# EigenVI: score-based variational inference with orthogonal function expansions

**Diana Cai** [1] **Chirag Modi** [1 2] **Charles C. Margossian** [1] **Robert M. Gower** [1] **David M. Blei** [3] **Lawrence K. Saul** [1]

## Abstract

We develop EigenVI, a new approach for black-box variational inference (BBVI). EigenVI fits a novel class variational approximations based on orthogonal function expansions. For distributions over $\mathbb{R}^D$, the lowest order term in these expansions provides a Gaussian variational approximation, while higher-order terms provide a systematic way to model non-Gaussianity. These variational approximations are flexible enough to model complex distributions (multimodal, asymmetric), but they are simple enough that one can calculate their low-order moments and draw samples from them. Further, by choosing different families of orthogonal functions, EigenVI can model different types of random variables (e.g., real-valued, nonnegative, bounded). To fit the approximation, EigenVI matches score functions by minimizing a Fisher divergence. Notably, this optimization reduces to solving a minimum eigenvalue problem, so that EigenVI effectively sidesteps the iterative gradient-based optimizations that are required for many other BBVI algorithms. (Gradient-based methods can be sensitive to learning rates, termination criteria, and other tunable hyperparameters.) We study EigenVI on a variety of target distributions, including a benchmark suite of Bayesian models from `posteriordb`. Compared to existing methods for BBVI, EigenVI is more accurate.

## 1. Introduction

Probabilistic modeling is a cornerstone of modern data analysis, uncertainty quantification, and decision making. A key challenge of probabilistic inference is computing a target distribution of interest; for instance, in Bayesian modeling, the goal is to compute a posterior distribution, which is often intractable. Variational inference (VI) (Jordan et al., 1999; Wainwright et al., 2008; Blei et al., 2017) is a popular method that has enabled scalable probabilistic inference across a range of applications. The idea behind VI is to target distribution with the closest member of a tractable family.

One major focus of variational inference is black-box variational inference (BBVI) algorithms (Ranganath et al., 2014; Kingma and Welling, 2014; Titsias and Lázaro-Gredilla, 2014; Kucukelbir et al., 2017; Locatello et al., 2018; Giordano et al., 2024; Wang et al., 2024; Modi et al., 2023; Cai et al., 2024). In BBVI, the target needs only to be available as the log of an unnormalized distribution, which is also typically assumed to be differentiable. Because it is so easily applicable, BBVI algorithms are now widely available in popular probabilistic programming languages, providing automated VI algorithms to data analysis practitioners (Salvatier et al., 2016; Carpenter et al., 2017; Ge et al., 2018; Bingham et al., 2019; Abril-Pla et al., 2023).

One thread of BBVI research focuses on Gaussian variational approximations (Ranganath et al., 2014; Kucukelbir et al., 2017). Traditionally, the approximation is optimized by minimizing the Kullback-Leibler (KL) divergence between the variational family and the target (equivalently, maximizing the ELBO). This strategy is powerful and scalable, but it relies on stochastic gradient descent, which can be difficult to tune (Dhaka et al., 2020; 2021; Zhang et al., 2022).

More recently, researchers have proposed Gaussian BBVI algorithms that do not require the use of SGD (Modi et al., 2023; Cai et al., 2024). These methods aim to match the *scores*, or the gradients of the log densities, between the variational distribution and the target density. Thanks to the Gaussian family, these algorithms implement closed-form proximal point updates for solving the score matching problem. The resulting methods are as inexpensive as SGD, but are not as brittle.

In this paper, we develop a new way to perform score-based BBVI. We propose a new class of variational families built from *orthogonal function expansions*, inspired by wave functions from quantum mechanics. This class is both expres-

[1]Center for Computational Mathematics, Flatiron Institute [2]Center for Computational Astrophysics, Flatiron Institute [3]Department of Statistics, Department of Computer Science, Columbia University. Correspondence to: Diana Cai <dcai@flatironinstitute.org>.

*Accepted by the Structured Probabilistic Inference & Generative Modeling workshop* of ICML 2024, Vienna, Austria. 

sive enough to capture a variety of target distributions, and tractable enough to be able to sample from and calculate low-order moments. For distributions supported on $\mathbb{R}^D$, a first-order expansion recovers the Gaussian distribution and higher-order expansions allow for increasing amounts of non-Gaussianity. Depending on the choice of basis set, this construction can also produce variational families over other spaces.

To optimize over a variational family from this class, we minimize an estimate of the Fisher divergence, which measures the scores of the variational distribution against those of the target distribution. We show that the optimization objective is equivalent to solving an eigenvalue problem, thus avoiding the need for gradient-based optimization. For this reason, we call our approach *EigenVI*.

We study EigenVI with a variational family constructed from weighted Hermite polynomials. We first demonstrate its expressiveness on a variety of complex target distributions, such as multimodal, asymmetric, and heavy-tailed target distributions. We then study a suite of non-Gaussian target distributions from `posteriordb` (Magnusson et al., 2022), a benchmark suite of Bayesian hierarchical models. Compared to leading implementations of Gaussian BBVI based on KL minimization and score matching, EigenVI provides more accurate posterior approximations.

In Section 2 we introduce the orthonormal family, a new variational family built from orthogonal function expansions, and we show how score matching with this family is equivalent to an eigenvalue problem. In Section 3, we evaluate EigenVI on a variety of synthetic and real-data targets. In Section 4, we discuss limitations and future work.

## 2. Score-based variational inference with the orthonormal family

In this section we develop a variational family for approximate probabilistic inference based on orthogonal function expansions. In Section 2.1, we show that the approximations in this family are expressive enough to model complex distributions, but also that they are simple enough to calculate their moments and draw samples from them. In Section 2.2, we show how to optimize these variational approximations by minimizing a score-based divergence; notably, for this divergence, the optimization reduces to an eigenvalue problem. Finally in Section 2.3, we consider how to use these variational approximations for unstandardized distributions; in these settings we must carefully manage the trade-off between expressiveness and computational cost.

### 2.1. Orthogonal function expansions

Let $\mathcal{Z} \subseteq \mathbb{R}^D$ denote the support of the target distribution $p$. Suppose that there exists a complete set of orthonormal basis functions $\{\phi_k(z)\}_{k=1}^{\infty}$ on this set. By *complete*, we mean that any sufficiently well-behaved function $f : \mathcal{Z} \to \mathbb{R}$ can be approximated, to arbitrary accuracy, by a particular weighted sum of these basis functions, and by *orthonormal*, we mean that the basis functions satisfy

$$\int \phi_k(z)\phi_{k'}(z)\, dz = \begin{cases} 1 & \text{if } k = k', \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where the integral is over $\mathcal{Z}$. We define the $K^{\text{th}}$-order variational family $\mathcal{Q}_K$ to be the set containing all distributions of the form

$$q(z) = \left(\sum_{k=1}^{K} \alpha_k \phi_k(z)\right)^2 \quad \text{where} \quad \sum_{k=1}^{K} \alpha_k^2 = 1, \tag{2}$$

where $\alpha_k \in \mathcal{R}$ for $k = 1, \ldots, K$ are the parameters of the family $\mathcal{Q}_K$. In words, $\mathcal{Q}_K$ contains all distributions that can be obtained by taking weighted sums of the first $K$ basis functions and then *squaring* the result.

Eq. 2 involves a squaring operation, a sum-of-squares constraint, and a weighted sum. The squaring operation ensures that the density functions in $\mathcal{Q}_K$ are nonnegative (i.e., with $q(z) \geq 0$ for all $z \in \mathcal{Z}$), while the sum-of-squares constraint ensures that they are normalized:

$$\begin{aligned} \int q(z)\, dz &= \int \left(\sum_{k=1}^{K} \alpha_k \phi_k(z)\right)^2 dz \\ &= \int \sum_{k,k'=1}^{K} \alpha_k \alpha_{k'} \phi_k(z)\phi_{k'}(z)\, dz \\ &= \sum_{k=1}^{K} \alpha_k^2 = 1. \end{aligned} \tag{3}$$

The weighted sum in Eq. 2 bears a superficial similarity to a mixture model, but we emphasize that neither the basis functions $\phi_k(z)$ nor the weights $\alpha_k$ in Eq. 2 are constrained to be nonnegative. Distributions of this form arise naturally in physics from the quantum-mechanical *wave functions* that satisfy Schrödinger's equation (Griffiths and Schroeter, 2018).

The simplest examples of orthogonal function expansions arise in one dimension. For example, functions on the interval $[-1, 1]$ can be represented as weighted sums of Legendre polynomials, while functions on the unit circle can be represented by Fourier series of sines and cosines; see Table 1. Distributions on unbounded intervals can also be represented in this way. On the real line, for example, we may consider approximations of the form in Eq. 2 where

$$\phi_{k+1}(z) = \left(\sqrt{2\pi}k!\right)^{-\frac{1}{2}} \left(e^{-\frac{1}{2}z^2}\right)^{\frac{1}{2}} \mathrm{H}_k(z), \tag{4}$$

and $\mathrm{H}_k(z)$ are the *probabilist's* Hermite polynomials given by

$$\mathrm{H}_k(z) = (-1)^k e^{\frac{z^2}{2}} \frac{d^k}{dz^k} \left[ e^{-\frac{z^2}{2}} \right]. \tag{5}$$

Note how the lowest-order basis function $\phi_1(z)$ in this family gives rise (upon squaring) to a Gaussian distribution with zero mean and unit variance.

Figure 1 shows how various multimodal distributions with one-dimensional support can be approximated by computing weighted sums of basis functions and squaring their result. We emphasize that *the more basis functions in the sum, the better the approximation*.

Orthogonal function expansions in one dimension are also important because their Cartesian products can be used to generate orthogonal function expansions in higher dimensions. For example, we can approximate distributions over (say) $\mathbb{R}^3$ by

$$q(z_1, z_2, z_3) = \left( \sum_{i=1}^{K_1} \sum_{j=1}^{K_2} \sum_{k=1}^{K_3} \beta_{ijk}\, \phi_i(z_1)\phi_j(z_2)\phi_k(z_3) \right)^2 \quad \text{where} \quad \sum_{ijk} \beta_{ijk}^2 = 1, \tag{6}$$

where $\beta_{i,j,k} \in \mathcal{R}$ now parametrize the family. Note that there are a total $K_1 K_2 K_3$ parameters in the above expansion, so that this method of Cartesian products does not scale well to high dimensions if multiple basis functions are used per dimension. Note that the same strategy can also be used for random variables of mixed type: for example, from Table 1, we can create a variational family of distributions over $\mathbb{R} \times [-1, 1] \times [0, \infty)$ from the Cartesian product of orthogonal function expansions involving Hermite, Legendre, and Laguerre polynomials.

As shown in Figure 1, the approximating distributions from $K^{\text{th}}$-order expansions can model the presence of multiple modes as well as many types of asymmetry, and this expressiveness also extends to higher dimensions. Nevertheless, it remains tractable to sample from these distributions and even to calculate (analytically) their low-order moments, as we show in Appendices B and C.

