are agnostic to the choice of divergence, although different objectives need not be equally amenable to optimisation (Geffner and Domke, 2021a,b). Alternative objectives that have been studied include the forward KL divergence (Minka, 2001; Naesseth et al., 2020), α-divergences (Hernandez-Lobato et al., 2016; Dieng et al., 2017; Daudel et al., 2023), α–β divergences (Regli and Silva, 2018), and Rényi divergences (Li and Turner, 2016).

A crucial drawback of variational methods lies in the fact that the variational family may not contain the target distribution, in which case the resulting approximation can be arbitrarily bad. This occurs, for example, when the variational family cannot match the target's number of modes, or is restricted to factorised distributions, as in mean-field variational inference (Peterson and Anderson, 1987; Hinton and Van Camp, 1993), despite the target exhibiting strong dependencies. As a result, guarantees on the quality of VI's estimates are crucial for assessing the reliability of the resulting approximations and understanding when and why they may be relied upon in practice.

In this work, we derive formal guarantees for the statistic recovery of variational approximations, even under severe model misspecification, in the presence of symmetries of the target distribution. Our goal is to develop a general framework for symmetry propagation in variational inference, together with a precise correspondence between symmetries and recoverable statistics, that allows us to answer the following open problem first articulated by Margossian and Saul (2025):

For a given symmetry of the target distribution,
which statistics does variational inference recover exactly?

Main contributions

1. We prove a general statistic recovery theorem showing that if the target is invariant under a group of symmetries that the variational family respects, a unique variational minimiser inherits this invariance and thus recovers all symmetry-determined statistics.
2. We show that the recovery results of Margossian and Saul (2025) in location–scale families under even and elliptical symmetry arise as special cases of our general framework.
3. We apply our theory to directional statistics on the sphere and derive novel guarantees for recovering the axis of symmetry in von Mises–Fisher families under rotational symmetry.

Our results are formulated at the level of f-divergences, thereby encompassing many of the aforementioned variational objectives, and are agnostic to the particular VI scheme used to optimise them (Hoffman et al., 2013; Ranganath et al., 2014). Conceptually, our theory isolates a general mechanism by which symmetry can propagate through variational inference and provides a modular framework for deriving new statistic recovery guarantees in symmetry settings beyond those considered here.

Related work Much work has focused on asymptotic guarantees for variational approximations. Wang and Blei (2019) proved a variational Bernstein–von Mises theorem, establishing frequentist consistency of variational estimators. Pati et al. (2018) and Yang et al. (2020) proved convergence rates for point estimates constructed from variational approximations with different objectives, while Zhang and Gao (2020) proved convergence rates for entire posteriors in nonparametric settings. Alquier and Ridgway (2020) similarly derived such rates for tempered posteriors.

A parallel line of work studies non-asymptotic guarantees for variational methods. Han and Yang (2019) derived explicit bounds on the KL divergence between a mean-field variational posterior and a normal distribution centred at the maximum likelihood estimator, and showed that the variational mean matches the MLE up to higher-order terms. Motivated by empirical evidence of VI's ability to recover the mean (MacKay, 2003; Giordano et al., 2018), subsequent work focused on the quality of variational estimates of the target's statistics. In particular, Katsevich and Rigollet (2024) derived non-asymptotic bounds on the mean and covariance error in Gaussian VI.

Most closely related to our work is that of Margossian and Saul (2025), who showed that, provided the variational minimiser is unique, reverse-KL divergence VI in location–scale families exactly recovers the mean when the target and variational family exhibit even symmetry, and additionally recovers the covariance up to a multiplicative constant when they exhibit elliptical symmetry.[1] Our work

[1] Margossian et al. (2026) extend these guarantees to f-divergence objectives and also establish conditions under which recovery is guaranteed at every stationary point.

generalises these results by moving beyond recovery guarantees tied to specific symmetry classes, distributional families, and statistics, and instead developing a unified symmetry-based framework for statistic recovery.

2 A symmetry-based framework for statistic recovery

Let $(\mathcal{X}, \mathcal{B})$ be a measurable space, and let $\mathcal{P}$ denote the set of associated probability measures. VI starts with a target distribution $P \in \mathcal{P}$, and seeks the best approximation to it inside a tractable variational family $\mathcal{Q} \subseteq \mathcal{P}$ by minimising a statistical divergence $D\colon \mathcal{P} \times \mathcal{P} \to [0, \infty]$:

$$\mathcal{Q}^\star \coloneqq \operatorname*{argmin}_{Q \in \mathcal{Q}} D(P \| Q). \tag{$\star$}$$

Throughout this paper, we consider f-divergences D_f, a broad class of discrepancy measures that encompasses many of the objectives used in variational inference, including the usual reverse KL divergence.[2] In most practical applications, $P \notin \mathcal{Q}$, meaning that the variational family is not rich enough to contain the true target and the variational problem is misspecified. For example, P may be a complicated multimodal distribution and $\mathcal{Q}$ a simple family of Gaussian distributions. Our goal in this section is therefore not necessarily to recover the full target distribution. Instead, we wish to characterise when, despite misspecification, VI can faithfully recover different statistics of the target.

We model a statistic as a partial map $S\colon \mathcal{P} \rightharpoonup \mathcal{Y}$, where $\mathcal{Y}$ is an output space. This reflects the fact that many statistics are only well-defined on a natural subclass of $\mathcal{P}$, which we call the *domain* of S and denote by $\mathrm{dom}(S) \subseteq \mathcal{P}$. For instance, for $\mathcal{X} = \mathbb{R}^d$, the mean is a statistic mapping distributions with finite first moment to vectors in $\mathbb{R}^d$, while the covariance is a statistic mapping distributions with finite second moment to positive-semidefinite $d \times d$ matrices. Since recovery statements are only meaningful when the statistic is well-defined, we assume that $P \in \mathrm{dom}(S)$ and $\mathcal{Q} \subseteq \mathrm{dom}(S)$.

We are particularly interested in statistic recovery guarantees that arise from symmetries, such as coordinate permutations, rotations, or sign flips. Such symmetries emerge from transformations of the sample space itself, and in our setting it is therefore natural to represent them by a group $\mathcal{G}$ of measurable bijections $g\colon \mathcal{X} \to \mathcal{X}$ with measurable inverses — a measure-theoretic generalisation of a change of coordinates — so that they may be composed and inverted.[3] Each $g \in \mathcal{G}$ acts on any probability measure $\pi \in \mathcal{P}$ via the pushforward map

$$g_\# \pi(A) \coloneqq \pi(g^{-1}(A)), \qquad A \in \mathcal{B}.$$

Equivalently, if $X \sim \pi$, then $g(X) \sim g_\# \pi$. We say that $\pi \in \mathcal{P}$ is *$\mathcal{G}$-invariant* if $g_\# \pi = \pi$ for all $g \in \mathcal{G}$, that is, if applying any symmetry in the group leaves the distribution unchanged. For example, a centred isotropic Gaussian in $\mathbb{R}^d$ is invariant under the action of the orthogonal group $\mathrm{O}(d)$ of real $d \times d$ orthogonal matrices, since it is left unchanged by rotations and reflections about the origin. Similarly, we say that a collection $\mathcal{F} \subseteq \mathcal{P}$ is *$\mathcal{G}$-closed* if for all $g \in \mathcal{G}$, we have $g_\# \mathcal{F} \subseteq \mathcal{F}$, meaning that $\mathcal{F}$ is closed under pushforward by elements of $\mathcal{G}$. For example, the Gaussian family in $\mathbb{R}^d$ is closed under the action of the general linear group $\mathrm{GL}(d, \mathbb{R})$ of real $d \times d$ invertible matrices, since every linear transformation of a Gaussian random variable is also a Gaussian random variable.

It turns out that symmetries of the target distribution can constrain the values that its statistics may take, and these constraints can be inherited by a variational minimiser, thus guaranteeing statistic recovery even when $P \notin \mathcal{Q}$. To make this possible, we impose three structural conditions describing how the target, the variational family, and the statistic must interact with $\mathcal{G}$:

Conditions

(1) The target $P \in \mathrm{dom}(S)$ is $\mathcal{G}$-invariant.

(2) The variational family $\mathcal{Q} \subseteq \mathrm{dom}(S)$ is $\mathcal{G}$-closed.

(3) The statistic $S\colon \mathcal{P} \rightharpoonup \mathcal{Y}$ has $\mathcal{G}$-closed domain, and for all $g \in \mathcal{G}$ and $\pi_1, \pi_2 \in \mathrm{dom}(S)$,

$$S(\pi_1) = S(\pi_2) \quad \Longrightarrow \quad S(g_\# \pi_1) = S(g_\# \pi_2).$$

[2] The definition and further background are deferred to Appendix A.1.

[3] Formally, $\mathcal{G}$ is a subgroup of the group of measurable automorphisms $\mathrm{Aut}(\mathcal{X}, \mathcal{B})$.

The final condition requires that if two distributions have the same statistic, then their pushforwards by any transformation in $\mathcal{G}$ also have the same statistic. This allows us to study the effect of symmetries of a distribution directly at the level of its statistic values. To this end, write $\mathrm{Im}(S) \coloneqq S(\mathrm{dom}(S))$ for the set of attainable statistic values and fix $y \in \mathrm{Im}(S)$. By definition, there exists at least one distribution $\pi \in \mathrm{dom}(S)$ such that $S(\pi) = y$. Now, for any $g \in \mathcal{G}$, consider the quantity $S(g_{\#}\pi)$, which is well-defined since $\mathrm{dom}(S)$ is assumed to be $\mathcal{G}$-closed. At first glance, $S(g_{\#}\pi)$ may appear to depend on the particular choice of π that generated y, since multiple distributions can share the same statistic. However, Condition (3) ensures precisely that if π_1 and π_2 both satisfy $S(\pi_1) = S(\pi_2) = y$ then $S(g_{\#}\pi_1) = S(g_{\#}\pi_2)$. We may thus introduce, for every $g \in \mathcal{G}$, a well-defined map $\rho_g : \mathrm{Im}(S) \to \mathrm{Im}(S)$ given by

$$\rho_g(y) \coloneqq S(g_{\#}\pi), \quad \text{for any } \pi \text{ such that } S(\pi) = y.$$

Since our ultimate goal is statistic recovery, we consider the statistic values fixed by every symmetry in $\mathcal{G}$. We collect these in the common fixed-point set of the family $(\rho_g)_{g\in\mathcal{G}}$ acting on $\mathrm{Im}(S)$:

$$\Gamma_S(\mathcal{G}) \coloneqq \{y \in \mathrm{Im}(S) \mid \rho_g(y) = y,\ \forall g \in \mathcal{G}\}.$$

We are now ready to state our main statistic recovery result. At a high level, f-divergences are invariant under simultaneous pushforward, meaning that $D_f(g_{\#}P \| g_{\#}Q) = D_f(P\|Q)$ for every $g \in \mathcal{G}$, as shown in Lemma A.1.[4] Together with Conditions (1) and (2), this ensures that pushing forward a variational minimiser by any element of $\mathcal{G}$ yields another minimiser. If the minimiser is unique, it must therefore inherit the $\mathcal{G}$-invariance of the target. Condition (3) then lifts this symmetry from distributions to statistic values, yielding the desired recovery guarantee.

Theorem 2.1 (Statistic recovery) *Let $S\colon \mathcal{P} \rightharpoonup \mathcal{Y}$ be a statistic, $P \in \mathrm{dom}(S)$ a target distribution, and $\mathcal{Q} \subseteq \mathrm{dom}(S)$ a variational family. Assume the variational problem ($\star$) has a unique global minimiser, denoted by $Q^\star$. If Conditions (1)–(3) are satisfied for some group of symmetries $\mathcal{G}$, then*

$$S(P), S(Q^\star) \in \Gamma_S(\mathcal{G}).$$

A proof can be found in Appendix A.2.[5] Theorem 2.1 shows that even when $P \notin \mathcal{Q}$, variational inference in a $\mathcal{G}$-closed family against a $\mathcal{G}$-invariant target forces every symmetry-compatible statistic of the target and a unique variational minimiser to lie in the corresponding fixed set. When the fixed set is a singleton, exact recovery of the statistic is achieved. In general, however, the fixed set may instead be a low-dimensional subset of the codomain, in which case the recovery is only partial. We will see examples of both cases in Section 3. The unique minimiser assumption is realised in a range of variational problems, including those considered below, with sufficient conditions often following from convexity properties of variational objectives, particularly in location and location–scale families; see Domke (2020, Theorem 9) and Margossian and Saul (2025, Lemma 13). This assumption also underpins several existing symmetry-based guarantees for statistic recovery (Margossian and Saul, 2025; Zellinger and Vergari, 2026).

