Symmetrized Variational Inference

David A. Moore
Computer Science Division
University of California, Berkeley
dmoore@cs.berkeley.edu

Abstract

We introduce a framework for modeling parameter symmetries in variational inference by explicitly mixing a base approximating density over a symmetry group. We show that this can be done tractably for the case of a Gaussian mixture over the orthogonal group under an isotropic variance assumption. Initial results show that inference with a symmetrized posterior avoids component collapse and leads to improved predictive performance.

1 Introduction

Many probability models commonly used in machine learning are not, strictly speaking, identifiable: they exhibit *parameter symmetries* in which the model density is invariant under some class of transformations of the latent parameters. The resulting multimodal posteriors are not captured well by most inference procedures: "label switching" is a well known problem in sampling mixture models (Neal, 1999; Celeux et al., 2000), while algorithms such as EP that attempt to match marginals may miss crucial joint structure (Nishihara et al., 2013).

It is sometimes thought that variational inference with a mode-seeking divergence such as $\mathrm{KL}[q\|p]$ breaks symmetries by modeling only a single mode of the posterior. However, symmetric modes are not isolated in latent space: nearby modes can bleed probability mass into each other to create new modes corresponding to degenerate solutions. This "implicit regularization" results in approximate inference failing to use a model's full representational capacity. This effect has been noted for Gaussian mixture models (MacKay, 2001) and analyzed extensively in the case of matrix factorization (Nakajima and Sugiyama, 2010; Nakajima et al., 2013). An analogous phenomenon may be observed in component collapse of variational autoencoders (Dinh and Dumoulin, 2014; Burda et al., 2015).

We propose modeling symmetries directly in variational inference using a *symmetrized posterior* formed by explicitly mixing a "base" distribution over the relevant symmetry group, so that our approximating class captures the same symmetries as the true posterior (Figures 1 and 2). This paper presents our general framework and demonstrates its application to signflip and orthogonal symmetries in matrix factorization models, with promising initial results.

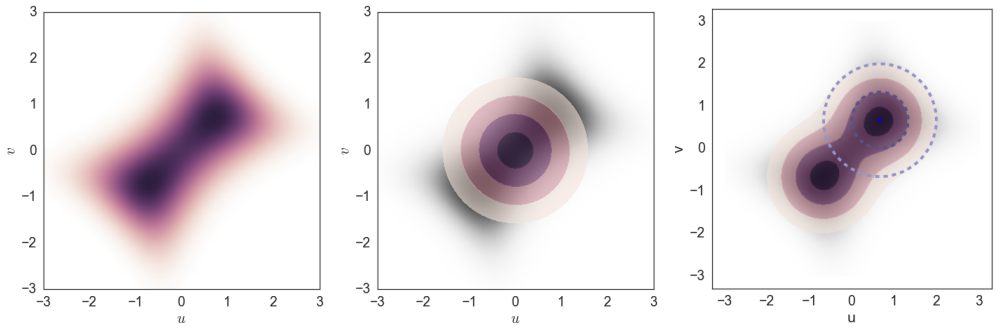
2 Matrix factorization and implicit regularization

We focus for concreteness on matrix factorization, although both the problem of parameter symmetries and our proposed solution framework are more general. We consider the model

$$\mathbf{U}, \mathbf{V} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$
$$\mathbf{R} \sim \mathcal{N}(\mathbf{U}\mathbf{V}^T/\sqrt{k}, \sigma_n^2 \mathbf{I})$$

in which $\mathbf{U} : n \times k$ and $\mathbf{V} : m \times k$ are latent trait matrices, representing n users and m movies (or other items) each described by k features, and $\mathbf{R}$ is an noisy ratings matrix. We assume $\mathbf{R}$ is fully observed;