For concreteness, consider the distribution over $\mathbb{R}^3$ in Eq. 6. The marginal distribution $q(z_1)$ is

$$\begin{aligned} q(z_1) &= \int q(z_1, z_2, z_3)\, dz_2\, dz_3 \\ &= \sum_{ii'} \left[ \sum_{jk} \beta_{ijk}\beta_{i'jk} \right] \phi_i(z_1)\phi_{i'}(z_1), \end{aligned} \tag{7}$$

and from this expression, moments such as $\mathbb{E}[z_1]$ and $\mathrm{Var}[z_1]$ can be calculated by evaluating integrals involving the elementary functions in Table 1. (In practice, these integrals are further simplified by recursion relations that relate basis functions of different orders.) We can sample $z_1 \sim q(z_1)$ by computing the cumulative distribution function (CDF) of this marginal distribution and then numerically inverting this CDF. Finally, extending these ideas, we can calculate higher-order moments and obtain joint samples via the nested draws

$$z_1 \sim q(z_1), \quad z_2 \sim q(z_2|z_1), \quad z_3 \sim q(z_3|z_1, z_2). \tag{8}$$

The overall complexity of these procedures scales no worse than quadratically in the number of basis functions in the expansion. These extensions are discussed further in Appendices B and C.

### 2.2. EigenVI

In variational inference, we posit a variational family of distributions and then fit the free variational parameters to find the member that is close to a target distribution of interest. Consider a target density $p(z)$, which is intractable to compute. Eq. 2 defines a variational family $\mathcal{Q}_K$ based on orthonormal functions, and where the free variational parameters are the weights $\{\alpha_k\}_{k=1}^K$. We now derive *EigenVI*, a method to find $q \in \mathcal{Q}_K$ that is close to $p(z)$.

We first define the measure of closeness that we will minimize. EigenVI measures the quality of an approximate density by the *Fisher divergence*,

$$\mathscr{D}(q, p) = \int \|\nabla \log q - \nabla \log p\|^2 \, dq, \tag{9}$$

where $\nabla \log q$ and $\nabla \log p$ are the score functions of the variational approximation and target, respectively. The Fisher divergence vanishes if and only if the scores of $q$ and $p$ are everywhere equal.

Though $p$ is, by assumption, intractable, in many applications it is possible to efficiently compute the score $\nabla \log p$ at any point $z \in \mathcal{Z}$. For example, in Bayesian models the score of the target posterior is equal to the gradient of the log joint. This is the main motivation for score-based methods in probabilistic modeling (Liu and Wang, 2016; Yu and Zhang, 2023; Modi et al., 2023; Cai et al., 2024).

Here we seek the $q \in \mathcal{Q}_K$ that minimizes $\mathscr{D}(q, p)$. But a challenge arises: it is generally difficult to evaluate the integral for $\mathscr{D}(q, p)$ in Eq. 9, let alone to minimize it as a function of $q$. To proceed, we construct an unbiased estimator of $\mathscr{D}(q, p)$ by importance sampling. Let $\{z^1, z^2, \ldots z^B\}$ denote a batch of $B$ samples drawn from some proposal distribution $\pi$ on $\mathcal{Z}$. From these samples we can form the unbiased estimator

$$\widehat{\mathscr{D}}_\pi(q, p) = \sum_{b=1}^{B} \frac{q(z^b)}{\pi(z^b)} \left\| \nabla \log q(z^b) - \nabla \log p(z^b) \right\|^2. \tag{10}$$

Table 1: Examples of orthogonal function expansions in one dimension. The basis functions in the table are not normalized, but they can be rescaled so that their squares integrate to one.

| **support** | **orthogonal family** | **basis functions** $\phi_k(\cdot)$ |
|---|---|---|
| $z \in [-1, 1]$ | Legendre polynomials | $\{1,\ z,\ 3z^2-1,\ 5z^3-3z, \ldots\}$ |
| $z = e^{i\theta} \in S^1$ | Fourier basis | $\{1, \cos\theta, \sin\theta, \cos 2\theta, \sin 2\theta, \ldots\}$ |
| $z \in [0, \infty)$ | weighted Laguerre polynomials | $e^{-\frac{z}{2}}\{1, 1-z, z^2-4z+2, \ldots\}$ |
| $z \in \mathbb{R}$ | weighted Hermite polynomials | $e^{-\frac{z^2}{4}}\{1, z, (z^2-1), (z^3-3z), \ldots\}$ |

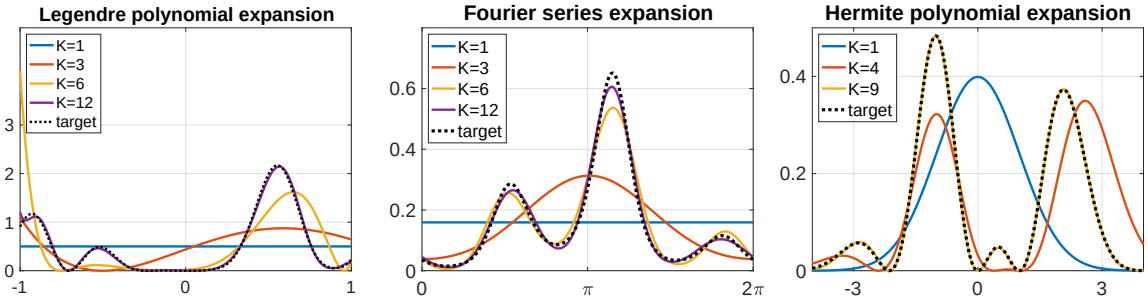



Figure 1: Target probability distributions (in black) on the interval $[-1, 1]$ (left), the unit circle (middle), and the real line (right), and their approximations by orthogonal function expansions from different families and of different orders; see Eq. 2 and Table 1.

This estimator should be accurate for appropriately broad proposal distributions and for sufficiently large batch sizes. We can therefore attempt to minimize Eq. 10 in place of Eq. 9.

Now we show that minimizing Eq. 10 over $q \in \mathcal{Q}_K$ simplifies to an eigenvalue problem for the weights $\{\alpha_k\}_{k=1}^K$ in Eq. 2. We note that this simplification arises from the particular choice of variational family (based on orthogonal function expansions) and the particular choice of divergence (based on score-matching). This eigenvalue problem stands in contrast to the gradient-based optimizations—involving learning rates, terminating criteria, and perhaps other algorithmic hyperparameters—that are typically required for ELBO-based BBVI (Dhaka et al., 2020; 2021).

To obtain the eigenvalue problem, we substitute the orthogonal function expansion in Eq. 2 into Eq. 10 for the unbiased estimator of $\mathscr{D}(q, p)$. As an intermediate step, we differentiate Eq. 2 to obtain the scores

$$\nabla \log q(z^b) = \frac{2\sum_k \alpha_k \nabla\phi_k(z^b)}{\sum_k \alpha_k \phi_k(z^b)}. \tag{11}$$

Further substitution of the scores provides the key result behind our approach: the unbiased estimator in Eq. 10 is a simple quadratic form in the weights $\alpha = [\alpha_1, \ldots, \alpha_K]^\top$ of the orthogonal function expansion,

$$\widehat{\mathscr{D}}_\pi(q, p) = \alpha^\top M \alpha, \tag{12}$$

where the coefficients of the quadratic form are given by

$$M_{jk} = \sum_{b=1}^{B} \frac{1}{\pi(z^b)} [2\nabla\phi_j(z^b) - \phi_j(z^b)\nabla \log p(z^b)] \cdot [2\nabla\phi_k(z^b) - \phi_k(z^b)\nabla \log p(z^b)]. \tag{13}$$

Note the elements of the $K \times K$ symmetric matrix $M$ capture all of the dependence on the batch of samples, the scores of $p$ and $q$ at these samples, and the choice of the family of orthogonal functions. Next we minimize the quadratic form in Eq. 12 subject to the sum-of-squares constraint $\sum_k \alpha_k^2 = 1$ in Eq. 2. In this way (see Appendix D for a proof) we obtain the eigenvalue problem (Courant and Hilbert, 1924)

$$\min_{q \in \mathcal{Q}_K} \left[\widehat{\mathscr{D}}(q, p)\right] = \min_{\|\alpha\|=1} \left[\alpha^\top M \alpha\right] =: \lambda_{\min}(M), \tag{14}$$

where $\lambda_{\min}(M)$ is the minimal eigenvalue of $M$, and the optimal weights are given (up to an arbitrary sign) by its corresponding eigenvector. EigenVI solves Eq. 14.

Computationally, the size of the eigenvalue problem is equal to the number of basis functions $K$ in the orthogonal expansion. The eigenvalue problem also generalizes to orthogonal function expansions that are formed from Cartesian products of one-dimensional families, but in this case, if multiple basis functions are used per dimension, then the overall basis size grows exponentially in the dimensionality. Thus, for example, the eigenvalue problem would be of size $K_1 K_2 K_3$ for the approximation in Eq. 6, as can be seen by "flattening"

the tensor of weights $\beta$ in Eq. 6 into the vector of weights $\alpha = \mathbf{vec}(\beta)$ in Eq. 2. Finally, we note that only the minimal eigenvector of $M$ needs to be computed, which can be much less expensive than computing a full diagonalization.

### 2.3. Standardization in $\mathbb{R}^D$

In the previous section, we discussed that the size of the eigenvalue problem grows linearly with the number of basis functions. In practice one should therefore use the least number of basis functions that are required for a suitable approximation. For target distributions on $\mathbb{R}^D$, this number can be reduced by simple linear transformations of the domain—transformations which standardize the random variables so that they have approximately zero mean and unit variance. Consider, for instance, the special case that the target distribution $p(z)$ is in fact multivariate Gaussian, with mean $\mu$ and covariance $\Sigma$. In this case, the standardized variable

$$\tilde{z} = \Sigma^{-\frac{1}{2}}(z-\mu) \tag{15}$$

will have zero mean and identity covariance, and the induced distribution $\tilde{p}(\tilde{z})$ will be perfectly fitted by a product of single basis functions in the orthogonal family of reweighted Hermite polynomials. More generally, when $p$ is not Gaussian, EigenVI can use a low-order expansion to approximate $\tilde{p}(\tilde{z})$ and then undo the change-of-variables to obtain the approximation $p(z) = \tilde{p}(\tilde{z})|\Sigma|^{-1/2}$.

Figure 2 shows why it is more difficult to approximate distributions that are badly centered or poorly scaled. The left panel shows the effect of translating a standard Gaussian *away* from the origin and *shrinking* its variance; note how a comparable approximation to the uncentered Gaussian now requires a 16th-order expansion. The right panel shows the similar effect of translating the mixture distribution in Figure 1 (right panel); comparing these panels, we see that twice as many basis functions ($K = 12$ versus $K = 6$) are required to provide a comparable fit of the uncentered mixture.

We emphasize that it is *not* necessary to know the *exact* mean and covariance of a target distribution over $\mathbb{R}^d$ for EigenVI to benefit from these types of standardizing transformations. Sometimes domain-specific knowledge may be enough to provide rough estimates of these quantities; if not, they might alternatively be estimated by a Laplace or Gaussian variational approximation (Shun and McCullagh, 1995; Ranganath et al., 2014; Kucukelbir et al., 2017; Galy-Fajou et al., 2021; Xu and Campbell, 2022; Modi et al., 2023; Cai et al., 2024). Indeed, the latter strategy suggests one compelling use case of EigenVI for target distributions on $\mathbb{R}^D$; after a standardizing transformation, it provides a systematic framework to model non-Gaussian effects via low-order orthogonal function expansions.