Remark 2.2 *In practice, variational inference is carried out over a parameterised variational family $\mathcal{Q} = \{\Pi(\theta) \mid \theta \in \Theta\}$, where Θ is a set of parameters and $\Pi\colon \Theta \to \mathcal{P}$ is a parameterisation map, by minimising $D_f(P\|\Pi(\theta))$ over Θ. While uniqueness of the minimising distribution need not imply uniqueness of the minimising parameter due to non-identifiability of the parameterisation, the converse does hold. As parameter-level uniqueness is often easier to verify in concrete models, this is the form in which we will use Theorem 2.1 in later sections.*

3 Location–scale families: guarantees under even and elliptical symmetry

In this section, we show that our framework generalises existing statistic recovery guarantees for VI in location–scale families. In particular, we demonstrate that the recovery results of Margossian and Saul (2025) under even and elliptical symmetry arise as direct corollaries of Theorem 2.1, obtained by verifying conditions (1)–(3) and computing the corresponding fixed set in each setting.[6]

[4] Indeed, this invariance is the only property of the objective used in the symmetry argument. Although f-divergences satisfy it under every measurable automorphism, invariance under $\mathcal{G}$ suffices. Under the same minimiser assumptions, the guarantees therefore extend to Rényi divergences, also invariant under every measurable automorphism, and to integral probability metrics whose test-function classes are preserved under composition with elements of $\mathcal{G}$.

[5] In fact, we prove a more general result that replaces uniqueness of the variational minimiser with the weaker assumptions that $\mathcal{Q}^\star \neq \varnothing$ and $S(\mathcal{Q}^\star)$ is a singleton, meaning that all global minimisers agree on the statistic S. Theorem 2.1 follows as a special case, since these assumptions hold whenever the global minimiser is unique.

[6] See Appendix B.1 for a comparison between our presentation and that of Margossian and Saul (2025).

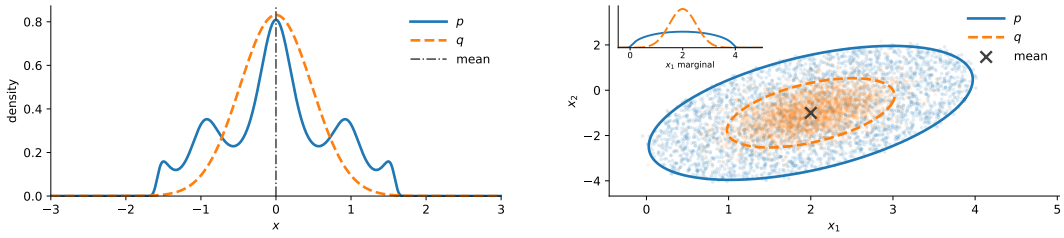


Figure 1: Unique variational minimisers q from a Gaussian family exactly recover symmetry-determined statistics of highly non-Gaussian targets p. Left: under even symmetry, the mean is recovered exactly as per Corollary 3.3. Right: under elliptical symmetry, the mean and correlation are recovered exactly, whereas the covariance is recovered only up to scale, as per Corollary 3.5.

3.1 Setup and notation

We begin by fixing the common setup and notation for this section. Throughout, we work on $\mathcal{X} = \mathbb{R}^d$ with $d \geq 1$, equipped with its Borel σ-algebra $\mathcal{B}$, and denote by $\mathcal{P}$ the associated set of probability measures. For $n \in \mathbb{N}$, we let $\mathcal{P}_n \subseteq \mathcal{P}$ denote the set of probability measures with finite n-th moment. We write $\mathcal{S}_+^d$ and $\mathcal{S}_{++}^d$ for the sets of real $d \times d$ positive-semidefinite and positive-definite matrices, respectively, and $\mathcal{C}_+^d \subseteq \mathcal{S}_+^d$ for the matrices with unit diagonal entries.

We study recovery of the mean statistic $\mu \colon \mathcal{P} \rightharpoonup \mathbb{R}^d$, with $\mathrm{dom}(\mu) = \mathcal{P}_1$, and the covariance statistic $\Sigma \colon \mathcal{P} \rightharpoonup \mathcal{S}_+^d$, with $\mathrm{dom}(\Sigma) = \mathcal{P}_2$, in different symmetry settings. We also consider the correlation statistic $\rho \colon \mathcal{P} \rightharpoonup \mathcal{C}_+^d$, whose domain $\mathcal{P}_2^\circ \subseteq \mathcal{P}_2$ consists of distributions with strictly positive marginal variances. In all examples, we consider variational families of location–scale form.

Definition 3.1 (Location–scale family) *Fix a base distribution $Q_0 \in \mathcal{P}$. The location–scale family generated by Q_0 is the collection of distributions $\mathcal{Q} := \{Q_{\nu,S} \mid \nu \in \mathbb{R}^d,\ S \in \mathcal{S}_{++}^d\}$ where*

$$Q_{\nu,S} := (T_{\nu,S})_{\#} Q_0, \qquad \text{with} \qquad T_{\nu,S}(x) := \nu + S^{1/2} x.$$

Equivalently, if $X_0 \sim Q_0$, then $\nu + S^{1/2} X_0 \sim Q_{\nu,S}$. Since each $Q_{\nu,S}$ is the pushforward of Q_0 under an invertible affine map, $\mathcal{Q} \subseteq \mathcal{P}_n$ if and only if $Q_0 \in \mathcal{P}_n$, meaning that the mean and covariance are well-defined for the whole variational family whenever they are well-defined for the base distribution.

3.2 Recovery guarantees

We first consider the setting in which the target and variational family exhibit even symmetry.

Definition 3.2 (Even symmetry) *A distribution $\pi \in \mathcal{P}$ is* even symmetric *about $a \in \mathbb{R}^d$ if*

$$\pi(A) = \pi(2a - A) \qquad \text{for all } A \in \mathcal{B},$$

where $2a - A := \{2a - x \mid x \in A\}$.

Specialising Theorem 2.1 to this setting shows that a unique variational minimiser exactly recovers the target mean, as illustrated in the left panel of Figure 1.[7] The resulting corollary coincides with the mean recovery theorem of Margossian and Saul (2025).

Corollary 3.3 (cf. Theorem 8 in Margossian and Saul (2025)) *Fix $m \in \mathbb{R}^d$ and let $P \in \mathcal{P}_1$ be even symmetric about m. Let $Q_0 \in \mathcal{P}_1$ be even symmetric about 0, and let $\mathcal{Q}$ be the location–scale family it generates. If $D_f(P \| Q_{\nu,S})$ has a unique minimiser $(\nu^\star, S^\star)$ over $\mathbb{R}^d \times \mathcal{S}_{++}^d$, then*

$$\mu(Q_{\nu^\star,S^\star}) = \mu(P) = m.$$

Proof sketch. Consider the group $\mathcal{G} = \{\mathrm{id}, r_m\}$, where $r_m(x) = 2m - x$ is reflection about m. The target P is $\mathcal{G}$-invariant by even symmetry. Moreover, even symmetry of the base distribution Q_0 gives

[7] Code to reproduce all figures can be found at https://github.com/MarksDaniel/vi-symmetry-guarantees. All scripts were run on an Apple M3 Pro MacBook Pro with an 11-core CPU, a 14-core GPU, and 18 GB of unified memory.

$(r_m)_{\#}Q_{\nu,S} = Q_{2m-\nu,S}$, showing that the variational family $\mathcal{Q}$ is $\mathcal{G}$-closed. Finally, under reflection, the mean transforms according to

$$\mu\big((r_m)_{\#}\pi\big) = 2m - \mu(\pi), \qquad \pi \in \mathcal{P}_1.$$

The induced map $\rho_{r_m}(x) = 2m - x$ therefore has the unique fixed point m, giving $\Gamma_\mu(\mathcal{G}) = \{m\}$. Exact mean recovery follows from the fact that Theorem 2.1 places the means of both the target and the unique variational minimiser in this fixed set. A full proof is given in Section 3.3. □

It is worth noting that since reflection leaves the scale matrix unchanged, Corollary 3.3 also applies *mutatis mutandis* to VI over a location family, i.e., a location–scale family with fixed scale matrix.

We next consider the setting where the target and variational family exhibit elliptical symmetry.

Definition 3.4 (Elliptical symmetry) *A distribution $\pi \in \mathcal{P}$ is* elliptically symmetric *about $\gamma \in \mathbb{R}^d$ with scale matrix $C \in \mathcal{S}^d_{++}$ if there exists an $\mathrm{O}(d)$-invariant distribution $\pi_0 \in \mathcal{P}$ such that*

$$\pi = (T_{\gamma,C})_{\#}\pi_0, \qquad T_{\gamma,C}(x) \coloneqq \gamma + C^{1/2}x.$$

Here, $\mathrm{O}(d)$ is the orthogonal group, which consists of real $d \times d$ orthogonal matrices. Equivalently, if $X_0 \sim \pi_0$, then $\gamma + C^{1/2}X_0 \sim \pi$. Specialising Theorem 2.1 as before, but now to the elliptically symmetric setting, yields exact recovery of the mean and recovery of the covariance up to scale, which in turn gives exact correlation recovery, provided the variational minimiser is unique. We record this result in the following corollary, which again coincides with the corresponding recovery theorem of Margossian and Saul (2025). An illustration is provided in the right panel of Figure 1.

Corollary 3.5 (cf. Theorem 10 in Margossian and Saul (2025)) *Fix $m \in \mathbb{R}^d$ and $M \in \mathcal{S}^d_{++}$, and let P be elliptically symmetric about m with scale matrix M. Let Q_0 be an $\mathrm{O}(d)$-invariant base distribution, and let $\mathcal{Q}$ be the location–scale family it generates. If $D_f(P\|Q_{\nu,S})$ has a unique minimiser $(\nu^\star, S^\star)$ over $\mathbb{R}^d \times \mathcal{S}^d_{++}$, then*

(i) if $P, Q_0 \in \mathcal{P}_1$, then

$$\mu(Q_{\nu^\star,S^\star}) = \mu(P) = m,$$

(ii) if $P, Q_0 \in \mathcal{P}_2$ and are not Dirac measures then

$$\Sigma(Q_{\nu^\star,S^\star}), \Sigma(P) \in \{\lambda M \mid \lambda > 0\}$$

and

$$\rho(Q_{\nu^\star,S^\star}) = \rho(P) = \mathrm{diag}(M)^{-1/2} M \,\mathrm{diag}(M)^{-1/2}.$$

Proof sketch. Elliptical symmetry implies even symmetry, so mean recovery follows directly from Corollary 3.3. For covariance recovery, consider the group of affine symmetries of the ellipsoid centred at m with scale matrix M, defined as

$$\mathcal{G} = \{g_R \mid R \in \mathrm{O}(d)\}, \qquad g_R(x) = m + A_R(x - m), \qquad A_R = M^{1/2} R M^{-1/2}.$$

The target P is $\mathcal{G}$-invariant by elliptical symmetry. Moreover, orthogonal invariance of the base distribution Q_0 ensures that the variational family $\mathcal{Q}$ is $\mathcal{G}$-closed. Under these transformations, the covariance satisfies

$$\Sigma((g_R)_{\#}\pi) = A_R \Sigma(\pi) A_R^\top, \qquad \pi \in \mathcal{P}_2,$$

giving the induced maps $\rho_{g_R}(V) = A_R V A_R^\top$. Their common fixed set is $\Gamma_\Sigma(\mathcal{G}) = \{\lambda M \mid \lambda \geq 0\}$. Indeed, a matrix $V \in \mathcal{S}^d_+$ is fixed by all these maps precisely when $M^{-1/2} V M^{-1/2}$ is invariant under conjugation by every orthogonal matrix. This is equivalent to $M^{-1/2} V M^{-1/2} = \lambda I_d$, or $V = \lambda M$, for some $\lambda \geq 0$. Theorem 2.1 places the covariance matrices of both the target and the unique variational minimiser in this fixed set, and the non-Dirac assumptions ensure that $\lambda > 0$ in each case, yielding covariance recovery up to a positive scalar multiple. Exact correlation recovery then follows because normalising the covariance matrices removes this scale ambiguity. A full proof is given in Section 3.3. □

Remark 3.6 *Note that although the conditions of Theorem 2.1 are sufficient for statistic recovery, they are not always necessary. Indeed, correlation is recovered in Corollary 3.5 through the covariance, even though it does not itself satisfy Condition (3) for $d \geq 2$. This is because correlation discards marginal variances, which can affect the transformed correlations under the action of $\mathcal{G}$. See Appendix B.2 for a discussion.*

3.3 Proofs of the recovery guarantees

We now prove the guarantees in Section 3.2, starting with mean recovery under even symmetry.