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

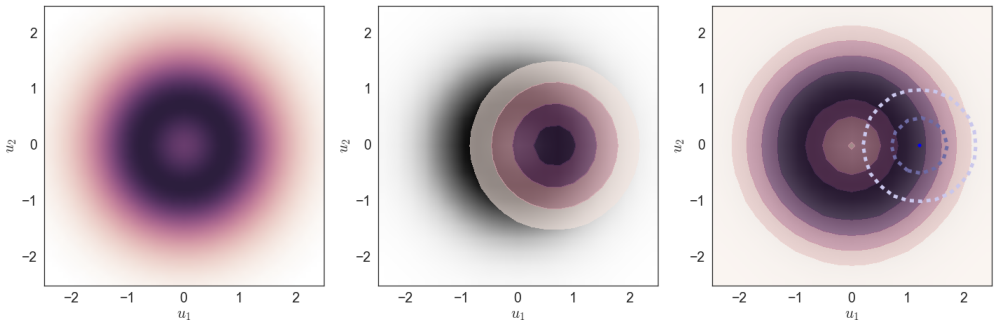


(a) True posterior $p(u, v|r)$ has signflip symmetry.

(b) Mean-field Gaussian approximation fits a degenerate mode at the origin.

(c) Symmetrized approximation $\tilde{q}$ mixing $q^*(u, v)$ (dotted) and $q^*(-u, -v)$.

Figure 1: Posteriors from the scalar factorization model $r = uv + \epsilon$; $u, v \in \mathbb{R}$, given observed $r = 1.5$. Implicit regularization results from the mean-field posterior attempting to cover both modes, ending up at the origin. The symmetrized posterior explicitly corrects this effect.



(a) True posterior $p(\mathbf{u}|r)$ yielding predictive mean $\mathbb{E}[\mathbf{u}\mathbf{v}^T|\mathbf{r}] = 1.30$.

(b) Mean-field approximation is pulled towards the origin, $\mathbb{E}[\mathbf{u}\mathbf{v}^T|\mathbf{r}] = 0.44$.

(c) Symmetrized approximation $\tilde{q}$, from base Gaussian q^* (dotted), $\mathbb{E}[\mathbf{u}\mathbf{v}^T|\mathbf{r}] = 1.38$.

Figure 2: Marginal posteriors $p(\mathbf{u}|r)$ from the overparameterized scalar model $r = \mathbf{u}\mathbf{v}^T + \epsilon$; $\mathbf{u}, \mathbf{v} \in \mathbb{R}^{1\times 2}$, given observed $r = 2$. This rotational symmetry is a special case of the more general matrix factorization symmetry in the joint column space of $\mathbf{U}$ and $\mathbf{V}$.

the inference problem is to recover the "true" low-rank ratings matrix $\mathbf{U}\mathbf{V}^T$ given noisy observations. Note that this model is subject to the posterior symmetry $p(\mathbf{U}, \mathbf{V}|\mathbf{R}) = p(\mathbf{U}\mathbf{T}, \mathbf{V}\mathbf{T}|\mathbf{R})$ where $\mathbf{T} \in \mathrm{O}(k)$ is any $k \times k$ orthogonal matrix. This includes as special cases a permutation symmetry between the latent columns, as well as signflip symmetries on each column.

Recent analysis by Nakajima et al. (2013) has obtained an analytic solution to the mean-field variational objective in this model. They show that latent traits obtained through variational and even MAP inference shrink each singular value by a factor that depends on the observation noise σ_n^2, so that all singular values below some threshold are zeroed out: the model effectively uses fewer traits than it was allotted. This "implicit regularization" arises directly from the use of approximate inference; it can be verified in simple cases that the true Bayesian posterior does not exhibit the same shrinkage (Figure 3).

We demonstrate that implicit regularization can be avoided by explicitly representing parameter symmetries in the variational posterior. Our symmetrized posteriors follow the true Bayes posterior in that they use the full model capacity with no unwanted shrinkage (figs. 3 and 4a); initial experiments suggest they also improve prediction quality. The next section describes our general framework, which we then specialize to model the orthogonal symmetries that arise in matrix factorization.