## 3. Experiments

We evaluate EigenVI on 9 synthetic targets and 8 real data targets. The focus of the experiments is the orthogonal family induced from *normalized Hermite polynomials* (see Table 1), which has a Gaussian base distribution. Thus, this variational family can be viewed as a more flexible Gaussian variational family. In what follows, we first study 2D synthetic targets, which allow us to show the expressiveness of the basis using higher order expansions. Next, we study this approach under varying amounts of non-Gaussianity. Finally, we evaluate this approach on Bayesian modeling applications with real data, comparing with several BBVI algorithms.

### 3.1. 2D synthetic targets

In Figure 3, we show the resulting fits for three 2D target distributions; details on the target densities are in the appendix. We report the forward KL divergence in Figure 6 for varying number of samples $B$ and families constructed using different orders $K = K_1 = K_2$; we report time comparisons between the different orders. These targets were not standardized before fitting, as they were already centered by construction. The Gaussian distribution is fit using a score-based divergence (Cai et al., 2024; Modi et al., 2023).

### 3.2. Non-Gaussianity: varying skew and tails in the sinh-arcsinh distribution

We now consider a non-Gaussian target, where we can control the amount of skew and tails in the target. The sinh-arcsinh normal distribution (Jones and Pewsey, 2009; 2019), which has parameters $s \in \mathbb{R}^D, \tau \in \mathbb{R}^D_+, \Sigma \in \mathbb{S}_{++}$, is the distribution induced by transforming a Gaussian $Z_0 \sim \mathcal{N}(0, \Sigma)$ to $Z = \mathcal{S}_{s,\tau}(Z_0)$, where

$$\begin{aligned} \mathcal{S}_{s,\tau}(z) &:= [S_{s_1,\tau_1}(z_1), \ldots, S_{s_D,\tau_D}(z_D)]^\top, \\ S_{s_d,\tau_d}(z_d) &:= \sinh\left(\tfrac{1}{\tau_d}\sinh^{-1}(z_d) + \tfrac{s_d}{\tau_d}\right). \end{aligned} \tag{16}$$

Here $s_d$ controls the amount of skew in the $d$th dimension, and $\tau_d$ controls the tail weight in that dimension. When $s_d = 0$ and $\tau_d = 1$ in all dimensions $d$, the distribution is the Gaussian distribution.

In Figure 4, we show the effect of changing the parameters to highlight increasing amounts of non-Gaussianity in the skew or the tails of the distribution (see the appendix for the precise parameters); the 2D targets are visualized in the top row. Before applying EigenVI, we standardize the target using a mean and covariance estimated from batch and match VI (Cai et al., 2024). For EigenVI, we measure the forward KL under varying numbers of samples $B$ and across increasing numbers of basis functions, given by $K_1 \times \ldots \times K_D$. We also present the forward KL resulting from

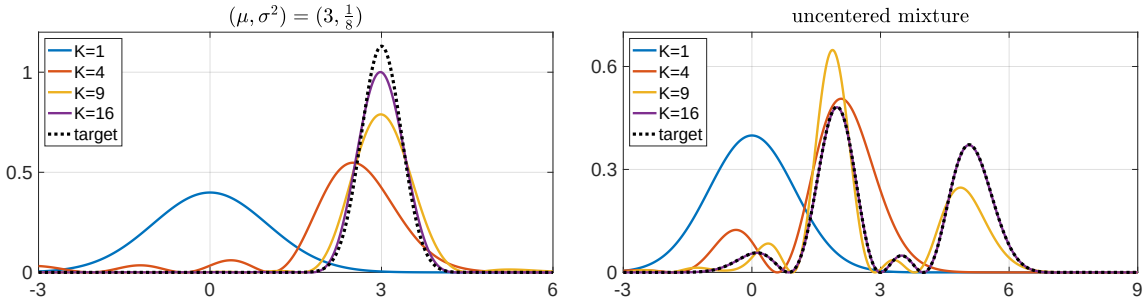



Figure 2: Higher-order expansions may be required to approximate target distributions that are not standardized. *Left:* approximation of a normal distribution. *Right:* approximation of the mixture distribution in Figure 1 after translating its largest modes away from the origin.

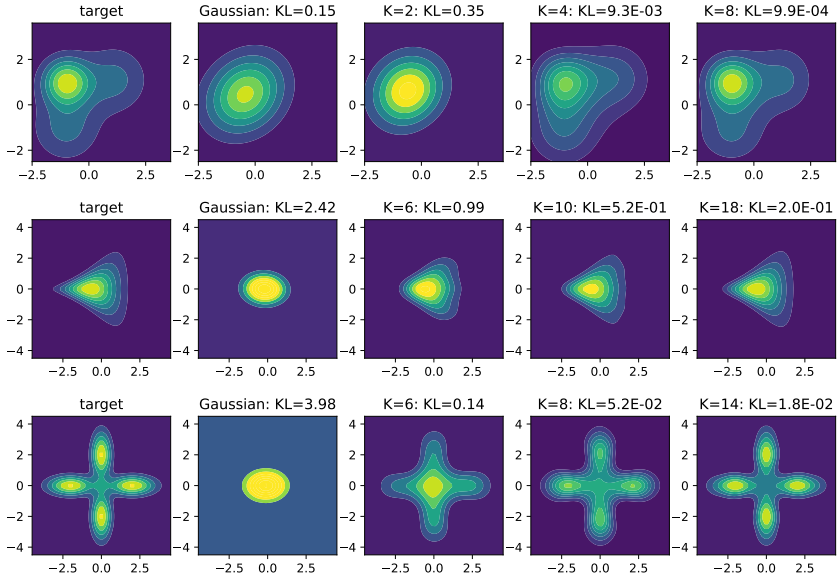



Figure 3: 2D target functions (column 1): a 3-component Gaussian mixture distribution (row 1), a funnel distribution (row 2), and a cross distribution (row 3). We report the $\mathrm{KL}(p; q)$ for the resulting optimal variational distributions obtained using score-based VI with a Gaussian variational family (second column) and the EigenVI variational family (columns 3–4), where $K_1{=}K_2{=}K$.

batch and match VI (BaM) and automatic differentiation VI (ADVI), which both use Gaussian variational families. Next we consider similar targets in 5 dimensions; for the parameters and visualizations of the targets and examples of the resulting EigenVI variational distributions, see Figure 7. In Figure 4c, we observe greater differences in the number of importance samples needed to lead to good approximations, especially as the number of basis functions increase.

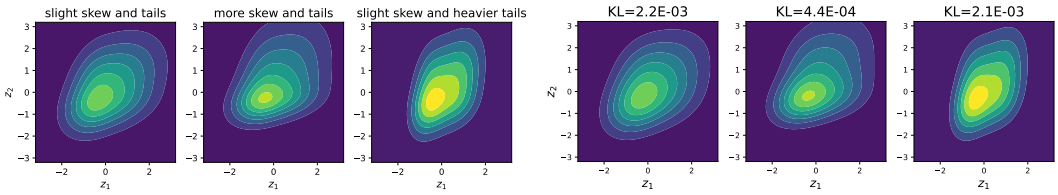
### 3.3. Hierarchical modeling benchmarks from posteriordb

We now evaluate EigenVI on a set of hierarchical Bayesian models (Carpenter et al., 2017; Magnusson et al., 2022; Roualdes et al., 2023), which are summarized in Table 2. The goal is posterior inference: given data observations $x_{1:N}$, the posterior of $z$ is

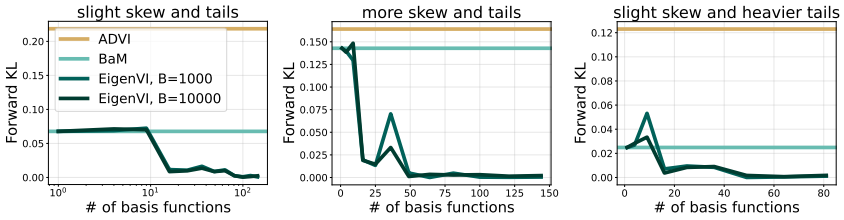
$$p(z \mid x_{1:N}) \propto p(z) L_z(x_{1:N}) =: \rho(z), \tag{17}$$

where $p(z)$ is the prior and $L_z(x_{1:N})$ denotes the likelihood.

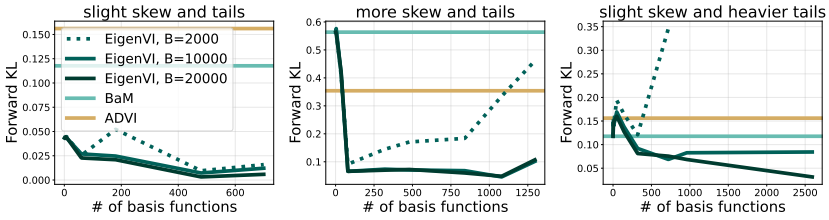
We compare EigenVI to 1) automatic differentation VI (ADVI) (Kucukelbir et al., 2017), which minimizes the ELBO over full covariance Gaussian family (ADVI-G), 2) Gaussian score matching (GSM) (Modi et al., 2023), a score-based BBVI approach with a full covariance Gaussian family, and 3) batch and match VI (BaM) (Cai et al., 2024), which minimizes a regularized score-based divergence over a full covariance Gaussian family. In these examples, we standardize the target using either GSM or BaM before applying EigenVI; for this reason, we do not compare the costs




(a) Example 2D targets (left) varying the skew $s$ or tail weight $\tau$ components and their EigenVI fits (right).




(b) $D = 2$




(c) $D = 5$

Figure 4: Sinh-arcsinh normal distribution synthetic target. Panel (a) shows the three targets we consider in 2D, and their resulting EigenVI fit. Panel (b) shows measures $\mathrm{KL}(p; q)$ for $D = 2$, and panel (c) shows $\mathrm{KL}(p; q)$ for $D = 7$; the $x$-axis shows the number of basis functions, $K_1 \times \ldots \times K_D$.

of these methods to EigenVI.

In these models, we do not have access to the target distribution, $p(z \mid x_{1:N})$, only the unnormalized target $\rho$. Thus, we cannot evaluate an estimate of the forward KL. Instead, to evaluate the fidelity of the fitted variational distributions, we compute the empirical Fisher divergence using reference samples from the posterior obtained via Hamiltonian Monte Carlo (HMC):

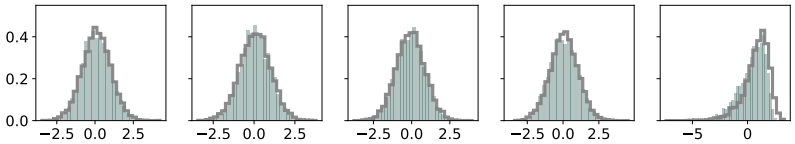
$$\frac{1}{S}\sum_{s=1}^{S} \|\nabla \log \rho(z^s) - \nabla \log q(z^s)\|^2, \quad z^s \sim p(z \mid x_{1:N}). \tag{18}$$

Note that this measure is not the objective that EigenVI minimizes; it is analogous to the forward KL divergence, as the expectation is taken with respect to $p$. We report the results in Figure 5, computing the Fisher divergence for EigenVI with increasing numbers of basis functions. We typically found that with more basis functions, the the scores becomes closer to that of the target.