Proof of Corollary 3.3. Fix $m \in \mathbb{R}^d$ and let $P \in \mathcal{P}_1$ be even symmetric about m. Further, let $\mathcal{Q} \subseteq \mathcal{P}_1$ be the location–scale family generated by a base distribution $Q_0 \in \mathcal{P}_1$ that is even symmetric about 0.[8] Let $\mathcal{G} \coloneqq \{\mathrm{id}, r_m\}$, where id denotes the identity map and $r_a(x) \coloneqq 2a - x$ for $a \in \mathbb{R}^d$. We begin by verifying Conditions (1)–(3):

(1) Trivially, we know that $\mathrm{id}_{\#}P = P$. Moreover, since $r_m = r_m^{-1}$, we have for every $A \in \mathcal{B}$,

$$(r_m)_{\#}P(A) = P(r_m^{-1}(A)) = P(r_m(A)) = P(A),$$

where the last equality follows from the even symmetry of P about m. Thus P is $\mathcal{G}$-invariant.

(2) For any $(\nu, S) \in \mathbb{R}^d \times \mathcal{S}_{++}^d$, we have $\mathrm{id}_{\#}Q_{\nu,S} = Q_{\nu,S} \in \mathcal{Q}$. Moreover, it is easy to verify that $r_m \circ T_{\nu,S} = T_{2m-\nu,S} \circ r_0$, and therefore

$$\begin{aligned}(r_m)_{\#}Q_{\nu,S} &= (r_m \circ T_{\nu,S})_{\#}Q_0 = (T_{2m-\nu,S} \circ r_0)_{\#}Q_0 \\ &= (T_{2m-\nu,S})_{\#}\big((r_0)_{\#}Q_0\big) = Q_{2m-\nu,S} \in \mathcal{Q},\end{aligned} \tag{1}$$

where the last equality follows from the even symmetry of Q_0 about 0. Hence, $\mathcal{Q}$ is $\mathcal{G}$-closed.

(3) First, the domain $\mathrm{dom}(\mu) = \mathcal{P}_1$ is $\mathcal{G}$-closed. Indeed, if $\pi \in \mathcal{P}_1$, then $\mathrm{id}_{\#}\pi = \pi \in \mathcal{P}_1$ and, moreover, $(r_m)_{\#}\pi \in \mathcal{P}_1$, since

$$\int_{\mathbb{R}^d} \|x\|\, d((r_m)_{\#}\pi)(x) = \int_{\mathbb{R}^d} \|2m - x\|\, d\pi(x) \le 2\|m\| + \int_{\mathbb{R}^d} \|x\|\, d\pi(x) < \infty.$$

Further, for every $\pi \in \mathcal{P}_1$, we can compute $\mu(g_{\#}\pi)$ for every $g \in \mathcal{G}$ as

$$\mu(g_{\#}\pi) = \begin{cases} \mu(\pi), & g = \mathrm{id}, \\ 2m - \mu(\pi), & g = r_m. \end{cases} \tag{2}$$

As the right-hand side of (2) depends only on $\mu(\pi)$, it follows that if $\mu(\pi_1) = \mu(\pi_2)$, then $\mu(g_{\#}\pi_1) = \mu(g_{\#}\pi_2)$ for all $\pi_1, \pi_2 \in \mathcal{P}_1$ and all $g \in \mathcal{G}$.

We have thus verified the conditions of Theorem 2.1. Moreover, by (2), Condition (3) allows us to define the maps $\rho_{\mathrm{id}}(x) = x$ and $\rho_{r_m}(x) = 2m - x$ on $\mathrm{Im}(\mu) = \mathbb{R}^d$ and compute the fixed set

$$\Gamma_\mu(\mathcal{G}) = \{x \in \mathbb{R}^d \mid \rho_g(x) = x,\ \forall g \in \mathcal{G}\} = \{x \in \mathbb{R}^d \mid 2m - x = x\} = \{m\}.$$

Applying Theorem 2.1 and using the assumed uniqueness of the variational minimiser therefore gives

$$\mu(Q_{\nu^\star,S^\star}) = \mu(P) = m.$$

□

We next turn to mean, covariance, and correlation recovery under elliptical symmetry.

Proof of Corollary 3.5. Let $P \in \mathcal{P}$ be elliptically symmetric about $m \in \mathbb{R}^d$ with scale matrix $M \in \mathcal{S}_{++}^d$, and let $\mathcal{Q} \subseteq \mathcal{P}$ be the location–scale family generated by an $\mathrm{O}(d)$-invariant base distribution $Q_0 \in \mathcal{P}$.[9]

Mean recovery. Observe that a distribution that is elliptically symmetric about $\gamma \in \mathbb{R}^d$ is also even symmetric about γ. Indeed, if $\pi = (T_{\gamma,C})_{\#}\pi_0$ for some $\mathrm{O}(d)$-invariant π_0, then for every $A \in \mathcal{B}$

$$\pi(2\gamma - A) = \pi_0(T_{\gamma,C}^{-1}(2\gamma - A)) = \pi_0(-T_{\gamma,C}^{-1}(A)) = \pi_0(T_{\gamma,C}^{-1}(A)) = \pi(A),$$

where the penultimate step follows from the fact that $\mathrm{O}(d)$-invariance of π_0 implies $(-I_d)_{\#}\pi_0 = \pi_0$. Therefore, if $P, Q_0 \in \mathcal{P}_1$, mean recovery is immediate by Corollary 3.3.

[8] This is equivalent to requiring that each $Q_{\nu,S} \in \mathcal{Q}$ is even symmetric about its location parameter ν.

[9] This is equivalent to requiring that each $Q_{\nu,S} \in \mathcal{Q}$ is elliptically symmetric about ν with scale matrix S.

Covariance recovery. For covariance recovery, we assume that $P, Q_0 \in \mathcal{P}_2$ and consider the group

$$\mathcal{G} \coloneqq \{g_R \mid R \in \mathrm{O}(d)\}, \qquad g_R(x) \coloneqq m + A_R(x-m), \qquad A_R \coloneqq M^{1/2} R M^{-1/2}.$$

We begin, once again, by verifying Conditions (1)–(3) under the action of this group:

(1) By elliptical symmetry, there exists an $\mathrm{O}(d)$-invariant $P_0 \in \mathcal{P}_2$ such that $P = (T_{m,M})_{\#}P_0$. It is easy to check that $g_R \circ T_{m,M} = T_{m,M} \circ R$ for every $R \in \mathrm{O}(d)$, and therefore

$$(g_R)_{\#}P = (g_R \circ T_{m,M})_{\#}P_0 = (T_{m,M} \circ R)_{\#}P_0 = (T_{m,M})_{\#}(R_{\#}P_0) = P.$$

(2) Let $(\nu, S) \in \mathbb{R}^d \times \mathcal{S}^d_{++}$ and $R \in \mathrm{O}(d)$. Define

$$\nu' \coloneqq m + A_R(\nu - m), \qquad S' \coloneqq A_R S A_R^\top, \qquad U \coloneqq (S')^{-1/2} A_R S^{1/2}.$$

Then $U \in \mathrm{O}(d)$ because $UU^\top = I_d$, and since Q_0 is $\mathrm{O}(d)$-invariant, we have $U_{\#}Q_0 = Q_0$. Moreover, we have that $g_R \circ T_{\nu,S} = T_{\nu',S'} \circ U$ and therefore

$$(g_R)_{\#}Q_{\nu,S} = (g_R \circ T_{\nu,S})_{\#}Q_0 = (T_{\nu',S'} \circ U)_{\#}Q_0 = (T_{\nu',S'})_{\#}(U_{\#}Q_0) = Q_{\nu',S'} \in \mathcal{Q}.$$

(3) First, the domain $\mathcal{P}_2$ is $\mathcal{G}$-closed. Indeed, for all $x \in \mathbb{R}^d$

$$\|g_R(x)\| \le \|m\| + \|A_R\|_{\mathrm{op}}(\|x\| + \|m\|). \tag{3}$$

Squaring (3) and using the inequality $(a+b)^2 \le 2a^2 + 2b^2$ gives, for any $\pi \in \mathcal{P}_2$,

$$\int_{\mathbb{R}^d} \|x\|^2 \, d((g_R)_{\#}\pi)(x) \le 2\|A_R\|^2_{\mathrm{op}} \int_{\mathbb{R}^d} \|x\|^2 \, d\pi(x) + 2(1 + \|A_R\|_{\mathrm{op}})^2 \|m\|^2 < \infty,$$

since $\|A_R\|_{\mathrm{op}} < \infty$ by the submultiplicativity of the operator norm. Moreover, for every $\pi \in \mathcal{P}_2$ and every $R \in \mathrm{O}(d)$, we have $\mu((g_R)_{\#}\pi) = g_R(\mu(\pi))$ and so

$$\begin{aligned} \Sigma((g_R)_{\#}\pi) &= \int_{\mathbb{R}^d} \big(g_R(x) - g_R(\mu(\pi))\big)\big(g_R(x) - g_R(\mu(\pi))\big)^\top \, d\pi(x) \\ &= A_R \left(\int_{\mathbb{R}^d} (x - \mu(\pi))(x - \mu(\pi))^\top \, d\pi(x) \right) A_R^\top \\ &= A_R \, \Sigma(\pi) \, A_R^\top, \end{aligned} \tag{4}$$

meaning that if $\Sigma(\pi_1) = \Sigma(\pi_2)$, then $\Sigma((g_R)_{\#}\pi_1) = \Sigma((g_R)_{\#}\pi_2)$ for all $R \in \mathrm{O}(d)$.

By (4), Condition (3) allows us to define for every $g_R \in \mathcal{G}$ the map $\rho_{g_R}(V) = A_R V A_R^\top$ on $\mathrm{Im}(\Sigma) = \mathcal{S}^d_+$ and find the positive-semidefinite matrices satisfying $\rho_{g_R}(V) = V$ for all $g_R \in \mathcal{G}$:

$$\Gamma_\Sigma(\mathcal{G}) = \{V \in \mathcal{S}^d_+ \mid A_R V A_R^\top = V, \ \forall R \in \mathrm{O}(d)\} = \{\lambda M \mid \lambda \ge 0\}.$$

The last equality follows from the fact that a matrix $V \in \mathcal{S}^d_+$ lies in the fixed set if and only if $M^{-1/2} V M^{-1/2}$ is invariant under conjugation by every $R \in \mathrm{O}(d)$, which in turn implies that $M^{-1/2} V M^{-1/2} = \lambda I_d$ for some $\lambda \ge 0$. It follows that if P is not a Dirac measure, then $\Sigma(P) = \alpha M$ for some $\alpha > 0$. Similarly, if Q_0 is not a Dirac measure and $(\nu^\star, S^\star)$ is a unique minimiser of $D_f(P \| Q_{\nu,S})$ over $\mathbb{R}^d \times \mathcal{S}^d_{++}$, then $\Sigma(Q_{\nu^\star,S^\star}) = bM$ for some $b > 0$. Therefore, outside the trivial point-mass case, a unique minimiser recovers the covariance up to a positive multiplicative constant.