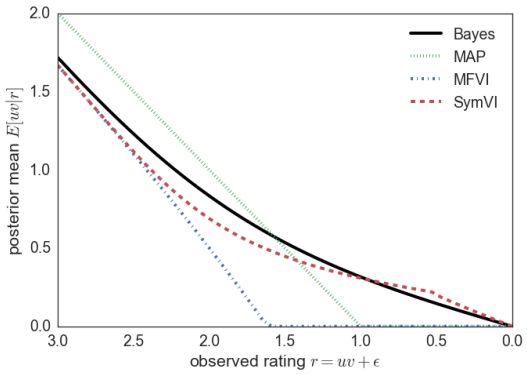


Figure 3: Predictive mean $\mathbb{E}[uv|r]$ in the scalar factorization model $r = uv+\epsilon$, with $u, v, \epsilon \sim \mathcal{N}(0, 1)$. Standard mean field (MFVI) and even MAP are subject to degenerate solutions at zero for small r, while inference with a symmetrized posterior invariant to signflips (SymVI) more closely recovers the Bayes predictive mean.

3 Symmetrized Variational Inference

We consider probability models of the form $p(\mathbf{x}, \mathbf{z})$ where $\mathbf{x}$ and $\mathbf{z}$ are observed and latent variables, and perform inference by minimizing the exclusive divergence $KL[q\|p]$ between an approximate posterior $q_\theta(\mathbf{z})$ and the true posterior $p(\mathbf{z}|\mathbf{x})$. This is equivalent to maximizing a lower bound on the log model evidence, known as the Evidence Lower Bound or ELBO (Bishop, 2006),

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{z}\sim q_\theta}\left[\log p(\mathbf{x}, \mathbf{z}) - \log q_\theta(\mathbf{z})\right] \leq \log p(\mathbf{x}), \tag{1}$$

which can equivalently be written in terms of the entropy $\mathcal{H}(q_\theta)$ of the approximating posterior:

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{z}\sim q_\theta}\left[\log p(\mathbf{x}, \mathbf{z})\right] + \mathcal{H}(q_\theta). \tag{2}$$

The so-called reparameterization trick enables practical inference via gradient-based stochastic optimization of $\mathcal{L}(\theta)$, as long as the sampling process $\mathbf{z} \sim q_\theta$ can be expressed as a differentiable transformation $\mathbf{z} = f_\theta(\epsilon)$ of a random source ϵ. (Kingma and Welling, 2013; Kucukelbir et al., 2016).

We are interested in cases where the model density in the latent space is invariant under the action of transformations from a group $\mathfrak{G}$, so that $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{x}, \mathbf{Tz}) \;\; \forall\, \mathbf{T} \in \mathfrak{G}$. We propose to exploit this structure by imposing the same invariance on our variational posterior.

We define a *symmetrized posterior* $\tilde{q}_\theta$ by a two-step sampling process: first sample $\mathbf{z}^*$ from some base posterior q^*_θ, then apply a random transformation $\mathbf{T}$ to generate $\mathbf{z} = \mathbf{Tz}^*$. Formally we write the density of $\tilde{q}_\theta$ as an explicit mixture with respect to the uniform (Haar) measure $V(\mathbf{T})$, given by

$$\tilde{q}_\theta(\mathbf{z}) = \int_{\mathbf{T}\in\mathfrak{G}} q^*_\theta(\mathbf{T}^{-1}\mathbf{z}) \left|\mathbf{T}^{-1}\right| dV(\mathbf{T}) \tag{3}$$

where $\left|\mathbf{T}^{-1}\right|$ is the Jacobian determinant (unity for orthogonal transformations). It is clear that this respects the symmetry $\tilde{q}_\theta(\mathbf{z}) = \tilde{q}_\theta(\mathbf{Tz})$ for all $\mathbf{T} \in \mathfrak{G}$.