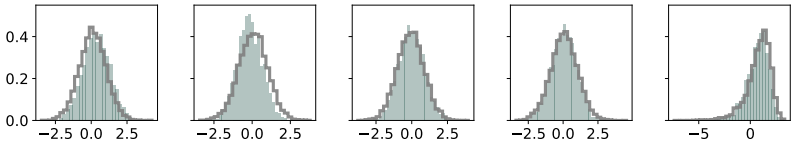
Finally, we also used Real-NVP normalizing flows (NFs) (Dinh et al., 2016) for VI by minimizing reverse KL. We found that by tuning the batch-size and learning rate, NFs generally could fit these models. However, these NFs do not have access to reliable scores (Köhler et al., 2021; Zeghal et al., 2022), hence we do not show their Fisher divergence in Figure 5. Instead we visualize the posterior marginals for a subset of dimensions from `8schools` in the top three rows, comparing EigenVI, VI with a normalizing flow, and BaM. In this example, we observe that the Gaussian struggles to fit the tails of this target distribution. While the normalizing flow appears to fit the tails a little better than EigenVI, it also includes other biases. In the appendix, we also show the full corner plot in Figure 8 and marginals of the `garch11` model in Figure 9.

## 4. Discussion of limitations and future work

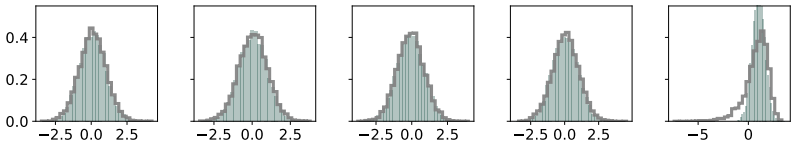
In this work, we introduced EigenVI, a new approach for score-based variational inference. We propose a new variational family built from orthogonal function expansions that supports analytical moments and exact sampling. We show that minimizing a score-based objective is equivalent to solving an eigenvalue problem, which leads to an alternative approach to gradient-based optimization. Importantly, in EigenVI many computations with respect to the batch of samples can be performed in parallel, unlike in iterative



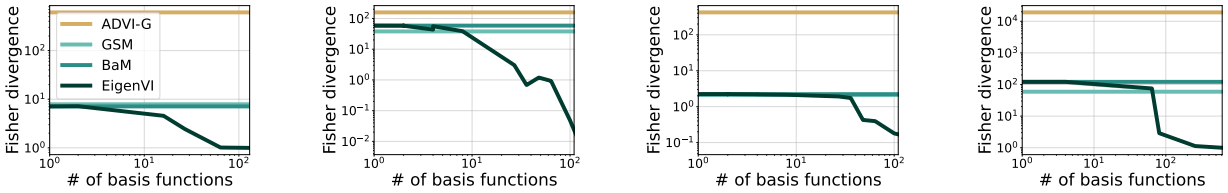
(a) EigenVI with normalized Hermite polynomial family



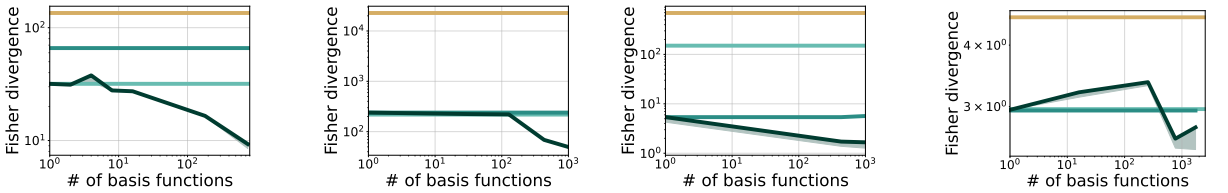
(b) VI with a normalizing flow family



(c) Batch and match VI with a full covariance Gaussian family




(d) `kidscore`, $D = 3$ (e) `sesame`, $D = 3$ (f) `gp-regr`, $D = 3$ (g) `logearn`, $D = 4$




(h) `garch11`, $D = 4$ (i) `arK-arK`, $D = 7$ (j) `logmesquite`, $D = 7$ (k) `8-schools`, $D = 10$

Figure 5: Results on `posteriordb` models. Top three rows: marginal distributions of the even dimensions from `8-schools`. Reference samples from HMC are outlined in gray, and the VI samples are in green. Bottom two rows: evaluation of methods with the (forward) Fisher divergence. The $x$-axis shows the number of basis functions, $K_1 \times \ldots \times K_D$. Shaded regions represent standard errors computed with respect to 5 random seeds.

methods. With experiments on synthetic and real-world targets, we show that EigenVI provides a principled way of improving upon Gaussian variational families.

Many future directions remain. First, the approach described in this paper uses importance sampling, and may benefit from using more sophisticated adaptive importance sampling methods. We leave this for future work. Second, our empirical study focused on a variational family built using normalized Hermite polynomials. Without utilizing higher-order function expansions, which are expensive in higher dimensions, this family is limited to target functions that are close to Gaussian. As we observed in our simulation studies, this was sufficient for many of the targets we considered. In future work, designing new orthogonal basis sets will be crucial for extension to highly non-Gaussian targets. Finally, a future direction is developing alternative families using orthogonal function expansions that have more favorable scaling with the dimension, such ones with low rank structure.

## References

O. Abril-Pla, V. Andreani, C. Carroll, L. Dong, C. J. Fonnesbeck, M. Kochurov, R. Kumar, J. Lao, C. C. Luhmann, O. A. Martin, et al. PyMC: a modern, and comprehensive probabilistic programming framework in Python. *PeerJ Computer Science*, 9:e1516, 2023.

A. Agrawal, D. R. Sheldon, and J. Domke. Advances in

black-box VI: Normalizing flows, importance weighting, and optimization. *Advances in Neural Information Processing Systems*, 33, 2020.

R. v. d. Berg, L. Hasenclever, J. M. Tomczak, and M. Welling. Sylvester normalizing flows for variational inference. *arXiv preprint arXiv:1803.05649*, 2018.

E. Bingham, J. P. Chen, M. Jankowiak, F. Obermeyer, N. Pradhan, T. Karaletsos, R. Singh, P. Szerlip, P. Horsfall, and N. D. Goodman. Pyro: Deep universal probabilistic programming. *The Journal of Machine Learning Research*, 20(1):973–978, 2019.

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

D. Cai, C. Modi, L. Pillaud-Vivien, C. Margossian, R. Gower, D. Blei, and L. Saul. Batch and match: black-box variational inference with a score-based divergence. In *International Conference on Machine Learning*, 2024.

B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1):1–32, 2017.

R. Courant and D. Hilbert. *Methoden der Mathematischen Physik*, volume 1. Julius Springer, Berlin, 1924.

B. Dai, H. Dai, A. Gretton, L. Song, D. Schuurmans, and N. He. Kernel exponential family estimation via doubly dual embedding. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2019.

A. K. Dhaka, A. Catalina, M. R. Andersen, M. Magnusson, J. Huggins, and A. Vehtari. Robust, accurate stochastic optimization for variational inference. *Advances in Neural Information Processing Systems*, 33, 2020.

A. K. Dhaka, A. Catalina, M. Welandawe, M. R. Andersen, J. Huggins, and A. Vehtari. Challenges and opportunities in high dimensional variational inference. *Advances in Neural Information Processing Systems*, 34, 2021.

L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using real NVP. *arXiv preprint arXiv:1605.08803*, 2016.

T. Galy-Fajou, V. Perrone, and M. Opper. Flexible and efficient inference with particles for the variational gaussian approximation. *Entropy*, 23(8):990, 2021.

H. Ge, K. Xu, and Z. Ghahramani. Turing: a language for flexible probabilistic inference. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2018.

R. Giordano, M. Ingram, and T. Broderick. Black box variational inference with a deterministic objective: Faster, more accurate, and even more black box. *Journal of Machine Learning Research*, 25(18):1–39, 2024.

D. J. Griffiths and D. F. Schroeter. *Introduction to Quantum Mechanics*. Cambridge University Press, 2018.

C. Jones and A. Pewsey. Sinh-arcsinh distributions. *Biometrika*, 96(4):761–780, 2009.

C. Jones and A. Pewsey. The sinh-arcsinh normal distribution. *Significance*, 16(2):6–7, 2019.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

T. Kim and Y. Bengio. Deep directed generative models with energy-based probability estimation. *arXiv preprint arXiv:1606.03439*, 2016.

D. P. Kingma and M. Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

D. P. Kingma, T. Salimans, R. Jozefowicz, X. Chen, I. Sutskever, and M. Welling. Improved variational inference with inverse autoregressive flow. *Advances in Neural Information Processing Systems*, 29, 2016.

I. Kobyzev, S. J. Prince, and M. A. Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3964–3979, 2020.

J. Köhler, A. Krämer, and F. Noé. Smooth normalizing flows. *Advances in Neural Information Processing Systems*, 34, 2021.

A. Kucukelbir, D. Tran, R. Ranganath, A. Gelman, and D. M. Blei. Automatic differentiation variational inference. *Journal of Machine Learning Research*, 2017.

J. Lawson, G. Tucker, B. Dai, and R. Ranganath. Energy-inspired models: Learning with sampler-induced distributions. *Advances in Neural Information Processing Systems*, 32, 2019.

Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. Huang. A tutorial on energy-based learning. *Predicting Structured Data*, 1(0), 2006.

Q. Liu and D. Wang. Stein variational gradient descent: a general purpose Bayesian inference algorithm. *Advances in Neural Information Processing Systems*, 29, 2016.

F. Locatello, G. Dresdner, R. Khanna, I. Valera, and G. Rätsch. Boosting black box variational inference. *Advances in Neural Information Processing Systems*, 31, 2018.

L. Loconte, A. M. Sladek, S. Mengel, M. Trapp, A. Solin, N. Gillis, and A. Vergari. Subtractive mixture models via squaring: Representation and learning. In *International Conference on Learning Representations*, 2024.

C. Louizos and M. Welling. Multiplicative normalizing flows for variational Bayesian neural networks. In *International Conference on Machine Learning*. PMLR, 2017.

M. Magnusson, P. Bürkner, and A. Vehtari. posteriordb: a set of posteriors for Bayesian inference and probabilistic programming. https://github.com/stan-dev/posteriordb, 2022.

C. Modi, C. Margossian, Y. Yao, R. Gower, D. Blei, and L. Saul. Variational inference with Gaussian score matching. *Advances in Neural Information Processing Systems*, 36, 2023.

G. S. Novikov, M. E. Panov, and I. V. Oseledets. Tensor-train density estimation. In *Uncertainty in Artificial Intelligence*, pages 1321–1331. PMLR, 2021.