Correlation recovery. Since M is positive definite and $\alpha, b > 0$, the covariance matrices αM and bM have strictly positive diagonal entries, so $P, Q_{\nu^\star,S^\star} \in \mathcal{P}_2^\circ$. Recall that, for any $\pi \in \mathcal{P}_2^\circ$, the correlation statistic is given by

$$\rho(\pi) = D(\pi)^{-1/2} \Sigma(\pi) D(\pi)^{-1/2}, \qquad D(\pi) \coloneqq \mathrm{diag}(\Sigma(\pi)).$$

Substituting the covariance expressions above and cancelling the scalar factors gives

$$\rho(Q_{\nu^\star,S^\star}) = \mathrm{diag}(M)^{-1/2} M \, \mathrm{diag}(M)^{-1/2} = \rho(P).$$

□

4 von Mises–Fisher family: new guarantees under rotational symmetry

Having established known statistic recovery results as special cases, we now use our framework to obtain a new guarantee for variational inference on the sphere. Specifically, we show that under rotational symmetry, VI in a von Mises–Fisher family can exactly recover the axis of symmetry. Unlike the location–scale examples, where the relevant symmetry, family, and statistic were known in advance, here, we instead start from the target symmetry, identify a statistic it pins down, and select a variational family closed under these symmetries to guarantee recovery. This section thus not only establishes a new result but also provides a blueprint for applying our framework to new symmetries and statistics. We highlight that these recovery guarantees are agnostic to the optimisation scheme used to compute the variational optimum. Practical implementation, including the choice of algorithm and the construction of suitable gradient estimators, requires separate consideration.

Let $\mathcal{X} = \mathcal{S}^{d-1}$ be the unit sphere in $\mathbb{R}^d$, $d \geq 3$, equipped with its Borel σ-algebra $\mathcal{B}$, and let $\mathcal{P}$ denote the associated set of probability measures. To ease the presentation, we work with distributions that are absolutely continuous with respect to the uniform probability measure σ on the unit sphere, which we collect inside the set $\mathcal{P}_\sigma \subseteq \mathcal{P}$. For any $\pi \in \mathcal{P}_\sigma$, we denote its density with respect to σ by f_π.

Definition 4.1 (Rotational symmetry) *A distribution $\pi \in \mathcal{P}_\sigma$ is* rotationally symmetric *about axis direction $w \in \mathcal{S}^{d-1}$ if there exists a measurable function $\psi\colon [-1,1] \to \mathbb{R}_{\geq 0}$ such that*

$$f_\pi(x) = \psi(w^\top x), \qquad \text{for } \sigma\text{-a.e. } x \in \mathcal{S}^{d-1}.$$

We call ψ the axial profile *of π. If ψ is constant, then π is the uniform distribution on $\mathcal{S}^{d-1}$.*

Suppose we are given a non-uniform target $P \in \mathcal{P}_\sigma$ that is rotationally symmetric about $u \in \mathcal{S}^{d-1}$. Observe that the level sets of the map $x \mapsto u^\top x$ are precisely the parallels orthogonal to u. Thus, a rotationally symmetric distribution may vary only with the polar angle away from u. In particular, it is invariant under the group $\mathcal{G} \coloneqq \{g_R \mid Ru = u,\ R \in \mathrm{SO}(d)\}$, where $\mathrm{SO}(d)$ is the special orthogonal group of orthogonal $d \times d$ matrices with determinant 1, and $g_R(x) = Rx$ for all $x \in \mathcal{S}^{d-1}$. Indeed, $(g_R)_\# P = P$ for all $g_R \in \mathcal{G}$, since by rotation invariance of σ we have for σ-a.e. $x \in \mathcal{S}^{d-1}$,

$$f_{(g_R)_\# P}(x) = f_P(R^{-1}x) = \psi(u^\top R^{-1}x) = \psi((Ru)^\top x) = \psi(u^\top x) = f_P(x).$$

Given this target symmetry, we now ask which statistic, if any, it pins down. A natural first candidate is the axis direction. However, if $\pi \in \mathcal{P}_\sigma$ is rotationally symmetric about w with axial profile ψ, it is also rotationally symmetric about $-w$ with axial profile $\tilde{\psi}(t) \coloneqq \psi(-t)$. Thus, the most the symmetry structure can uniquely determine is a one-dimensional subspace of $\mathbb{R}^d$. More precisely, if $\pi \in \mathcal{P}_\sigma$ is non-uniform and rotationally symmetric about both $w_1, w_2 \in \mathcal{S}^{d-1}$, then $\mathrm{span}(w_1) = \mathrm{span}(w_2)$; proof in Appendix C.1. We therefore consider the space of lines through the origin in $\mathbb{R}^d$ denoted by $\mathbb{RP}^{d-1}$, let $\mathcal{P}_\mathcal{A} \subseteq \mathcal{P}_\sigma$ be the set of all non-uniform rotationally symmetric distributions, and define the axis statistic $\mathcal{A}\colon \mathcal{P} \rightharpoonup \mathbb{RP}^{d-1}$ on $\mathrm{dom}(\mathcal{A}) = \mathcal{P}_\mathcal{A}$ by $\mathcal{A}(\pi) = \mathrm{span}(w)$, where w is any axis direction of rotational symmetry of π. $\mathcal{P}_\mathcal{A}$ is $\mathcal{G}$-closed, since for any $\pi \in \mathcal{P}_\mathcal{A}$ rotationally symmetric about w with axial profile ψ and any $g_R \in \mathcal{G}$,

$$f_{(g_R)_\#\pi}(x) = f_\pi\left(R^{-1}x\right) = \psi(w^\top R^{-1}x) = \psi\left((Rw)^\top x\right), \qquad \text{for } \sigma\text{-a.e. } x \in \mathcal{S}^{d-1},$$

meaning $(g_R)_\#\pi$ is rotationally symmetric about Rw with profile ψ and thus $(g_R)_\#\pi \in \mathcal{P}_\mathcal{A}$. Hence,

$$\mathcal{A}((g_R)_\#\pi) = \mathrm{span}(Rw) = R\,\mathrm{span}(w) = R\mathcal{A}(\pi), \tag{5}$$

which implies that if $\mathcal{A}(\pi_1) = \mathcal{A}(\pi_2)$, then $\mathcal{A}((g_R)_\#\pi_1) = \mathcal{A}((g_R)_\#\pi_2)$ for all $g_R \in \mathcal{G}$. By (5), we can define for every $g_R \in \mathcal{G}$ the map $\rho_{g_R}(L) \coloneqq RL$ on $\mathrm{Im}(\mathcal{A}) = \mathbb{RP}^{d-1}$ to obtain the fixed set

$$\Gamma_\mathcal{A}(\mathcal{G}) = \{L \in \mathbb{RP}^{d-1} \mid \rho_{g_R}(L) = L,\ \forall g_R \in \mathcal{G}\} = \{\mathrm{span}(u)\}.$$

By construction, $P \in \mathcal{P}_\mathcal{A}$. We thus know that choosing a $\mathcal{G}$-closed family will guarantee the recovery of the axis statistic by a unique minimiser. One such family is the *von Mises–Fisher family*. Each of its members is rotationally symmetric about a mean direction $\nu \in \mathcal{S}^{d-1}$, and indexed by a concentration parameter $\kappa > 0$ controlling how quickly the density decays as the angle from ν increases.

Definition 4.2 (von Mises–Fisher family) *A von Mises–Fisher (vMF) family $\mathcal{Q} \subseteq \mathcal{P}_\sigma$ is a two-parameter family $\{Q_{\nu,\kappa} \mid \nu \in \mathcal{S}^{d-1}, \kappa \in \mathbb{R}_{>0}\}$, where for a normalising constant $c_d(\kappa) > 0$,*

$$f_{Q_{\nu,\kappa}}(x) = c_d(\kappa)\exp(\kappa\nu^\top x), \qquad \text{for } \sigma\text{-a.e. } x \in \mathcal{S}^{d-1}.$$

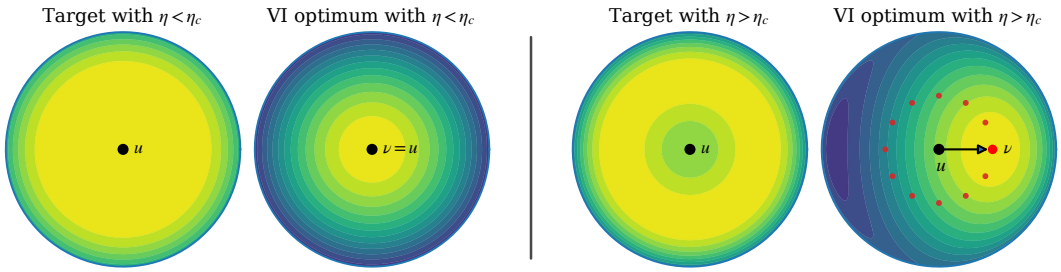


Figure 2: Log-density contours of the target and variational posterior are shown using a Lambert azimuthal equal-area projection centred at the true direction u. Left: $\eta < \eta_c$ and the reverse KL has a unique minimiser at $\nu = u$, recovering the symmetry axis. Right: $\eta > \eta_c$ and the minimisers form a latitude circle. None of them, including the one found during optimisation, recovers the statistic.

Indeed, the von Mises–Fisher family is $\mathcal{G}$-closed. For all $(\nu, \kappa) \in \mathcal{S}^{d-1} \times \mathbb{R}_{>0}$ and all $g_R \in \mathcal{G}$, $(g_R)_{\#} Q_{\nu,\kappa} = Q_{R\nu,\kappa} \in \mathcal{Q}$ since by rotation invariance of σ, we have that for σ-a.e. $x \in \mathcal{S}^{d-1}$

$$f_{(g_R)_{\#} Q_{\nu,\kappa}}(x) = c_d(\kappa) \exp\big(\kappa \nu^\top R^{-1} x\big) = c_d(\kappa) \exp\big(\kappa (R\nu)^\top x\big) = f_{Q_{R\nu,\kappa}}(x), \tag{6}$$

and since $\kappa > 0$, it follows that $\mathcal{Q} \subseteq \mathcal{P}_{\mathcal{A}}$. We thus obtain the following new recovery guarantee.

Corollary 4.3 *Let $P \in \mathcal{P}_{\mathcal{A}}$ be rotationally symmetric about $u \in \mathcal{S}^{d-1}$, and let $\mathcal{Q} \subseteq \mathcal{P}_{\mathcal{A}}$ be the von Mises–Fisher family. If $D_f(P \| Q_{\nu,\kappa})$ has a unique minimiser $(\nu^\star, \kappa^\star)$ over $\mathcal{S}^{d-1} \times \mathbb{R}_{>0}$, then*

$$\mathcal{A}(Q_{\nu^\star,\kappa^\star}) = \mathcal{A}(P) = \operatorname{span}(u).$$

Analogously to location and location–scale families, it is natural to consider a fixed-concentration von Mises–Fisher family, namely, the one-parameter subfamily of distributions that share the same concentration parameter $\kappa_0 > 0$. Since the pushforward action leaves the concentration parameter unchanged as per (6), Corollary 4.3 equally applies to fixed-concentration vMF families.

Remark 4.4 *As with all corollaries of Theorem 2.1, recovery can fail when the variational minimiser is not unique. In what follows, we highlight this with a concrete example.*

Fix $\lambda \neq 0$ and $u \in \mathcal{S}^{d-1}$ and let $(P_{\lambda,\eta})_{\eta>0} \subseteq \mathcal{P}_{\mathcal{A}}$ be a one-parameter family of target distributions such that for each $\eta > 0$, $P_{\lambda,\eta}$ is rotationally symmetric about u with axial profile

$$\psi_{\lambda,\eta}(t) \coloneqq Z_{\lambda,\eta}^{-1} \exp\left(\lambda t - \eta t^2\right), \qquad t \in [-1, 1],$$

where $Z_{\lambda,\eta} > 0$ is a normalising constant. Let $\mathcal{Q}_{\kappa_0}$ be a fixed-concentration von Mises–Fisher family for some $\kappa_0 > 0$ and consider the reverse KL objective $D_{\mathrm{KL}}(Q_{\nu,\kappa_0} \| P_{\lambda,\eta})$. Since $\eta > 0$, $P_{\lambda,\eta}$ is not a vMF distribution and the model is misspecified. Despite that, there exists a critical threshold $\eta_c(d, \lambda, \kappa_0) > 0$ such that for all $0 < \eta \leq \eta_c$, the objective $D_{\mathrm{KL}}(Q_{\nu,\kappa_0} \| P_{\lambda,\eta})$ has a unique minimiser $\nu^\star$ over $\mathcal{S}^{d-1}$, and exact recovery holds with $\mathcal{A}(Q_{\nu^\star,\kappa_0}) = \mathcal{A}(P_{\lambda,\eta})$. By contrast, for every $\eta > \eta_c$, the minimiser set is non-singleton and no minimiser recovers the true axis statistic. This behaviour is visualised for $d = 3$ in Figure 2, with an explicit construction given in Appendix C.2.