Plugging $\tilde{q}_\theta$ into the ELBO, we note that expectations under $\tilde{q}_\theta$ can be written as a nested expectation over a base sample $\mathbf{z}^*$ and transformation $\mathbf{T}$:

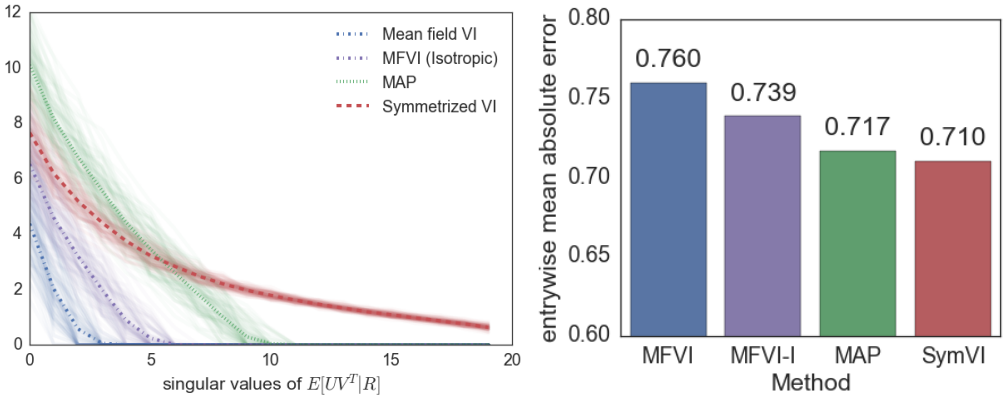
$$\begin{aligned}\mathcal{L}(\tilde{q}_\theta) &= \mathbb{E}_{\mathbf{z}\sim\tilde{q}_\theta}\left[\log p(\mathbf{x}, \mathbf{z}) - \log \tilde{q}_\theta(\mathbf{z})\right] \\ &= \mathbb{E}_{\mathbf{z}^*\sim q^*_\theta}\left[\mathbb{E}_{\mathbf{T}}\left[\log p(\mathbf{x}, \mathbf{Tz}^*) - \log \tilde{q}_\theta(\mathbf{Tz}^*)\right]\right]\end{aligned}$$

Since the model density p is invariant to $\mathbf{T}$ by assumption, and the symmetrized posterior $\tilde{q}_\theta$ is also invariant by construction, the expectation over $\mathbf{T}$ is vacuous. Dropping it yields the objective

$$\mathcal{L}(\tilde{q}_\theta) = \mathbb{E}_{\mathbf{z}^*\sim q^*_\theta}\left[\log p(\mathbf{x}, \mathbf{z}^*) - \log \tilde{q}_\theta(\mathbf{z}^*)\right], \tag{4}$$

which differs from (1) only in that we are now evaluating the mixture density $\log \tilde{q}_\theta$ in place of the original $\log q_\theta$. By straightforward manipulation we obtain an equivalent form,

$$\mathcal{L}(\tilde{q}_\theta) = \mathbb{E}_{\mathbf{z}^*\sim q^*_\theta}\left[\log p(\mathbf{x}, \mathbf{z}^*)\right] + \mathcal{H}(q^*_\theta) + KL[q^*_\theta\|\tilde{q}_\theta], \tag{5}$$



(a) Singular values of predictive mean $\mathbb{E}[\mathbf{U}\mathbf{V}^T|\mathbf{R}]$.

(b) Mean absolute entrywise reconstruction error $\|\mathbb{E}[\mathbf{U}\mathbf{V}^T|\mathbf{R}] - \mathbf{U}\mathbf{V}^T\|_1$.