G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

R. Ranganath, S. Gerrish, and D. Blei. Black box variational inference. In *Artificial Intelligence and Statistics*, pages 814–822. PMLR, 2014.

D. Rezende and S. Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*. PMLR, 2015.

E. Roualdes, B. Ward, S. Axen, and B. Carpenter. BridgeStan: Efficient in-memory access to Stan programs through Python, Julia, and R. https://github.com/roualdes/bridgestan, 2023.

J. Salvatier, T. V. Wiecki, and C. Fonnesbeck. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2:e55, 2016.

Z. Shun and P. McCullagh. Laplace approximation of high dimensional integrals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 57(4):749–760, 1995.

M. Titsias and M. Lázaro-Gredilla. Doubly stochastic variational Bayes for non-conjugate inference. In *International Conference on Machine Learning*. PMLR, 2014.

M. J. Wainwright, M. I. Jordan, et al. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008.

X. Wang, T. Geffner, and J. Domke. Dual control variate for faster black-box variational inference. In *International Conference on Artificial Intelligence and Statistics*, 2024.

Z. Xu and T. Campbell. The computational asymptotics of Gaussian variational inference and the Laplace approximation. *Statistics and Computing*, 32(4):63, 2022.

Y. Yang, R. Martin, and H. Bondell. Variational approximations using Fisher divergence. *arXiv preprint arXiv:1905.05284*, 2019.

L. Yu and C. Zhang. Semi-implicit variational inference via score matching. In *International Conference on Learning Representations*, 2023.

J. Zeghal, F. Lanusse, A. Boucaud, B. Remy, and E. Aubourg. Neural Posterior Estimation with Differentiable Simulators. In *International Conference on Machine Learning Conference*, 2022.

C. Zhang, B. Shahbaba, and H. Zhao. Variational Hamiltonian Monte Carlo via score matching. *Bayesian Analysis*, 13(2):485, 2018.

L. Zhang, B. Carpenter, A. Gelman, and A. Vehtari. Pathfinder: Parallel quasi-newton variational inference. *Journal of Machine Learning Research*, 23(306):1–49, 2022.

S. C. Zhu, Y. Wu, and D. Mumford. Filters, random fields and maximum entropy (FRAME): Towards a unified theory for texture modeling. *International Journal of Computer Vision*, 27:107–126, 1998.

D. Zoltowski, D. Cai, and R. P. Adams. Slice sampling reparameterization gradients. *Advances in Neural Information Processing Systems*, 34:23532–23544, 2021.

## A. Related work

Several recent works have considered black-box variational inference methods based on minimizing the scores between the variational distribution and the target distribution. However, in many cases, the focus of these methods all are limited to Gaussian variational families (Modi et al., 2023; Cai et al., 2024). The Fisher divergence has been previously studied as a divergence for variational inference by several others, e.g., Yang et al. (2019). In the context of non-Gaussian variational families, Yu and Zhang (2023) propose minimizing a Fisher divergence for semi-implicit variational families; the divergence is minimized using gradient-based optimization. In another line of work, Zhang et al. (2018) minimize the Fisher divergence with an energy-based model as the variational family, and they show it can be minimized with a closed-form solution.

More generally, variational inference has been applied to more expressive variational families such as energy-based models (Zhu et al., 1998; LeCun et al., 2006; Kim and Bengio, 2016; Dai et al., 2019; Lawson et al., 2019; Zoltowski et al., 2021) and normalizing flows (Rezende and Mohamed, 2015; Kingma et al., 2016; Louizos and Welling, 2017; Berg et al., 2018; Kobyzev et al., 2020; Papamakarios et al., 2021). However the performance of these models, especially the normalizing flows, is often sensitive to the hyperparameters of the flow architecture, optimization algorithm and parameters of the base distribution (Dhaka et al., 2021; Agrawal et al., 2020). Futhermore, these models do not support closed-form computations for lower-order moments. In the case of energy-based models, they cannot be sampled from or evaluated, and for normalizing flows, the scores are not available, thus precluding the use of score-based divergences.

The idea of using a squaring a sum of basis functions as a way to model a distribution has appeared elsewhere in the machine literature. Novikov et al. (2021) propose a tensor train-based model for density estimation, but they do not consider orthogonal basis sets. Similarly, Loconte et al. (2024) consider squaring a set of basis functions as a mixture model with negative weights, studying this model in conjunction with probabilistic circuits.

## B. Sampling from orthogonal function expansions

In this appendix we show how to sample from a density on $\mathbb{R}^D$ constructed from a Cartesian product of orthogonal function expansions. Specifically, we assume that the density is of the form

$$q(z_1, z_2, \ldots, z_D) = \left( \sum_{k_1=1}^{K_1} \cdots \sum_{k_D=1}^{K_D} \alpha_{k_1 k_2 \ldots k_D} \phi_{k_1}(z_1) \phi_{k_2}(z_2) \cdots \phi_{k_D}(z_D) \right)^2, \tag{19}$$

where $\{\phi_k(\cdot)\}_{k=1}^{\infty}$ define a family of orthonormal functions on $\mathbb{R}$ and where the density is normalized by requiring that

$$\sum_{k_1 k_2 \ldots k_D} \alpha^2_{k_1 k_2 \ldots k_D} = 1. \tag{20}$$

To draw samples from this density, we describe a sequential procedure based on inverse transform sampling. In particular, we obtain a sample $z \in \mathcal{R}^D$ by the sequence of draws

$$\begin{aligned} z_1 &\sim q(z_1), && (21)\\ z_2 &\sim q(z_2|z_2), && (22)\\ \vdots &\quad \vdots \quad \vdots && (23)\\ z_D &\sim q(z_D|z_1, z_2, \ldots, z_{D-1}). && (24) \end{aligned}$$

This basic strategy can also be used to sample from distributions whose domains are Cartesian products of different one-dimensional spaces.

### Core primitive

First we describe the core primitive that we will use for each of the draws in eqs. (21–24). To begin, we observe the following: if $S$ is any positive semidefinite matrix with $\text{trace}(S) = 1$, then

$$\rho(\xi) = \sum_{k,\ell=1}^{K} S_{k\ell} \phi_k(\xi) \phi_\ell(\xi), \tag{25}$$

defines a normalized density over $\mathbb{R}$. In particular, since $S \succeq 0$, it follows that $\rho(\xi) \geq 0$ for all $\xi \in \mathbb{R}$, and since $\text{trace}(S) = 1$, it follows that

$$\int_{-\infty}^{\infty} \rho(\xi)\, d\xi = \sum_{k,\ell=1}^{K} S_{k\ell} \int_{-\infty}^{\infty} \phi_k(\xi)\phi_\ell(\xi)\, d\xi = \sum_{k,\ell=1}^{K} S_{k\ell}\delta_{kl} = \text{trace}(S) = 1. \tag{26}$$

The core primitive that we need is an efficient procedure to sample from a normalized density of this form. We will see later that all of the densities in eq. (21–24) can be expressed in this form.

### Inverse transform sampling

Since the density in eq. (25) is one-dimensional, we can obtain the draw we need by inverse transform sampling. In particular, let $C(\xi)$ denote the cumulative distribution function (CDF) associated to (25), which is given by

$$\mathcal{C}(\xi) = \int_{-\infty}^{\xi} \rho(z)\, dz, \tag{27}$$

and let $C^{-1}(\xi)$ denote the inverse CDF. Then at least in principle, we can draw a sample from $\rho$ by the two-step procedure

$$u \sim \text{Uniform}[0, 1], \tag{28}$$

$$\xi = C^{-1}(u). \tag{29}$$

Next we consider how to implement this procedure efficiently in practice, and in particular, how to calculate the definite integral for the CDF in eq. (27). As shorthand, we define the doubly-indexed set of real-valued functions

$$\Phi_{k\ell}(\xi) = \int_{-\infty}^{\xi} \phi_k(z)\phi_\ell(z)\, dz. \tag{30}$$

It follows from orthogonality that $\Phi_{kl}(+\infty) = \delta_{kl}$ and from the Cauchy-Schwartz inequality that $|\Phi_{k\ell}(\xi)| \leq 1$ for all $\xi \in \mathbb{R}$. Our interest in these functions stems from the observation that

$$C(\xi) = \sum_{k,\ell=1}^{K} S_{k\ell}\Phi_{kl}(\xi) = \text{trace}[S\Phi(\xi)], \tag{31}$$

so that if we have already computed the functions $\Phi_{k\ell}(\xi)$, then we can use eq. (31) to compute the CDF whose inverse we need in eq. (29). In practice, we can use numerical quadrature to pre-compute $\Phi_{k\ell}(\xi)$ for many values along the real line and then solve eq. (29) quickly by interpolation; that is, given $u$, we find $\xi$ satisfying $\text{trace}[S\Phi(\xi)] = u$. The result is an unbiased sample drawn from the density $\rho(\xi)$ in eq. (25).

### Sequential sampling

Finally we show that each draw in eqs. (21–24) reduces to the problem described above. As in section 2.1, we work out the steps specifically for an example in $D = 3$, where we must draw the samples $z_1 \sim q(z_1)$, $z_2 \sim q(z_2|z_1)$ and $z_3 \sim q(z_3|z_1, z_2)$. This example illustrates all the ideas needed for the general case but with a minimum of indices. Consider the joint distribution given by

$$q(z_1, z_2, z_3) = \left( \sum_{i=1}^{K_1} \sum_{j=1}^{K_2} \sum_{k=1}^{K_3} \beta_{ijk}\, \phi_i(z_1)\phi_j(z_2)\phi_k(z_3) \right)^2 \quad \text{where} \quad \sum_{ijk} \beta_{ijk}^2 = 1. \tag{32}$$

From this joint distribution, we can compute marginal distributions by integrating out subsets of variables, and each integration over $\mathbb{R}$ gives rise to a contraction of indices, as in eq. (7), due to the property of orthogonality. In this way we find

$$q(z_1, z_2) = \sum_{j,j'=1}^{K_2} \left[ \sum_{i,i'=1}^{K_1} \sum_{k=1}^{K_3} \beta_{ijk}\beta_{i'j'k}\phi_i(z_1)\phi_{i'}(z_1) \right] \phi_j(z_2)\phi_{j'}(z_2), \tag{33}$$

$$q(z_1) = \sum_{i,i'=1}^{K_1} \left[ \sum_{j=1}^{K_2} \sum_{k=1}^{K_3} \beta_{ijk}\beta_{i'jk} \right] \phi_i(z_1)\phi_{i'}(z_1). \tag{34}$$

Now we note from the brackets in eq. (34) that this marginal distribution is already in the quadratic form of eq. (25) with coefficients

$$S^{(1)}_{ii'} = \sum_{j=1}^{K_2} \sum_{k=1}^{K_3} \beta_{ijk} \beta_{i'jk}. \tag{35}$$

From this first quadratic form, we can therefore use inverse transform sampling to obtain a draw $z_1 \sim q(z_1)$ Next we consider how to sample from $q(z_2|z_1) = q(z_1, z_2)/q(z_1)$. Again, from the brackets in eq. (33), we see that this conditional distribution is also in the quadratic form of eq. (25) with coefficients

$$S^{(2)}_{jj'} = \frac{\sum_{i,i'=1}^{K_1} \sum_{k=1}^{K_3} \beta_{ijk} \beta_{i'j'k} \phi_i(z_1) \phi_{i'}(z_1)}{q(z_1)}. \tag{36}$$

From this second quadratic form, we can therefore use inverse transform sampling to obtain a draw $z_2 \sim q(z_2|z_1)$. Finally, we consider how to sample from $q(z_3|z_1, z_2) = q(z_1, z_2, z_3)/q(z_1, z_2)$. From eq. (32, we see that this conditional distribution is also in the quadratic form of eq. (25) with coefficients

$$S^{(3)}_{kk'} = \frac{\sum_{i,i'=1}^{K_1} \sum_{j,j'=1}^{K_2} \beta_{ijk} \beta_{i'j'k'} \phi_i(z_1) \phi_{i'}(z_1) \phi_j(z_2) \phi_{j'}(z_2)}{q(z_1, z_2)} \tag{37}$$

From this third quadratic form, we can therefore use inverse transform sampling to obtain a draw $z_3 \sim q(z_3|z_1, z_2)$. Finally, from the sums in eq. (37, we see that the overall cost of this procedure is $O(K_1^2 K_2^2 K_3^2)$, or quadratic in the total number of basis functions.