5 Discussion

In this work, we have developed a general symmetry-based theory of statistic recovery in variational inference, and have used it both to recover existing results and to derive novel guarantees. With our framework providing a modular blueprint for establishing statistic recovery, we see significant opportunities to apply it across different domains, symmetry classes, and variational families. At the same time, we identify at least two important directions for future work. First, while our theory provides sufficient conditions for statistic recovery, we have shown that these conditions are not always necessary. Deriving necessary and sufficient conditions would therefore yield a complete characterisation of statistic recovery under symmetry in variational problems. Second, it would be valuable to extend the theory to settings where the target is only approximately symmetric. Such an extension could unlock new guarantees for a broad class of inference problems that exhibit substantial structure but do not exactly satisfy our assumptions — a direction we leave for future work.

Acknowledgments and Disclosure of Funding

D. Marks acknowledges funding from a G-Research graduate scholarship. D. Paccagnan was partially supported by the EPSRC grant EP/Y001001/1, funded by the International Science Partnerships Fund (ISPF) and UKRI; and by an Imperial-MIT seed fund.

References

Pierre Alquier and James Ridgway. Concentration of tempered posteriors and of their variational approximations. *The Annals of Statistics*, 48(3):1475–1497, June 2020.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, April 2017.

Kamélia Daudel, Joe Benton, Yuyang Shi, and Arnaud Doucet. Alpha-divergence variational inference meets importance weighted auto-encoders: Methodology and asymptotics. *Journal of Machine Learning Research*, 24(243):1–83, 2023.

Adji Bousso Dieng, Dustin Tran, Rajesh Ranganath, John Paisley, and David Blei. Variational inference via χ-upper bound minimization. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Justin Domke. Provable smoothness guarantees for black-box variational inference. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 2587–2596. PMLR, July 2020.

E. B. Dynkin. The maximal subgroups of the classical groups. In *American Mathematical Society Translations, Series 2*, volume 6, pages 245–378. American Mathematical Society, 1957.

Tomas Geffner and Justin Domke. On the difficulty of unbiased alpha divergence minimization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 3650–3659. PMLR, July 2021a.

Tomas Geffner and Justin Domke. Empirical evaluation of biased methods for alpha divergence minimization. arXiv preprint arXiv:2105.06587, May 2021b.

Ryan Giordano, Tamara Broderick, and Michael I. Jordan. Covariances, robustness and variational Bayes. *J. Mach. Learn. Res.*, 19(1):1981–2029, January 2018.

Wei Han and Yun Yang. Statistical Inference in Mean-Field Variational Bayes. arXiv preprint arXiv:1911.01525, November 2019.

Jose Hernandez-Lobato, Yingzhen Li, Mark Rowland, Thang Bui, Daniel Hernandez-Lobato, and Richard Turner. Black-box alpha divergence minimization. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 1511–1520, June 2016.

Geoffrey E. Hinton and Drew Van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory - COLT '93*, pages 5–13, Santa Cruz, California, United States, 1993. ACM Press.

Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic Variational Inference. *Journal of Machine Learning Research*, 14(40):1303–1347, 2013.

Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, November 1999.

Anya Katsevich and Philippe Rigollet. On the approximation accuracy of Gaussian variational inference. *The Annals of Statistics*, 52(4):1384–1409, August 2024.

Yingzhen Li and Richard E Turner. Rényi divergence variational inference. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

David J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, May 1992.

David J. C. MacKay. *Information theory, inference, and learning algorithms*. Cambridge University Press, Cambridge, UK ; New York, 2003.

Kanti V. Mardia and Peter E. Jupp. *Directional statistics*. John Wiley & Sons, September 2009.

Charles Margossian and Lawrence K. Saul. Variational inference in location-scale families: Exact recovery of the mean and correlation matrix. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, pages 3466–3474, April 2025.

Charles C. Margossian, Isaac E. Rankin, and Lawrence K. Saul. Generalized guarantees for variational inference in the presence of even and elliptical symmetry. arXiv preprint arXiv:2511.01064v3, May 2026.

Thomas P Minka. Expectation propagation for approximate Bayesian inference. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, 2001.

Dragoslav S. Mitrinović. *Analytic inequalities*. Springer Berlin Heidelberg, Berlin, Heidelberg, 1970.

Christian A. Naesseth, Fredrik Lindsten, and David Blei. Markovian score climbing: Variational inference with KL(p||q). In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 15499–15510, Red Hook, NY, USA, December 2020. Curran Associates Inc.

Constantin P. Niculescu and Lars-Erik Persson. *Convex functions and their applications*. CMS Books in Mathematics. Springer International Publishing, Cham, 2018.

Debdeep Pati, Anirban Bhattacharya, and Yun Yang. On statistical optimality of variational Bayes. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, pages 1579–1588, March 2018.

Carsten Peterson and James R Anderson. A mean field theory learning algorithm for neural networks. *Complex Systems*, 1:995–1019, 1987.

Yury Polyanskiy and Yihong Wu. Information theory: From coding to learning, January 2025.

Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, pages 814–822, April 2014.

Jean-Baptiste Regli and Ricardo Silva. Alpha-beta divergence for variational inference. arXiv preprint arXiv:1805.01045, May 2018.

Christian P. Robert and George Casella. *Monte Carlo statistical methods*. Springer Texts in Statistics. Springer, New York, NY, 2004.

Igal Sason and Sergio Verdu. f-divergence inequalities. *IEEE Trans. Inf. Theor.*, 62(11):5973–6006, November 2016.

Terence Tao. *An introduction to measure theory*. Number v. 126 in Graduate studies in mathematics. American Mathematical Society, Providence, R.I, 2011.

Fuichi Uchida. Classification of real analytic $\mathrm{SL}(n, \mathbb{R})$ actions on n-sphere. *Osaka Journal of Mathematics*, 16(3):561–579, 1979.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, December 2008.

Yixin Wang and David M. Blei. Frequentist consistency of variational Bayes. *Journal of the American Statistical Association*, 114(527):1147–1161, July 2019.

Yun Yang, Debdeep Pati, and Anirban Bhattacharya. α-variational inference with statistical guarantees. *The Annals of Statistics*, 48(2), April 2020.

Lena Zellinger and Antonio Vergari. Even more guarantees for variational inference in the presence of symmetries. arXiv preprint arXiv:2604.21407v2, June 2026.

Fengshuo Zhang and Chao Gao. Convergence rates of variational posterior distributions. *The Annals of Statistics*, 48(4):2180–2207, August 2020.

A Additional material for Section 2

A.1 Overview of f-divergences

In this appendix, we give a formal definition of f-divergences, which underpin our formulation of the variational inference problem throughout the paper. Let $(\mathcal{X}, \mathcal{B})$ be a measurable space, and denote by $\mathcal{P}$ the set of associated probability measures. Let $P, Q \in \mathcal{P}$ be two such probability measures, and let μ be any σ-finite measure on $(\mathcal{X}, \mathcal{B})$ such that $P \ll \mu$ and $Q \ll \mu$, e.g. $\mu = P + Q$. Write $p = dP/d\mu$ and $q = dQ/d\mu$ for the corresponding densities. Let $f\colon \mathbb{R}_{>0} \to \mathbb{R}$ be a convex function with $f(1) = 0$, and define the extensions:

$$f(0) \coloneqq \lim_{t \downarrow 0} f(t), \qquad f'(\infty) \coloneqq \lim_{t \downarrow 0} t f\left(\frac{1}{t}\right).$$

The f-divergence generated by f, denoted $D_f\colon \mathcal{P} \times \mathcal{P} \to [0, \infty]$, is defined by

$$D_f(P\|Q) \coloneqq \int_{\{q>0\}} q f\left(\frac{p}{q}\right) d\mu + f'(\infty)\, P(\{q = 0\}), \tag{7}$$

with the convention that $0 \cdot \infty = 0$. This definition follows Polyanskiy and Wu (2025, Def. 7.1).[10] A number of popular divergences can be recovered as special cases of f-divergences with different generators. These include the Kullback–Leibler (KL) divergence with $f(t) = t \log t$, the χ^2-divergence with $f(t) = (t-1)^2$, the total variation distance with $f(t) = \frac{1}{2}|t-1|$, and the squared Hellinger distance with $f(t) = (1-\sqrt{t})^2$. For any f-divergence, swapping the order of the arguments again yields an f-divergence, now generated by $\tilde{f}(t) = tf(1/t)$. That is, $D_f(Q\|P) = D_{\tilde{f}}(P\|Q)$. For example, since the forward KL divergence is generated by $f(t) = t \log t$, the reverse KL divergence is generated by $\tilde{f}(t) = tf(1/t) = -\log t$.

For a detailed overview of f-divergences, see Sason and Verdu (2016).

A.2 Proof of Theorem 2.1

In this appendix, we prove a generalisation of Theorem 2.1 that allows multiple global minimisers, provided they all agree on the statistic of interest. The proof is divided into three parts. In the first part, we show that f-divergences are invariant under simultaneous pushforward by measurable automorphisms. In the second part, we use this invariance and Conditions (1) and (2) to show that the set of global minimisers is $\mathcal{G}$-closed. In the third part, using the maps ρ_g introduced in Section 2 and guaranteed to be well defined by Condition (3), we show that the target's statistic and the common statistic value of the global minimisers lie in $\Gamma_S(\mathcal{G})$. Theorem 2.1 then follows as a special case.

Setup. We formally recall the setup of Section 2. Let $(\mathcal{X}, \mathcal{B})$ be a measurable space, and denote by $\mathcal{P}$ the set of probability measures on $(\mathcal{X}, \mathcal{B})$. Let $\mathrm{Aut}(\mathcal{X}, \mathcal{B})$ be the group of measurable automorphisms, acting on $\mathcal{P}$ by pushforward, and fix a subgroup $\mathcal{G} \leq \mathrm{Aut}(\mathcal{X}, \mathcal{B})$. Throughout this section, we consider the variational inference problem ($\star$) with f-divergences

$$\mathcal{Q}^\star \coloneqq \operatorname*{argmin}_{Q \in \mathcal{Q}} D_f(P\|Q).$$

For the first two parts of the proof, which concern only the variational problem and the action of $\mathcal{G}$ on distributions, we work with a general $P \in \mathcal{P}$ and $\mathcal{Q} \subseteq \mathcal{P}$. For the third part, we additionally consider a statistic $S\colon \mathcal{P} \rightharpoonup \mathcal{Y}$, with domain $\mathrm{dom}(S)$, and assume that $P \in \mathrm{dom}(S)$ and $\mathcal{Q} \subseteq \mathrm{dom}(S)$. Whenever Conditions (1)–(3) are invoked below, we mean the three conditions stated in Section 2.

We begin with the first part of the proof, and show that f-divergences are invariant under simultaneous pushforward by measurable automorphisms.

Lemma A.1 (f-divergence invariance) *For any $P, Q \in \mathcal{P}$ and $g \in \mathrm{Aut}(\mathcal{X}, \mathcal{B})$,*

$$D_f\big(g_\# P \,\big\|\, g_\# Q\big) = D_f\big(P \,\big\|\, Q\big).$$

[10] Note that Polyanskiy and Wu (2025) work with the standing assumption that $(\mathcal{X}, \mathcal{B})$ is a standard Borel space. However, their f-divergence definition extends verbatim to the case of an arbitrary measurable space, such as the one considered here.