Figure 4: Inference on 80 synthetic matrices sampled with $n = m = 40$, $k = 20$, and $\sigma_n = 2$. Orthogonally symmetrized VI produces full-rank posteriors (left) that yield more accurate predictions than MFVI or MAP (right).

which is equal to (2) plus a nonnegative term $KL[q^*_\theta \| \tilde{q}_\theta]$, so this bound is at least as tight as (2). The KL term measures the additional information needed to code a sample from the symmetrized $\tilde{q}_\theta$ given a sample from q^*_θ; in the case of well-separated discrete modes, this is simply the the (log) number of modes.[1] When modes overlap, it measures the discrepancy between q^* and its symmetrized modes, going to zero in the case where q^* is already symmetric (e.g., an isotropic Gaussian at the origin). The symmetrized bound (5) therefore discourages q^* from straddling multiple modes, so that it instead attempts to capture a single mode as well as possible.

3.1 Signflip symmetry

The framework described above is applicable generally to any base family q^*_θ and symmetry group $\mathfrak{G}$; the challenge in practice is to efficiently compute or bound the symmetrized divergence $KL[q^*_\theta \| \tilde{q}_\theta] = \mathbb{E}[\log q^*_\theta(\mathbf{z}^*) - \log \tilde{q}_\theta(\mathbf{z}^*)]$. Under a Monte Carlo approximation the challenge reduces to computing the log density $-\log \tilde{q}_\theta(\mathbf{z}^*)$ at values $\mathbf{z}^*$ sampled from the base posterior. In the simple case of signflip symmetries, this can be done tractably with an explicit sum $\log \tilde{q}_\theta(\mathbf{z}) = \log \frac{1}{2}\left(q^*_\theta(\mathbf{z}) + q^*_\theta(-\mathbf{z})\right)$.

Signflip symmetries arise in the special case of matrix factorization where all quantities are scalar (Figure 1). In this case it is also tractable to numerically compute the Bayes predictive mean. Figure 3 compares predictive means as a function of the observed scalar value r, showing that while standard methods exhibit degenerate regularization, predictions using the symmetrized posterior track the true Bayes prediction much more closely.

3.2 General orthogonal symmetry

Our main technical contribution is an efficient approximation to the continuous mixture density

$$\log \tilde{q}(\mathbf{X}) = \log \int_{\mathbf{T} \in \mathrm{O}(k)} \mathcal{N}(\mathbf{X}; \mathbf{M}\mathbf{T}, \mathbf{T}^T \mathbf{\Sigma} \mathbf{T}) dV(\mathbf{T})$$

in which the columns of an elementwise matrix Gaussian $q^*(\mathbf{X}) = \mathcal{N}(\mathbf{X}; \mathbf{M}, \mathbf{\Sigma})$ are mixed over the orthogonal group $\mathrm{O}(k)$ (Figure 2). For space considerations we defer details to Appendix A. Our current analysis assumes an isotropic covariance ($\mathbf{\Sigma} = c\mathbf{I}$) which we hope to relax in future work.

Informally, any rotation $\mathbf{T}$ acting within the nullspace of $\mathbf{M}$ generates a mixture component $\mathbf{M}\mathbf{T} = \mathbf{M}$ equivalent to the base distribution; such rotations are "wasted" in the sense that they do not

[1] It is common to adjust naïve evidence bounds by a symmetry-counting term, e.g., $\log 2^n$ for signflip symmetries or $\log n!$ for permutation symmetries, but this is incorrect when the approximate posterior straddles multiple modes. The KL term in (5) can be seen as the generally correct form of this adjustment, which always yields a valid evidence bound.

increase the mixture entropy $\mathcal{H}(\tilde{q})$ relative to the base $\mathcal{H}(q)$. Inference using the mixture entropy $\mathcal{H}(\tilde{q})$ therefore penalizes the nullspace dimension of the posterior mean, avoiding component collapse by encouraging nonzero singular values. This effect is seen in Figure 4a, which plots singular values of the predictive mean $\mathbb{E}[\mathbf{U}\mathbf{V}^T|\mathbf{R}]$ from a relatively noisy ($\sigma_n = 2$) synthetic 20-trait model. While the mean field VI and MAP solutions (and MFVI with isotropic covariance, included for comparison) shrink many traits to zero, symmetrized VI uses the model's full capacity. Figure 4b shows that this is reflected in predictive performance: symmetrized VI yields better reconstructions of the true generative ratings $\mathbf{U}\mathbf{V}^T$ than the implicitly-regularized MAP and (especially) mean field estimates.