## C. Calculation of moments

In this appendix we show how to calculate the low-order moments of a density constructed from the Cartesian product of orthogonal function expansions. In particular, we assume that the density is over $\mathbb{R}^D$ and of the form

$$q(z_1, z_2, \ldots, z_D) = \left( \sum_{k_1=1}^{K_1} \cdots \sum_{k_D=1}^{K_D} \alpha_{k_1 k_2 \ldots k_D} \phi_{k_1}(z_1) \phi_{k_2}(z_2) \cdots \phi_{k_D}(z_D) \right)^2, \tag{38}$$

where $\{\phi_k(\cdot)\}_{k=1}^{\infty}$ are orthogonal functions on $\mathbb{R}$ and where the coefficients are properly normalized so that the density integrates to one. For such a density, we show that the calculation of first and second-order moments boils down to evaluating *one-dimensional* integrals of the form

$$\mu_{ij} = \int_{-\infty}^{\infty} \phi_i(z) \phi_j(z)\, z\, dz, \tag{39}$$

$$\nu_{ij} = \int_{-\infty}^{\infty} \phi_i(z) \phi_j(z)\, z^2\, dz. \tag{40}$$

We also show how to evaluate these integrals specifically for the orthogonal family of weighted Hermite polynomials.

First we consider how to calculate moments such as $\mathbb{E}_q[z_d^p]$, where $p \in \{1, 2\}$, and without loss of generality we focus on calculating $\mathbb{E}_q[z_1^p]$. We start from the joint distribution in eq. (38) and proceed by marginalizing over the variables $(z_2, z_3, \ldots, z_D)$. Exploiting orthogonality, we find that

$$\mathbb{E}_q[z_1^p] = \int q(z_1, z_2, \ldots, z_D)\, z_1^p\, dz_1\, dz_2\, \ldots dz_D, \tag{41}$$

$$= \int \left( \sum_{k_1=1}^{K_1} \cdots \sum_{k_D=1}^{K_D} \alpha_{k_1 k_2 \ldots k_D} \phi_{k_1}(z_1) \phi_{k_2}(z_2) \cdots \phi_{k_D}(z_D) \right)^2 z_1^p\, dz_1\, dz_2\, \ldots dz_D, \tag{42}$$

$$= \sum_{k_1, k_1'=1}^{K_1} \left[ \sum_{k_2=1}^{K_2} \cdots \sum_{k_D=1}^{K_D} \alpha_{k_1 k_2 \ldots k_D} \alpha_{k_1' k_2 \ldots k_D} \right] \int \phi_{k_1}(z_1) \phi_{k_1'}(z_1)\, z_1^p\, dz_1. \tag{43}$$

We can rewrite this expression more compactly as a quadratic form over integrals of the form in eqs. (39–40). To this end, we define the coefficients

$$A_{ij} = \sum_{k_2=1}^{K_2} \cdots \sum_{k_D=1}^{K_D} \alpha_{ik_2\ldots k_D}\alpha_{jk_2\ldots k_D} \tag{44}$$

which simply encapsulate the bracketed term in eq. (43). Note that there are $K_1^2$ of these coefficients, each of which can be computed in $O(K_2K_3\ldots K_D)$. With this shorthand, we can write

$$\mathbb{E}_q[z_1] = \sum_{i,j=1}^{K_1} A_{ij}\mu_{ij}, \tag{45}$$

$$\mathbb{E}_q[z_1^2] = \sum_{i,j=1}^{K_1} A_{ij}\nu_{ij}, \tag{46}$$

where $\mu_{ij}$ and $\nu_{ij}$ are the integrals defined in eqs. (39–40). Thus the problem has been reduced to a weighted sum of one-dimensional integrals.

A similar calculation gives the result we need for correlations. Again, without loss of generality, we focus on calculating $\mathbb{E}_q[z_1z_2]$. Analogous to eq. (44), we define the tensor of coefficients

$$B_{ijk\ell} = \sum_{k_3=1}^{K_3} \cdots \sum_{k_D=1}^{K_D} \alpha_{ikk_3\ldots k_D}\alpha_{j\ell k_3\ldots k_D}, \tag{47}$$

which arises from marginalizing over the variables $(z_3, z_4, \ldots, z_D)$. There are $K_1^2K_2^2$ of these coefficients, each of which can be computed in $O(K_3K_4\ldots K_D)$. With this shorthand, we can write

$$\mathbb{E}_q[z_1z_2] = \sum_{i,j=1}^{K_1}\sum_{k,\ell=1}^{K_2} B_{ijk\ell}\mu_{ij}\mu_{k\ell}. \tag{48}$$

where $\mu_{ij}$ is again the integral defined in eq. (39). Thus the problem has been reduced to a weighted sum of (the product of) one-dimensional integrals.

Finally, we show how to evaluate the integrals in eqs. (39–40) for the specific case of orthogonal function expansions with weighted Hermite polynomials. Recall in this case that

$$\phi_{k+1}(z) = \left(\sqrt{2\pi}k!\right)^{-\frac{1}{2}}\left(e^{-\frac{1}{2}z^2}\right)^{\frac{1}{2}}\,\mathrm{H}_k(z), \tag{49}$$

where $\mathrm{H}_k(z)$ are the *probabilist's* Hermite polynomials given by

$$\mathrm{H}_k(z) = (-1)^k e^{\frac{z^2}{2}}\frac{d^k}{dz^k}\left[e^{-\frac{z^2}{2}}\right]. \tag{50}$$

To evaluate the integrals for this particular family, we can exploit the following recursions that are satisfied by Hermite polynomials:

$$H_{k+1}(z) = zH_k(z) - H_k'(z), \tag{51}$$

$$H_k'(z) = kH_{k-1}(z). \tag{52}$$

Eliminating the derivatives $H_k'(z)$ in eqs. (51–52), we see that $zH_k(z) = H_{k+1}(z) + kH_{k-1}(z)$. We can then substitute eq. (49) to obtain a recursion

$$z\phi_k(z) = \sqrt{k}\phi_{k+1}(z) + \sqrt{k-1}\phi_{k-1}(z) \tag{53}$$

for the orthogonal basis functions themselves. With the above recursion, we can now read off these integrals from the property of orthogonality. For example, starting from eq. (39), we find that

$$\mu_{ij} = \int_{-\infty}^{\infty} \phi_i(z)\phi_j(z)\, z\, dz, \tag{54}$$

$$= \int_{-\infty}^{\infty} \phi_i(z) \left[\sqrt{j}\phi_{j+1}(z) + \sqrt{j-1}\phi_{j-1}(z)\right] dz, \tag{55}$$

$$= \delta_{i,j+1}\sqrt{j} + \delta_{i,j-1}\sqrt{i}, \tag{56}$$

where $\delta_{ij}$ is the Kronecker delta function. Next we consider the integral in eq. (40), which involves a power of $z^2$ in the integrand. In this case we can make repeated use of the recursion:

$$\nu_{ij} = \int_{-\infty}^{\infty} \phi_i(z)\phi_j(z)\, z^2\, dz, \tag{57}$$

$$= \int_{-\infty}^{\infty} \left[\sqrt{i}\phi_{i+1}(z) + \sqrt{i-1}\phi_{i-1}(z)\right] \left[\sqrt{j}\phi_{j+1}(z) + \sqrt{j-1}\phi_{j-1}(z)\right] dz, \tag{58}$$

$$= \delta_{ij}\left[\sqrt{ij} + \sqrt{(i-1)(j-1)}\right] + \delta_{i-1,j+1}\sqrt{j(j+1)} + \delta_{j-1,i+1}\sqrt{i(i+1)}. \tag{59}$$

Note that the matrices in eqs. (56) and (59) can be computed for whatever size is required by the orthogonal basis function expansion in eq. (38). Once these matrices are computed, it is a simple matter of substitution[1] to compute the moments $\mathbb{E}_q[z_1]$, $\mathbb{E}_q[z_1^2]$, and $\mathbb{E}_q[z_1 z_2]$ from eqs. (45–46) and eq. (48). Finally, we can compute other low-order moments (such as $\mathbb{E}_q[z_5]$ or $\mathbb{E}_q[z_3 z_7]$) by an appropriate permutation of indices.

## D. Eigenvalue problem

In this appendix we show in detail how the optimization for EigenVI reduces to a minimum eigenvalue problem. In particular we prove the following.

**Lemma D.1.** *Let $\{\phi_k(z)\}_{k=1}^{\infty}$ be an orthogonal function expansion, and let $q \in \mathcal{Q}_K$ be the variational approximation parameterized by*

$$q(z) = \left[\sum_{k=1}^{K} \alpha_k \phi_k(z)\right]^2, \tag{60}$$

*where the weights satisfy $\sum_{k=1}^{K} \alpha_k^2 = 1$, thus ensuring that the distribution is normalized. Suppose furthermore that $q$ is chosen to minimize the empirical estimate of the Fisher divergence given, as in eq. (10), by*

$$\widehat{\mathscr{D}}_\pi(q,p) = \sum_{b=1}^{B} \frac{q(z^b)}{\pi(z^b)} \left\|\nabla \log q(z^b) - \nabla \log p(z^b)\right\|^2.$$

*Then the optimal variational approximation $q$ in this family can be computed by solving the minimum eigenvalue problem*

$$\min_{q\in\mathcal{Q}_K} \left[\widehat{\mathscr{D}}(q,p)\right] = \min_{\|\alpha\|=1} \left[\sum_{j,k=1}^{K} M_{jk}\alpha_j\alpha_k\right] =: \lambda_{min}(M), \tag{61}$$

*and the optimal weights $\alpha$ are given (up to an arbitrary sign) by the corresponding eigenvector of this minimal eigenvalue.*

*Proof.* The scores of $q$ in this variational family are given by

$$\nabla \log q(z^b) = \frac{2\sum_k \alpha_k \nabla\phi_k(z^b)}{\sum_k \alpha_k \phi_k(z^b)}.$$

[1] With further bookkeeping, one can also exploit the *sparsity* of $\mu_{ij}$ and $\nu_{ij}$ to derive more efficient calculations of these moments.