Proof . Let μ be a σ-finite common dominating measure of P and Q, write $p = dP/d\mu$ and $q = dQ/d\mu$ for the corresponding densities, and define $\tilde{\mu} = g_{\#}\mu$. It is easy to verify that $\tilde{\mu}$ is σ-finite and that $g_{\#}P, g_{\#}Q \ll \tilde{\mu}$. Define $\tilde{p} := p \circ g^{-1}$ and $\tilde{q} := q \circ g^{-1}$. Then for every $A \in \mathcal{B}$,

$$\int_A \tilde{p}\, d\tilde{\mu} = \int_{\mathcal{X}} \mathbf{1}_A\, (p \circ g^{-1})\, d\tilde{\mu} = \int_{\mathcal{X}} (\mathbf{1}_A \circ g)\, p\, d\mu = \int_{g^{-1}(A)} p\, d\mu = P(g^{-1}(A)) = (g_{\#}P)(A),$$

meaning $\tilde{p} = d(g_{\#}P)/d\tilde{\mu}$, and similarly $\tilde{q} = d(g_{\#}Q)/d\tilde{\mu}$. To simplify the argument below, we first reduce to the case $f \geq 0$ by showing that f can be replaced by a nonnegative generator without changing the divergence. To this end, for any $a \in \mathbb{R}$, define

$$h_a(t) := f(t) + a(t-1), \qquad t > 0.$$

Then h_a is convex, $h_a(1) = 0$, and its extensions satisfy

$$h_a(0) = f(0) - a, \qquad h_a'(\infty) = f'(\infty) + a.$$

Moreover, by the definition of f-divergence in (7),

$$\begin{aligned} D_{h_a}(P\|Q) &= \int_{\{q>0\}} q\Big(f(p/q) + a(p/q - 1)\Big)\, d\mu + \big(f'(\infty) + a\big)P(\{q=0\}) \\ &= D_f(P\|Q) + a\left[\int_{\{q>0\}} (p-q)\, d\mu + P(\{q=0\})\right]. \end{aligned}$$

Now, by linearity of the integral

$$\int_{\{q>0\}} (p-q)\, d\mu + P(\{q=0\}) = P(\{q>0\}) - Q(\{q>0\}) + P(\{q=0\}),$$

and since $Q(\{q=0\}) = 0$ we have $Q(\{q>0\}) = 1$. Hence

$$\int_{\{q>0\}} (p-q)\, d\mu + P(\{q=0\}) = 1 - 1 = 0,$$

and therefore

$$D_{h_a}(P\|Q) = D_f(P\|Q).$$

Now choose $\alpha \in \partial f(1)$. By convexity (Niculescu and Persson, 2018),

$$f(t) \geq f(1) + \alpha(t-1) = \alpha(t-1), \qquad t > 0,$$

so $h_{-\alpha}(t) = f(t) - \alpha(t-1) \geq 0$. Therefore, replacing f by the equivalent generator $h_{-\alpha}$ if necessary, we may assume from the outset that $f \geq 0$. With this reduction in hand, define

$$\Phi(x) := \begin{cases} \tilde{q}(x)\, f\big(\tilde{p}(x)/\tilde{q}(x)\big), & \tilde{q}(x) > 0, \\ 0, & \tilde{q}(x) = 0. \end{cases}$$

Since $f \geq 0$, Φ is nonnegative and measurable, so a change of variables yields (Tao, 2011)

$$\int_{\mathcal{X}} \Phi\, d\tilde{\mu} = \int_{\mathcal{X}} \Phi \circ g\, d\mu.$$

Using $\tilde{p} \circ g = p$ and $\tilde{q} \circ g = q$, we get $\Phi \circ g = \mathbf{1}_{\{q>0\}}\, q\, f(p/q)$. Therefore,

$$\int_{\{\tilde{q}>0\}} \tilde{q}\, f(\tilde{p}/\tilde{q})\, d\tilde{\mu} = \int_{\{q>0\}} q\, f(p/q)\, d\mu.$$

For the singular term, since $\tilde{q} = q \circ g^{-1}$, we have $g^{-1}(\{\tilde{q}=0\}) = \{q=0\}$, and hence

$$(g_{\#}P)(\{\tilde{q}=0\}) = P(g^{-1}(\{\tilde{q}=0\})) = P(\{q=0\}).$$

Combining the two identities yields

$$D_f(g_{\#}P\|g_{\#}Q) = \int_{\{q>0\}} q\, f(p/q)\, d\mu + f'(\infty)\, P(\{q=0\}) = D_f(P\|Q).$$

This proves the claim. □

We now proceed with the second part of the proof, where we isolate the role of Conditions (1) and (2), namely the $\mathcal{G}$-invariance of the target and the $\mathcal{G}$-closedness of the variational family, to show that the set of variational minimisers is $\mathcal{G}$-closed.

Lemma A.2 (Symmetry propagation) *Let $\mathcal{G} \leq \mathrm{Aut}(\mathcal{X}, \mathcal{B})$. Assume $P \in \mathcal{P}$ is $\mathcal{G}$-invariant, and let $\mathcal{Q} \subseteq \mathcal{P}$ be $\mathcal{G}$-closed. Then $\mathcal{Q}^\star$ is $\mathcal{G}$-closed.*

Proof . For any $g \in \mathcal{G}$, the objective is invariant under simultaneous pushforward by Lemma A.1. Now if P is $\mathcal{G}$-invariant, then $g_\# P = P$ and hence

$$D_f(P \| g_\# Q) = D_f(g_\# P \,\|\, g_\# Q) = D_f(P \,\|\, Q), \tag{8}$$

for all $Q \in \mathcal{Q}$. Thus the variational objective is constant along $\mathcal{G}$-orbits.

If $\mathcal{Q}^\star$ is empty, then it is trivially $\mathcal{G}$-closed. Let $Q^\star \in \mathcal{Q}^\star$ and take $g \in \mathcal{G}$. Since $\mathcal{Q}$ is $\mathcal{G}$-closed, we have $g_\# Q^\star \in \mathcal{Q}$. Moreover, by (8),

$$D_f(P \,\|\, g_\# Q^\star) = D_f(P \,\|\, Q^\star).$$

Because $Q^\star$ is a minimiser over $\mathcal{Q}$, the right-hand side equals $\inf_{Q \in \mathcal{Q}} D_f(P\|Q)$, so $g_\# Q^\star$ also attains this infimum. Thus $g_\# Q^\star \in \mathcal{Q}^\star$, proving that $\mathcal{Q}^\star$ is $\mathcal{G}$-closed.

□

We conclude the proof by finally turning to Condition (3). As shown in Section 2, this condition guarantees that for each $g \in \mathcal{G}$, the map $\rho_g \colon \mathrm{Im}(S) \to \mathrm{Im}(S)$ is well defined. We now combine this observation with the preceding lemmas to obtain our statistic recovery guarantee.

Theorem A.3 (Statistic recovery, generalised) *Let $S \colon \mathcal{P} \rightharpoonup \mathcal{Y}$ be a statistic, $P \in \mathrm{dom}(S)$ a target distribution, and $\mathcal{Q} \subseteq \mathrm{dom}(S)$ a variational family. Assume that $\mathcal{Q}^\star \neq \varnothing$ and that $S(\mathcal{Q}^\star) \coloneqq \{S(Q) \mid Q \in \mathcal{Q}^\star\}$ is a singleton. If Conditions (1)–(3) are satisfied for some group of symmetries $\mathcal{G}$, then*

$$S(P),\, S(Q^\star) \in \Gamma_S(\mathcal{G}) \qquad \textit{for every } Q^\star \in \mathcal{Q}^\star.$$

Proof . By Condition (3), for every $g \in \mathcal{G}$, the map ρ_g introduced in Section 2 is well defined on $\mathrm{Im}(S)$. Since P is $\mathcal{G}$-invariant by Condition (1), for every $g \in \mathcal{G}$ we have

$$\rho_g(S(P)) = S(g_\# P) = S(P).$$

Hence, by definition of $\Gamma_S(\mathcal{G})$,

$$S(P) \in \Gamma_S(\mathcal{G}).$$

Fix any $Q^\star \in \mathcal{Q}^\star$, which is nonempty by assumption. By Conditions (1) and (2), Lemma A.2 implies that $g_\# Q^\star \in \mathcal{Q}^\star$ for every $g \in \mathcal{G}$. Since $S(\mathcal{Q}^\star)$ is a singleton, it follows that $S(g_\# Q^\star) = S(Q^\star)$. Condition (3) therefore gives, for every $g \in \mathcal{G}$,

$$\rho_g(S(Q^\star)) = S(g_\# Q^\star) = S(Q^\star),$$

and hence

$$S(Q^\star) \in \Gamma_S(\mathcal{G}).$$

Because $Q^\star$ was arbitrary, this holds for every global minimiser. Combining the two inclusions yields

$$S(P),\, S(Q^\star) \in \Gamma_S(\mathcal{G}) \qquad \text{for every } Q^\star \in \mathcal{Q}^\star.$$

□

Theorem 2.1 follows immediately from Theorem A.3, since a unique global minimiser $Q^\star$ gives $\mathcal{Q}^\star = \{Q^\star\} \neq \varnothing$ and $S(\mathcal{Q}^\star) = \{S(Q^\star)\}$, which is a singleton.

B Additional material for Section 3

B.1 Presentation differences relative to Margossian and Saul

In this appendix, we discuss the differences between our presentation of the recovery guarantees in location–scale families in Section 3 and that of Margossian and Saul (2025), whose results we recover. The first difference is that we choose to work with probability measures rather than densities, in order to follow the general blueprint of Section 2. Accordingly, our definitions of location–scale families, as well as of even and elliptical symmetry, are the measure-theoretic analogues of those

presented in Margossian and Saul (2025). This places the results in a more general setting, but at the same time forfeits some of the convenient properties that follow from the density-level formulation. For example, for covariance recovery, in addition to the necessary assumption that the target and the base distribution of the variational family have finite second moments, we explicitly assume that both have non-zero variance — a property that automatically follows from the existence of a density with respect to Lebesgue measure. Any similar minor differences in presentation are simply artefacts of our choice to work more generally with probability distributions rather than just probability densities.

The other, and arguably more important, difference in presentation is that, whereas we state our main results under the assumption that the variational minimiser is unique, Margossian and Saul (2025) take the extra step of imposing additional regularity conditions, such as differentiability and somewhere-strict log-concavity of the target, in order to guarantee uniqueness. Because our theory focuses on the fundamental mechanism underpinning symmetry-induced statistic recovery, we are less concerned with the specific assumptions needed to ensure uniqueness in concrete settings. Accordingly, we retain uniqueness as a convenient sufficient assumption in the main text. Any regularity conditions that guarantee uniqueness of the variational minimiser can be substituted for this assumption in all our corollaries. In Appendix A.2, we further show that uniqueness can be weakened to requiring that the global minimiser set is nonempty and all global minimisers agree on the statistic of interest.

B.2 Correlation does not satisfy Condition (3)

Here, we demonstrate the problems that arise when trying to apply Theorem 2.1 directly to the correlation statistic. First, we show that the statistic does not satisfy Condition (3) because its domain is not $\mathcal{G}$-closed. We then show that even after restricting the domain to distributions with finite second moment and positive-definite covariance to guarantee $\mathcal{G}$-closedness, the correlation statistic still fails to satisfy the condition.

Recall the setting of Section 3 and, in particular, the correlation statistic $\rho\colon \mathcal{P}_2 \rightharpoonup \mathcal{C}_+^d$, defined by

$$\rho(\pi) \coloneqq D(\pi)^{-1/2}\Sigma(\pi)D(\pi)^{-1/2}, \qquad D(\pi) \coloneqq \operatorname{diag}(\Sigma(\pi)), \tag{9}$$

where $\operatorname{dom}(\rho) = \mathcal{P}_2^\circ$ is the set of distributions with finite second moment and positive marginal variances, and $\mathcal{C}_+^d$ is the set of positive semidefinite matrices with unit diagonal entries. Further, recall the relevant symmetry group:

$$\mathcal{G} \coloneqq \{g_R \mid R \in \mathrm{O}(d)\}, \qquad g_R(x) \coloneqq m + A_R(x-m), \qquad A_R \coloneqq M^{1/2}RM^{-1/2}.$$

Assume $d \geq 2$ and choose $v \in \mathbb{R}^d$ with $v_i \neq 0$ for every i. Define

$$\pi \coloneqq \frac{1}{2}\delta_v + \frac{1}{2}\delta_{-v}.$$

Then $\pi \in \mathcal{P}_2$ since it is finitely supported, and moreover $\mu(\pi) = 0$ and $\Sigma(\pi) = vv^\top$. Hence, for all $i \in \{1, \dots, d\}$, we have

$$[\Sigma(\pi)]_{ii} = v_i^2 > 0.$$

So $\pi \in \mathcal{P}_2^\circ$. Set $z \coloneqq M^{-1/2}v \neq 0$. Since $d \geq 2$, the hyperplane

$$(M^{1/2}e_1)^\perp = \{w \in \mathbb{R}^d \mid \langle w, M^{1/2}e_1\rangle = 0\}$$

has dimension $d-1 \geq 1$, and therefore intersects the sphere of radius $\|z\|$. Choose $w \in (M^{1/2}e_1)^\perp$ such that $\|w\| = \|z\|$. By transitivity of $\mathrm{O}(d)$ on the sphere, there exists $R \in \mathrm{O}(d)$ such that

$$Rz = w.$$

For this choice of R,

$$[A_R v]_1 = e_1^\top M^{1/2}RM^{-1/2}v = e_1^\top M^{1/2}w = \langle M^{1/2}e_1, w\rangle = 0.$$

Since covariance transforms under the affine map g_R by congruence with A_R,

$$\Sigma((g_R)_\# \pi) = A_R\Sigma(\pi)A_R^\top = A_R vv^\top A_R^\top = (A_R v)(A_R v)^\top.$$

Its first diagonal entry is therefore

$$[\Sigma((g_R)_\#\pi)]_{11} = [A_R v]_1^2 = 0.$$

Hence $(g_R)_\# \pi \notin \mathcal{P}_2^\circ$. This proves that $\mathcal{P}_2^\circ$ is not $\mathcal{G}$-closed, and thus ρ does not satisfy Condition (3).