Substituting the above into the empirical divergence, we find that

$$\begin{aligned}
\widehat{\mathscr{D}}_\pi(q,p) &= \sum_{b=1}^{B} \frac{q(z^b)}{\pi(z^b)} \left\|\nabla \log q(z^b) - \nabla \log p(z^b)\right\|^2 \\
&= \sum_{b=1}^{B} \frac{\left(\sum_k \alpha_k \phi_k(z^b)\right)^2}{\pi(z^b)} \left\|\frac{2\sum_k \alpha_k \nabla\phi_k(z^b)}{\sum_k \alpha_k \phi_k(z^b)} - \nabla \log p(z^b)\right\|^2 \\
&= \sum_{b=1}^{B} \frac{1}{\pi(z^b)} \left\|2\sum_k \alpha_k \nabla\phi_k(z^b) - \sum_k \alpha_k \phi_k(z^b) \nabla \log p(z^b)\right\|^2 \\
&= \sum_{b=1}^{B} \frac{1}{\pi(z^b)} \left\|\sum_k \alpha_k \left(2\nabla\phi_k(z^b) - \phi_k(z^b)\nabla \log p(z^b)\right)\right\|^2 \\
&= \alpha^\top M \alpha,
\end{aligned}$$

where $M$ is given in (13) and $\alpha = [\alpha_1, \ldots, \alpha_K] \in \mathcal{R}^K$. Thus the empirical divergence is a convex quadratic function of $\alpha$. Furthermore, since the gradient of the constraint $\alpha^\top \alpha = 1$ is always non-zero, it follows that the constraint qualification holds and the solution to eq. (61) must satisfy the KKT equations.

The Lagrangian associated with eq. (61) is given by

$$L(\alpha, \mu) := \widehat{\mathscr{D}}_\pi(q,p) + \mu \left(\sum_k \alpha_k^2 - 1\right),$$

where $\mu \in \mathcal{R}$ is the Lagrange multiplier. The associated KKT equations are

$$0 = \nabla_\alpha \widehat{\mathscr{D}}_\pi(q,p) + \mu \nabla_\alpha \left(\sum_k \alpha_k^2 - 1\right), \tag{62}$$

$$0 = \sum_k \alpha_k^2 - 1. \tag{63}$$

Computing the gradients in $\alpha$ of the above, we find that

$$M\alpha + 2\mu\alpha = 0. \tag{64}$$

Left multiplying the above by $\alpha^\top$, then enforcing the constraint, we find that

$$\alpha^\top M \alpha + 2\mu = 0,$$

or equivalently that

$$\mu = -\tfrac{1}{2}\alpha^\top M \alpha.$$

Finally, isolating $\mu$ and substituting back into (64) gives

$$M\alpha = (\alpha^\top M \alpha)\alpha.$$

Consequently $\alpha$ is an eigenvector of $M$, with an associated eigenvalue $(\alpha^\top M \alpha)$. Though any eigenvalue-eigenvector pair provides a valid solution to the KKT equations, we want the solution that minimizes the objective function. Since ths objective is equivalent to $\alpha^\top M \alpha$, we have that $\alpha$ should be the eigenvector associated to the smallest eigenvalue. Since $M$ is also a symmetric matrix, we have that the smallest eigenvalue is given by minimizing the Rayleigh quotient (Courant and Hilbert, 1924), which is equivalent to the right hand side of eq. (61). □

## E. Additional experiments and details

### E.1. Computational resources

The experiments were run on a Linux workstation with a 32-core Intel(R) Xeon(R) w5-3435X processor and with 503 GB of memory. Experiments were run on CPU. In the sinh-arcsinh and posteriordb experiments, computations to construct the matrix $M$ were parallelized over 28 threads.

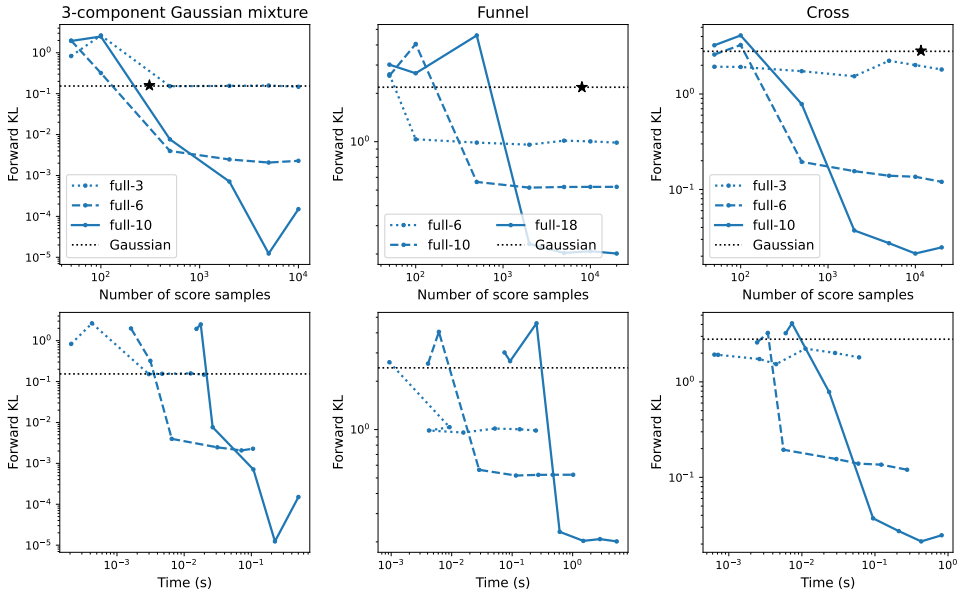



Figure 6: We compare # of score evaluations wallclock vs FKL divergence for the target distributions in Figure 3: the Gaussian mixture (column 1), the funnel (column 2), and the cross (column 3) distributions. The $K$ used for EigenVI is reported in each figure legend, where $K_1 = K_2 = K$. The black star denotes the number of gradient evaluations for the Gaussian method.

### E.2. 2D synthetic targets

We considered the following targets.

**3-component Gaussian mixture:**

$$p(z) = 0.4\mathcal{N}(z \,|\, [-1,1]^\top, \Sigma) + 0.3\mathcal{N}(z \,|\, [1.1,1.1]^\top, 0.5I) + 0.3\mathcal{N}(z \,|\, [-1,-1]^\top, 0.5I),$$

where $\Sigma = \begin{bmatrix} 2 & 0.1 \\ 0.1 & 2 \end{bmatrix}$.

**Funnel distribution with $\sigma^2 = 1.2$:**

$$p(z) = \mathcal{N}(z_1 \,|\, 0, \sigma^2)\mathcal{N}(z_2 \,|\, 0, \exp(z_1/2)).$$

**Cross distribution:**

$$p(z) = \tfrac{1}{4}\mathcal{N}(z \,|\, [0,2]^\top, \Sigma_1) + \tfrac{1}{4}\mathcal{N}(z \,|\, [-2,0]^\top, \Sigma_2) + \tfrac{1}{4}\mathcal{N}(z \,|\, [2,0]^\top, \Sigma_2) + \tfrac{1}{4}\mathcal{N}(z \,|\, [0,-2]^\top, \Sigma_1),$$

where $\Sigma_1 = \begin{bmatrix} 0.15^{0.9} & 0 \\ 0 & 1 \end{bmatrix}$ and $\Sigma_2 = \begin{bmatrix} 1 & 0 \\ 0 & 0.15^{0.9} \end{bmatrix}$.

These experiments were conducted without standardization with a Gaussian VI estimate. The EigenVI proposal distribution $\pi$ used was a uniform$([-5,5]^2)$.

In Figure 7, we run EigenVI for increasing numbers of importance samples $B$ and report the resulting forward KL divergence. The blue curves denote variational families with different $K_1 = K_2 = K$ values used, i.e., 3, 6, and 10 (resulting in a total number of basis functions of $3^2$, $6^2$, and $10^2$). In the bottom row of the plot, we also show wall clock timings (computed without parallelization) to show how the cost grows with the increase in the number of basis functions and importance samples. The horizontal dotted line denotes the result from batch and match VI, which fits a Gaussian via score matching; here a batch size of 16 was used and a learning rate of $\lambda_t = \frac{BD}{t+1}$.

The black star denotes the number of score evaluations used by the Gaussian VI method.

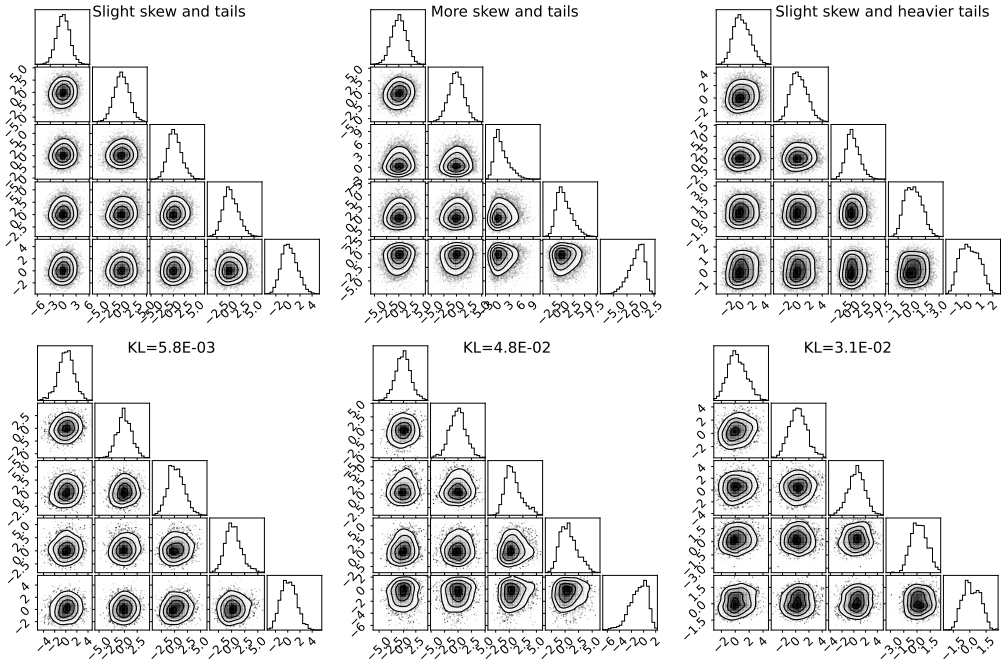



Figure 7: Targets (top) for the 5D sinh-arcsinh normal distribution example and EigenVI fits (bottom) with the KL divergence in the figure title.