Since the problem already arises with the domain, one may try to suitably restrict it to ensure $\mathcal{G}$-closedness, with the hope that the remaining part of the condition will go through. We now show this not to be the case. Let $\mathcal{P}_2^\square$ be the set of distributions with finite second moment and positive-definite covariance, and let $\mathcal{C}_{++}^d$ be the set of positive-definite matrices with unit diagonal entries. Clearly $\mathcal{P}_2^\square \subset \mathcal{P}_2^\circ$ and $\mathcal{C}_{++}^d \subset \mathcal{C}_+^d$, so we can define the correlation statistic in the same way as in (9).

Let $\pi \in \mathcal{P}_2^\square$ and $g_R \in \mathcal{G}$. Since g_R is affine, $(g_R)_\# \pi \in \mathcal{P}_2$. Moreover, by (4),

$$\Sigma((g_R)_\# \pi) = A_R \Sigma(\pi) A_R^\top.$$

Because $M \in \mathcal{S}_{++}^d$ and $R \in \mathrm{O}(d)$, A_R is invertible. Therefore, since $\Sigma(\pi) \in \mathcal{S}_{++}^d$, it follows that

$$A_R \Sigma(\pi) A_R^\top \in \mathcal{S}_{++}^d.$$

Thus $(g_R)_\# \pi \in \mathcal{P}_2^\square$, proving $\mathcal{G}$-closedness. However, the correlation statistic still does not satisfy Condition (3). It suffices to give a counterexample. We work in $\mathbb{R}^2$ and consider the spherical subcase $m = 0$, $M = I_2$, for which

$$\mathcal{G} = \{g_R \mid R \in \mathrm{O}(2)\}, \qquad g_R(x) = Rx.$$

Define

$$\pi_1 \coloneqq \mathcal{N}(0, I_2), \qquad \pi_2 \coloneqq \mathcal{N}\left(0, \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}\right).$$

Clearly, both belong to $\mathcal{P}_2^\square$, and

$$\rho(\pi_1) = I_2 = \rho(\pi_2),$$

since both covariance matrices are diagonal with positive entries. Now let

$$R \coloneqq \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} \in \mathrm{O}(2).$$

Then $\Sigma((g_R)_\# \pi_1) = I_2$, so

$$\rho((g_R)_\# \pi_1) = I_2.$$

By contrast,

$$\Sigma((g_R)_\# \pi_2) = R \begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix} R^\top = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}.$$

Therefore

$$D((g_R)_\# \pi_2) = \mathrm{diag}(\Sigma((g_R)_\# \pi_2)) = 2I_2,$$

and hence

$$\rho((g_R)_\# \pi_2) = (2I_2)^{-1/2} \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} (2I_2)^{-1/2} = \begin{pmatrix} 1 & -1/2 \\ -1/2 & 1 \end{pmatrix}.$$

Thus, there exists $g_R \in \mathcal{G}$ such that

$$\rho(\pi_1) = \rho(\pi_2) \qquad \text{but} \qquad \rho((g_R)_\# \pi_1) \neq \rho((g_R)_\# \pi_2).$$

This violates the implication required by Condition (3).

C Additional material for Section 4

C.1 Uniqueness of the axis of symmetry

In this appendix, we formally prove that if a non-uniform distribution is rotationally symmetric about two axis directions, then these determine the same line through the origin in $\mathbb{R}^d$.

Proposition C.1 (Uniqueness of the rotational axis) *Fix $d \geq 3$. Let $\pi \in \mathcal{P}_\sigma$ be a non-uniform distribution on $\mathcal{S}^{d-1}$ with density f_π, and suppose it is rotationally symmetric about both $w_1, w_2 \in \mathcal{S}^{d-1}$, i.e., there exist measurable functions $\psi, \varphi : [-1, 1] \to \mathbb{R}_{\geq 0}$ such that*

$$f_\pi(x) = \psi(w_1^\top x) = \varphi(w_2^\top x) \qquad \textit{for } \sigma\textit{-a.e. } x \in \mathcal{S}^{d-1}.$$

Then

$$\mathrm{span}(w_1) = \mathrm{span}(w_2).$$

Proof . Let $\mathscr{T}(\pi) \coloneqq \{R \in \mathrm{SO}(d) \mid (g_R)_{\#}\pi = \pi\}$ denote the subgroup of $\mathrm{SO}(d)$ under which π is invariant. $\mathscr{T}(\pi)$ is closed in $\mathrm{SO}(d)$: if $R_n \to R$ in $\mathrm{SO}(d)$ with $(g_{R_n})_{\#}\pi = \pi$ for all n, then for every $h \in C(\mathcal{S}^{d-1})$, the functions $x \mapsto h(R_n x)$ converge uniformly on $\mathcal{S}^{d-1}$ to $x \mapsto h(Rx)$ and hence

$$\int_{\mathcal{S}^{d-1}} h(x)d\pi(x) = \int_{\mathcal{S}^{d-1}} h(R_n x)d\pi(x) \longrightarrow \int_{\mathcal{S}^{d-1}} h(Rx)d\pi(x) = \int_{\mathcal{S}^{d-1}} h(x)d((g_R)_{\#}\pi)(x),$$

which guarantees that $(g_R)_{\#}\pi = \pi$ and therefore $R \in \mathscr{T}(\pi)$. Let $\mathscr{T}(\pi)^0 \leq \mathscr{T}(\pi)$ denote the identity component of $\mathscr{T}(\pi)$, which is a connected and closed subgroup. As $\mathscr{T}(\pi)$ is closed in $\mathrm{SO}(d)$, $\mathscr{T}(\pi)^0$ is also closed in $\mathrm{SO}(d)$. Letting $\mathcal{G}_{w_i} \coloneqq \{R \in \mathrm{SO}(d) \mid Rw_i = w_i\}$ for $i = 1, 2$, we know that since π is rotationally symmetric about both $w_1, w_2 \in \mathcal{S}^{d-1}$, we have $\mathcal{G}_{w_1}, \mathcal{G}_{w_2} \subseteq \mathscr{T}(\pi)$. Moreover, $\mathrm{SO}(d-1)$ is a maximal connected closed subgroup of $\mathrm{SO}(d)$ (Dynkin, 1957; Uchida, 1979), and $\mathcal{G}_{w_1}, \mathcal{G}_{w_2}$ are conjugate in $\mathrm{SO}(d)$ to the canonical embedding of $\mathrm{SO}(d-1)$. Since conjugation is a group automorphism and a topological homeomorphism, it preserves maximal connected closed subgroups, and so $\mathcal{G}_{w_1}, \mathcal{G}_{w_2}$ are themselves maximal connected closed subgroups of $\mathrm{SO}(d)$. As they are connected, and both contain the identity, it follows that $\mathcal{G}_{w_1}, \mathcal{G}_{w_2} \leq \mathscr{T}(\pi)^0$.

Assume, towards a contradiction, that $w_1 \neq \pm w_2$. Then there exists $R \in \mathcal{G}_{w_2} \setminus \mathcal{G}_{w_1}$. Indeed, otherwise every rotation fixing w_2 would also fix w_1, so w_1 would belong to the common fixed-point subspace of $\mathcal{G}_{w_2}$, namely $\mathrm{span}(w_2)$, which would force $w_1 = \pm w_2$. Since $\mathcal{G}_{w_2} \leq \mathscr{T}(\pi)^0$, it follows that $R \in \mathscr{T}(\pi)^0 \setminus \mathcal{G}_{w_1}$. Hence $\mathcal{G}_{w_1}$ is a proper subgroup of $\mathscr{T}(\pi)^0$. But $\mathcal{G}_{w_1}$ is a maximal connected closed subgroup of $\mathrm{SO}(d)$ and so it must be that $\mathscr{T}(\pi)^0 = \mathrm{SO}(d)$. It follows that $\mathscr{T}(\pi) = \mathrm{SO}(d)$. Therefore π is invariant under every rotation of $\mathcal{S}^{d-1}$. The only rotation-invariant probability measure on $\mathcal{S}^{d-1}$ is the uniform measure, a contradiction. So $\mathrm{span}(w_1) = \mathrm{span}(w_2)$. □

C.2 Explicit construction of the unique minimiser example

Here, we justify the claim made at the end of Section 4 for the reverse KL divergence objective. In particular, we show that there exists a threshold $\eta_c(d, \lambda, \kappa_0) > 0$ such that the KL has a unique minimiser for $0 < \eta \leq \eta_c(d, \lambda, \kappa_0)$, whereas for $\eta > \eta_c(d, \lambda, \kappa_0)$ the minimiser set is non-singleton.

Fix $\lambda \neq 0$, $u \in \mathcal{S}^{d-1}$, $\kappa_0 > 0$, and $\eta > 0$. For $\nu \in \mathcal{S}^{d-1}$, let $X \sim Q_{\nu,\kappa_0}$, and write $\mathscr{L}_{\lambda,\eta,\kappa_0}(\nu)$ for the reverse KL objective $D_{\mathrm{KL}}(Q_{\nu,\kappa_0} \| P_{\lambda,\eta})$. Using

$$\log f_{Q_{\nu,\kappa_0}}(x) = \log c_d(\kappa_0) + \kappa_0 \nu^\top x$$

and

$$\log f_{P_{\lambda,\eta}}(x) = -\log Z_{\lambda,\eta} + \lambda u^\top x - \eta(u^\top x)^2,$$

we obtain

$$\begin{aligned}\mathscr{L}_{\lambda,\eta,\kappa_0}(\nu) &= \mathbb{E}\left[\log \frac{f_{Q_{\nu,\kappa_0}}(X)}{f_{P_{\lambda,\eta}}(X)}\right] \\ &= \log c_d(\kappa_0) + \log Z_{\lambda,\eta} + \kappa_0 \underbrace{\mathbb{E}[\nu^\top X]}_{(E_1)} - \lambda \underbrace{\mathbb{E}[u^\top X]}_{(E_2)} + \eta \underbrace{\mathbb{E}[(u^\top X)^2]}_{(E_3)}. \end{aligned} \tag{10}$$

We begin by reducing each expectation in (10) to a function of the single scalar $c \coloneqq u^\top \nu \in [-1, 1]$. Set $T \coloneqq \nu^\top X$. With $\rho_d(t) \coloneqq (1-t^2)^{(d-3)/2}$, the density of T is (Mardia and Jupp, 2009, Eq. 9.3.12)

$$f_T(t) \propto e^{\kappa_0 t}\rho_d(t), \qquad t \in [-1, 1]. \tag{11}$$

Now define

$$A_{d,\kappa_0} \coloneqq \mathbb{E}[T], \qquad m_{d,\kappa_0} \coloneqq \mathbb{E}[T^2].$$

Then the mean and variance of X are given by (Mardia and Jupp, 2009, Eqs. 9.3.33–9.3.34)

$$\mathbb{E}[X] = A_{d,\kappa_0}\nu, \qquad \mathrm{Var}(X) = \mathrm{Var}(T)\,\nu\nu^\top + \frac{1 - m_{d,\kappa_0}}{d-1}(I_d - \nu\nu^\top).$$

Therefore, the expectations (E_1) and (E_2) in (10) are

$$\begin{aligned}\mathbb{E}[\nu^\top X] &= A_{d,\kappa_0}, \\ \mathbb{E}[u^\top X] &= u^\top \mathbb{E}[X] = A_{d,\kappa_0}\, u^\top \nu = A_{d,\kappa_0}\, c. \end{aligned} \tag{12}$$