### E.3. Sinh-arcsinh targets

The sinh-arcsinh normal distribution has the following density:

$$p(z; s, \tau, \Sigma) = [(2\pi)^D |\Sigma|]^{-\frac{1}{2}} \prod_{d=1}^{D} \left\{ (1 + z_d^2)^{-\frac{1}{2}} \tau_d \, C_{s_d,\tau_d}(z_d) \right\} \exp\left( -\frac{1}{2} S_{s,\tau}(z)^\top \Sigma^{-1} S_{s,\tau} \right), \tag{65}$$

where we define the functions

$$C_{s_d,\tau_d}(z_d) := (1 + S^2_{s_d,\tau_d}(z))^{\frac{1}{2}}, \tag{66}$$

and

$$S_{s_d,\tau_d}(z_d) := \sinh(\tau_d \sinh^{-1}(z_d) - s_d), \quad S_{s,\tau}(z) = [S_{s_1,\tau_1}(z_1), \ldots, S_{s_D,\tau_D}(z_D)]^\top. \tag{67}$$

We constructed 3 targets in 2 dimensions and 3 targets in 5 dimensions, each with varying amounts of non-Gaussianity. The details of each target are below. In all experiments, EigenVI was applied with the standardization, where a Gaussian was fit using batch and match VI with a batch size of 16 and a learning rate $\lambda_t = \frac{BD}{t+1}$.

For all experiments, we used a proposal distribution $\pi$ that was uniform on $[-5, 5]^2$.

**2D sinh-arcsinh normal experiment** For $D = 2$ (Figure 4b), we consider the *slight skew and tails target* with parameters $s = [0.2, 0.2], \tau = [1.1, 1.1]$, the *more skew and tails target* with $s = [0.2, 0.5], \tau = [1.1, 1.1]$, and *the slight skew and heavier tails* with $s = [0.2, 0.2], \tau = [1.4, 1.1]$. Note that $s = [0, 0], \tau = [1, 1]$ recovers the multivariate Gaussian. These three target are visualized in Figure 4a.

Table 2: Summary of `posteriordb` models

| Name | Dimension | Model description |
|---|---|---|
| `kidscore` | 3 | linear model with a Cauchy noise prior |
| `sesame` | 3 | linear model with uniform prior |
| `gp_regr` | 3 | Gaussian process regression with squared exponential kernel |
| `garch11` | 4 | generalized autoregressive conditional heteroscedastic model |
| `logearn` | 4 | log-log linear model with multiple predictors |
| `arK-arK` | 7 | autoregressive model for time series |
| `logmesquite` | 7 | multiple predictors log-log model |
| `8-schools` | 10 | non-centered hierarchical model for 8-schools |

**5D sinh-archsinh normal experiment** We constructed three targets $P_1$ (slight skew and tails), $P_2$ (more skew and tails), and $P_3$ (slight skew and heavier tails) each with

$$\Sigma = \begin{bmatrix} 2.2 & 0.3 & 0 & 0 & 0.3 \\ 0.3 & 2.2 & 0 & 0 & 0 \\ 0 & 0 & 2.2 & 0.3 & 0 \\ 0 & 0 & 0.3 & 2.2 & 0 \\ 0.3 & 0 & 0 & 0 & 2.2 \end{bmatrix}. \tag{68}$$

The skew and tail weight parameters used were: $s_1 = [0., 0., 0.2, 0.2, 0.2]; \tau_1 = [1., 1., 1., 1., 1.1]$, $s_2 = [0.0, 0.0, 0.6, 0.4, -0.5]; \tau_2 = [1., 1., 1., 1., 1.1]$, and $s_3 = [0.2, 0.2, 0.2, 0.2, 0.2]; \tau_3 = [1.1, 1.1, 1., 1.4, 1.6]$. See Figure 7 for a visualization of the marginals of each target distribution. In the second row, we show examples of resulting EigenVI fit (visualized using samples from $q$) from $B = 20{,}000$ and $K = 10$.

### E.4. Posteriordb experiments

We consider 8 real data targets from `posteriordb`, a suite of benchmark Bayesian models for real data problems. In Table 2, we summarize the models considered in the study. These target distributions are non-Gaussian, typically with some skew or different tails. To access the log target probability and their gradients, we used the BridgeStan library (Roualdes et al., 2023), which by default transforms the target to be supported on $\mathbb{R}^D$.

For all experiments, we fixed the number of importance samples to be $B = 40{,}000$; to construct the EigenVI matrix $M$, the computations were parallelized over the samples. These experiments were repeated over 5 random seeds, and we report the mean and standard errors in Figure 5; for lower dimensions, there was little variation between runs.

The target distributions were standardized using a Gaussian fit from score matching before applying EigenVI. In most cases, the proposal distribution $\pi$ was chosen to be uniform over $[-6, 6]^D$. For the models `8-schools`, which has a longer tail, we used a multivariate Gaussian proposal with zero mean and a scaled diagonal covariance $\sigma I$, with $\sigma = 3^2$.

For the Gaussian score matching (GSM) method (Modi et al., 2023), we chose a batch size of 16 for all experiments. We generally found the results were not too sensitive in comparison to other batch sizes of 4,8, and 32. For the batch and match (BaM) method (Cai et al., 2024), we chose a batch size of 16. The learning rate was fixed at $\lambda_t = \frac{BD}{t+1}$, which was a recommended schedule for non-Gaussian targets.

For all ELBO optimization methods (full covariance Gaussian family and normalizing flow family), we used Adam to optimize the ELBO. We performed a grid search over the learning rate $0.01, 0.02, 0.05, 0.1$ and batch size $B = 4, 8, 16, 32$. For the normalizing flow model, we used a real NVP (Dinh et al., 2016) with 8 layers and 32 neurons.

In Figure 8 and Figure 9, we show the corner plots that compare an EigenVI fit, a normalizing flow fit, and a Gaussian fit (BaM). In each plot, we plot the samples from the variational distribution against samples from Hamiltonian Monte Carlo. We observe that the two more expressive families EigenVI and the normalizing flow are able to model the tails of the distribution better than the Gaussian fit.

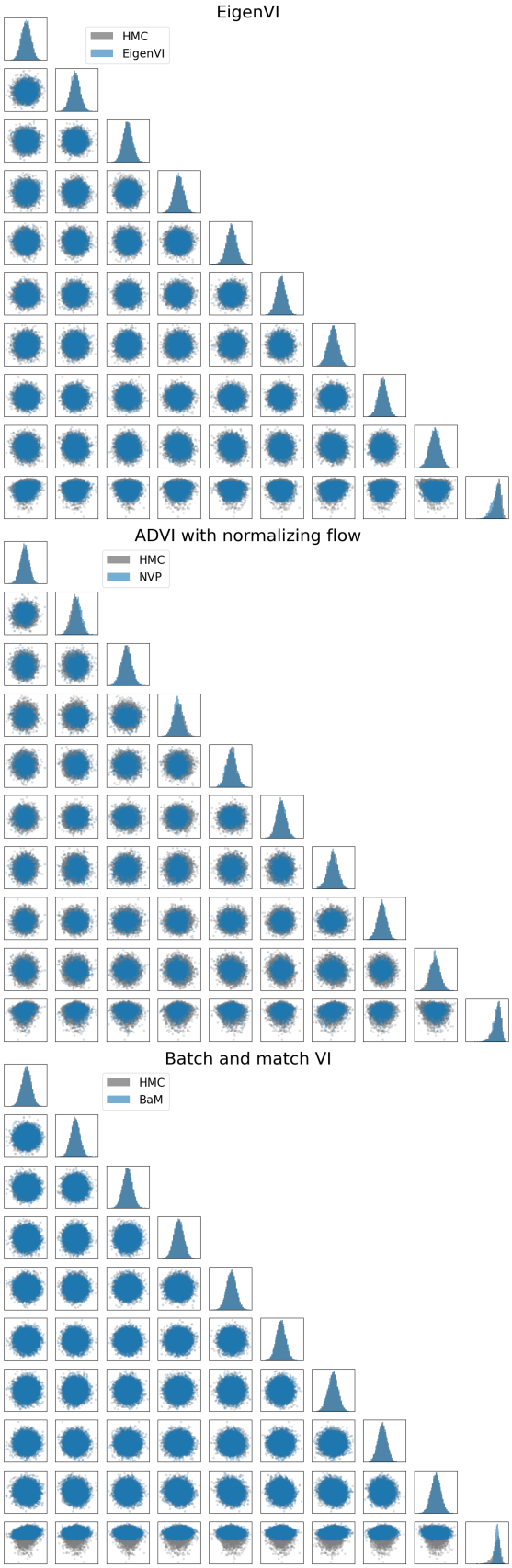



Figure 8: Comparison of EigenVI, normalizing flow, and Gaussian score-based BBVI methods on `8schools`.

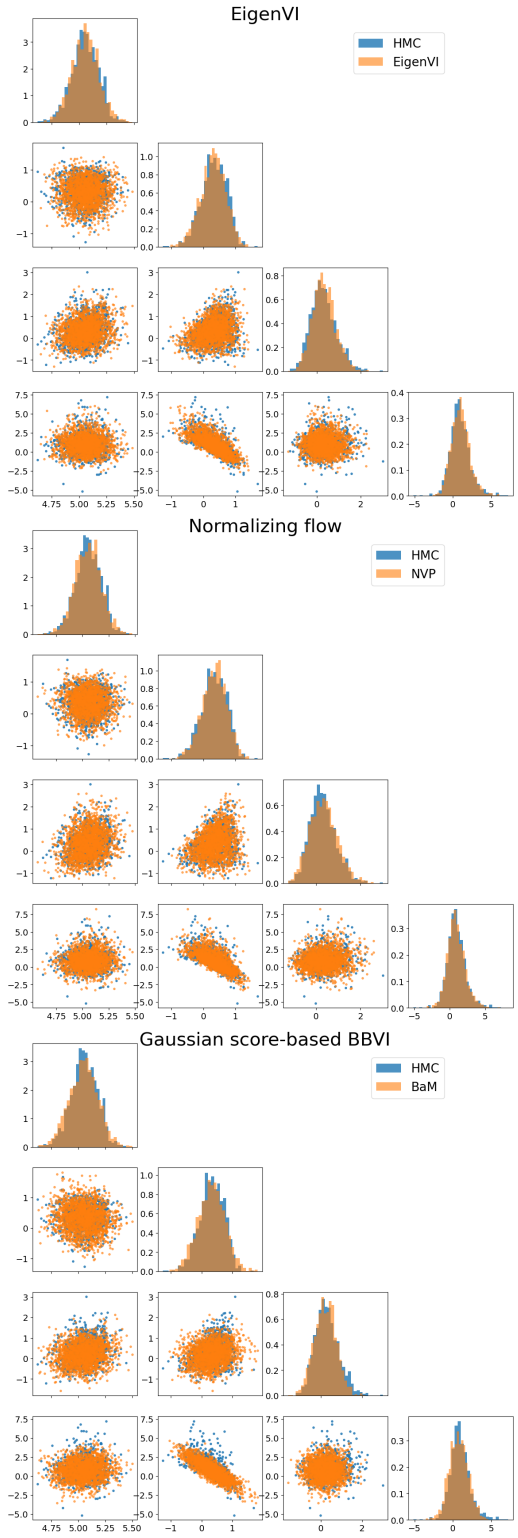



Figure 9: Comparison of EigenVI, normalizing flow, and Gaussian score-based BBVI methods on `garch11`. Note that the Gaussian approximation over/underestimates the tails, while the more expressive families fit the tails better.