Moreover, the second moment matrix is

$$\begin{aligned}\mathbb{E}[XX^\top] &= \mathrm{Var}(X) + \mathbb{E}[X]\mathbb{E}[X]^\top \\ &= \big(\mathrm{Var}(T) + (\mathbb{E}[T])^2\big)\nu\nu^\top + \frac{1-\mathbb{E}[T^2]}{d-1}(I_d - \nu\nu^\top) \\ &= m_{d,\kappa_0}\,\nu\nu^\top + \frac{1-m_{d,\kappa_0}}{d-1}(I_d - \nu\nu^\top),\end{aligned}$$

and hence, expectation (E_3) in (10) reduces to

$$\begin{aligned}\mathbb{E}[(u^\top X)^2] &= u^\top \mathbb{E}[XX^\top]u \\ &= m_{d,\kappa_0}(u^\top\nu)^2 + \frac{1-m_{d,\kappa_0}}{d-1}\big(1-(u^\top\nu)^2\big) \\ &= \frac{1-m_{d,\kappa_0}}{d-1} + \underbrace{\frac{d\,m_{d,\kappa_0}-1}{d-1}}_{B_{d,\kappa_0}} c^2. \end{aligned} \tag{13}$$

Substituting (12) and (13) into (10) yields

$$\mathscr{L}_{\lambda,\eta,\kappa_0}(\nu) = \eta B_{d,\kappa_0}\, c^2 - \lambda A_{d,\kappa_0}\, c + C_{\lambda,\eta,\kappa_0},$$

where

$$C_{\lambda,\eta,\kappa_0} := \log c_d(\kappa_0) + \log Z_{\lambda,\eta} + \kappa_0 A_{d,\kappa_0} + \eta\frac{1-m_{d,\kappa_0}}{d-1},$$

is a constant independent of ν. We have thus reduced the optimisation problem to the scalar quadratic

$$\widehat{\mathscr{L}}_{\lambda,\eta,\kappa_0}(c) := \eta B_{d,\kappa_0} c^2 - \lambda A_{d,\kappa_0} c, \qquad c \in [-1,1]. \tag{14}$$

Since $\eta > 0$, once $B_{d,\kappa_0} > 0$ is established, the quadratic in (14) is strictly convex and therefore has a unique unconstrained minimiser. Because $d \geq 3$, the denominator of B_{d,κ_0} is positive, so it suffices to show that

$$d\,m_{d,\kappa_0} - 1 > 0,$$

that is, $m_{d,\kappa_0} > 1/d$. To this end, let $Y \sim \sigma$ be uniform on $\mathcal{S}^{d-1}$ and define $T_0 := u^\top Y$. Then the density of T_0 is (Mardia and Jupp, 2009, Eq. 9.3.1)

$$f_{T_0}(t) \propto \rho_d(t), \qquad t \in [-1,1]. \tag{15}$$

We use (15) to lower bound m_{d,κ_0}. Expanding the definition, observe that $t^2\rho_d(t)$ is even while $e^{\kappa_0 t} = \cosh(\kappa_0 t) + \sinh(\kappa_0 t)$, so the odd terms integrate to 0 on $[-1,1]$, and we obtain

$$m_{d,\kappa_0} = \frac{\displaystyle\int_{-1}^{1} t^2 e^{\kappa_0 t}\rho_d(t)\,dt}{\displaystyle\int_{-1}^{1} e^{\kappa_0 t}\rho_d(t)\,dt} = \frac{\displaystyle\int_{0}^{1} t^2\cosh(\kappa_0 t)\rho_d(t)\,dt}{\displaystyle\int_{0}^{1} \cosh(\kappa_0 t)\rho_d(t)\,dt} > \frac{\displaystyle\int_{0}^{1} t^2\rho_d(t)\,dt}{\displaystyle\int_{0}^{1} \rho_d(t)\,dt} = \mathbb{E}[T_0^2], \tag{16}$$

where the strict inequality follows from Chebyshev's integral inequality (Mitrinović, 1970, Thm. 10, Ch. 2.5), since $t \mapsto t^2$ and $t \mapsto \cosh(\kappa_0 t)$ are strictly increasing on $[0,1]$. To compute $\mathbb{E}[T_0^2]$, note that since $Y \sim \sigma$ is uniform on $\mathcal{S}^{d-1}$, its law is rotation invariant and therefore $\mathbb{E}[YY^\top] = \alpha I_d$ for some $\alpha \in \mathbb{R}$. Taking traces,

$$\mathrm{tr}(\mathbb{E}[YY^\top]) = \mathbb{E}[\mathrm{tr}(YY^\top)] = \mathbb{E}[\|Y\|^2] = 1, \qquad \mathrm{tr}(\alpha I_d) = \alpha d,$$

and hence $\alpha = 1/d$. The second moment of T_0 is thus

$$\mathbb{E}[T_0^2] = u^\top \mathbb{E}[YY^\top]u = u^\top\left(\frac{1}{d}I_d\right)u = \frac{1}{d}\|u\|^2 = \frac{1}{d}.$$

Combining this with (16) gives $m_{d,\kappa_0} > 1/d$, and therefore $B_{d,\kappa_0} > 0$. Consequently, the scalar objective in (14) has the unique unconstrained minimiser

$$c^\star = \frac{\lambda A_{d,\kappa_0}}{2\eta B_{d,\kappa_0}}.$$

Define

$$\eta_c(d, \lambda, \kappa_0) \coloneqq \frac{|\lambda| A_{d,\kappa_0}}{2B_{d,\kappa_0}}.$$

Then, $\eta_c(d, \lambda, \kappa_0) > 0$ since $|\lambda|, B_{d,\kappa_0} > 0$ and moreover $A_{d,\kappa_0} > 0$. Indeed, since $\rho_d(t)$ is even,

$$A_{d,\kappa_0} = \frac{\displaystyle\int_{-1}^{1} t e^{\kappa_0 t} \rho_d(t)\, dt}{\displaystyle\int_{-1}^{1} e^{\kappa_0 t} \rho_d(t)\, dt} = \frac{\displaystyle\int_{0}^{1} t \left(e^{\kappa_0 t} - e^{-\kappa_0 t}\right) \rho_d(t)\, dt}{\displaystyle\int_{-1}^{1} e^{\kappa_0 t} \rho_d(t)\, dt} > 0.$$

If $0 < \eta \le \eta_c(d, \lambda, \kappa_0)$, then $|c^\star| \ge 1$, and so the unique constrained minimiser of (14) over $[-1, 1]$ is $c = \operatorname{sgn}(\lambda)$. Since $u^\top \nu = \operatorname{sgn}(\lambda)$ if and only if $\nu = \operatorname{sgn}(\lambda) u$, the original objective $\mathscr{L}_{\lambda,\eta,\kappa_0}(\nu)$ has the unique minimiser $\nu^\star = \operatorname{sgn}(\lambda) u$. Consequently, exact statistic recovery holds with

$$\mathcal{A}(Q_{\nu^\star,\kappa_0}) = \operatorname{span}(u) = \mathcal{A}(P_{\lambda,\eta}).$$

If $\eta > \eta_c(d, \lambda, \kappa_0)$, then $c^\star \in (-1, 1)$. Hence

$$\operatorname*{argmin}_{\nu \in \mathcal{S}^{d-1}} \mathscr{L}_{\lambda,\eta,\kappa_0}(\nu) = \{\nu \in \mathcal{S}^{d-1} \mid u^\top \nu = c^\star\}. \tag{17}$$

Since $d \ge 3$ and $c^\star \in (-1, 1)$, the set in (17) is a $(d-2)$-sphere and hence is not a singleton. Moreover, no minimiser ν satisfies $\mathcal{A}(Q_{\nu,\kappa_0}) = \mathcal{A}(P_{\lambda,\eta})$, because $\mathcal{A}(Q_{\nu,\kappa_0}) = \operatorname{span}(\nu)$, while $\operatorname{span}(\nu) = \operatorname{span}(u)$ would force $\nu = \pm u$ and hence $u^\top \nu \in \{\pm 1\}$, contradicting $c^\star \in (-1, 1)$.

To visualise the axis recovery, we work with $d = 3$. For fixed $\lambda \neq 0$ and $\kappa_0 > 0$, we can explicitly compute the constants A_{3,κ_0} and B_{3,κ_0}, and evaluate $\eta_c(3, \lambda, \kappa_0)$. First observe that since $\rho_3(t) = 1$ for all $t \in [-1, 1]$, the marginal density of $T = \nu^\top X$ in dimension 3 is proportional to $e^{\kappa_0 t}$. Writing

$$Z(\kappa_0) \coloneqq \int_{-1}^{1} e^{\kappa_0 t}\, dt = \frac{2 \sinh \kappa_0}{\kappa_0},$$

the first two moments of T can be obtained by differentiating the log-partition function:

$$A_{3,\kappa_0} = \mathbb{E}[T] = \frac{d}{d\kappa_0} \log Z(\kappa_0), \qquad m_{3,\kappa_0} = \mathbb{E}[T^2] = \frac{d^2}{d\kappa_0^2} \log Z(\kappa_0) + \left(\frac{d}{d\kappa_0} \log Z(\kappa_0)\right)^2.$$

Since $\log Z(\kappa_0) = \log 2 + \log \sinh \kappa_0 - \log \kappa_0$, this yields

$$A_{3,\kappa_0} = \coth \kappa_0 - \frac{1}{\kappa_0}, \qquad m_{3,\kappa_0} = \mathbb{E}[T^2] = 1 - \frac{2}{\kappa_0} \coth \kappa_0 + \frac{2}{\kappa_0^2}.$$

Substituting m_{3,κ_0} into the definition of B_{3,κ_0} gives

$$B_{3,\kappa_0} = 1 - \frac{3}{\kappa_0} \coth \kappa_0 + \frac{3}{\kappa_0^2}$$

and the critical threshold therefore becomes

$$\eta_c(3, \lambda, \kappa_0) = \frac{|\lambda| \kappa_0 (\kappa_0 \coth \kappa_0 - 1)}{2(\kappa_0^2 - 3\kappa_0 \coth \kappa_0 + 3)}.$$

In the example of Figure 2, we take $\lambda = 1$ and $\kappa_0 = 2.5$, so

$$A_{3,2.5} \approx 0.6135, \qquad B_{3,2.5} \approx 0.2637,$$
$$\eta_c(3, 1, 2.5) \approx 1.1632.$$

In the left panel, we use $\eta = 1 < \eta_c(3, 1, 2.5)$, and the unique minimiser $\nu^\star$ of $\mathscr{L}_{\lambda,\eta,\kappa_0}(\nu)$ satisfies

$$\mathcal{A}(Q_{\nu^\star,2.5}) = \mathcal{A}(P_{1,1}) = \operatorname{span}(u),$$

while in the right panel, we use $\eta = 2 > \eta_c(3, 1, 2.5)$, so the minimiser set is the latitude circle

$$\left\{\nu \in \mathcal{S}^2 \;\middle|\; u^\top \nu = \frac{A_{3,2.5}}{4B_{3,2.5}} \approx 0.5816\right\},$$

and $\mathcal{A}(Q_{\nu,2.5}) \neq \mathcal{A}(P_{1,2})$ for all minimisers ν.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims in the abstract and introduction are reflected in the paper's body: Section 2 gives the general statistic recovery theory, Section 3 recovers known location-scale guarantees as special cases, and Section 4 develops the new rotational symmetry guarantees.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper addresses limitations by discussing settings in which the assumptions of the main theorem are not met. Section 4 and Appendix C.2 give a concrete failure mode, visualised in Figure 2. Section 3 and Appendix B.2 further clarify the scope of the result by showing that the theorem's conditions are sufficient but not always necessary.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All mathematical results are presented through explicit definitions, theorems, and corollaries, with assumptions stated clearly and proofs provided either in the main text or in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The details required to reproduce the numerical results supporting the paper's main claims are provided in the main text and appendices, together with the code.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper provides a link to the code used to reproduce all figures. The repository includes all the necessary instructions and dependencies.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Sufficient details to understand the numerical results are provided in Sections 3 and 4 and Appendices B and C, together with the code.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: Our numerics do not involve statistical estimates over repeated trials, so error bars and statistical significance tests are not applicable.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The compute resources are described in the main text.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This is theoretical work on variational inference, and we do not identify any direct societal impacts beyond those more broadly associated with approximate inference.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper does not pose such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.