4 Future work

These preliminary results demonstrate that explicitly accounting for parameter symmetries can yield better inferences. In future work we hope to consider symmetrization of non-isotropic Gaussians and more flexible approximating classes (Salimans et al., 2015; Rezende and Mohamed, 2015; Tran et al., 2016), as well as permutation, translation, and scaling symmetries. We are especially interested in the use of symmetrized posteriors for automatic relevance determination (ARD) and model selection, as well as extensions to stochastic inference in "deep" models such as variational autoencoders.

Acknowledgements

This work was inspired by (though not performed during) an internship with the Infer.NET team at Microsoft Research Cambridge, and in particular by many illuminating discussions with Tom Minka. The author is supported by the Defense Threat Research Agency (DTRA) under grant #HDTRA-1111-0026.

References

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*.

Burda, Y., Grosse, R., and Salakhutdinov, R. (2015). Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*.

Butler, R. W. and Wood, A. T. (2003). Laplace approximation for bessel functions of matrix argument. *Journal of Computational and Applied Mathematics*, 155(2):359–382.

Celeux, G., Hurn, M., and Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association*, 95(451):957–970.

Dinh, L. and Dumoulin, V. (2014). Training neural bayesian nets. `http://www.iro.umontreal.ca/~bengioy/cifar/NCAP2014-summerschool/slides/Laurent_dinh_cifar_presentation.pdf`. CIFAR NCAP summer school presentation.

Herz, C. S. (1955). Bessel functions of matrix argument. *Annals of Mathematics*, pages 474–523.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. (2016). Automatic differentiation variational inference. *arXiv preprint arXiv:1603.00788*.

MacKay, D. J. (2001). Local minima, symmetry-breaking, and model pruning in variational free energy minimization. `http://www.inference.phy.cam.ac.uk/mackay/minima.pdf`. Inference Group, Cavendish Laboratory, Cambridge, UK.

Muirhead, R. J. (1982). *Aspects of multivariate statistical theory*. John Wiley & Sons.

Nakajima, S. and Sugiyama, M. (2010). Implicit regularization in variational bayesian matrix factorization. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 815–822.

Nakajima, S., Sugiyama, M., Babacan, S. D., and Tomioka, R. (2013). Global analytic solution of fully-observed variational bayesian matrix factorization. *Journal of Machine Learning Research*, 14(Jan):1–37.

Neal, R. M. (1999). Erroneous results in "marginal likelihood from the gibbs output". `http://www.cs.utoronto.ca/~radford/chib-letter.html`. University of Toronto.

Nishihara, R., Minka, T., and Tarlow, D. (2013). Detecting parameter symmetries in probabilistic models. *arXiv preprint arXiv:1312.5386*.

Rezende, D. and Mohamed, S. (2015). Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1530–1538.

Salimans, T., Kingma, D. P., Welling, M., et al. (2015). Markov chain monte carlo and variational inference: Bridging the gap. In *International Conference on Machine Learning*, pages 1218–1226.

Tran, D., Ranganath, R., and Blei, D. M. (2016). The variational gaussian process. In *International Conference on Learning Representations (ICLR)*.

A Orthogonal Gaussian mixture densities

We consider as a base density the elementwise (mean-field) Gaussian

$$q^*(\mathbf{X}) = \mathcal{N}(\mathbf{X}; \mathbf{M}, \mathbf{S}) = \prod_{i=1}^{n+m} \prod_{j=1}^{k} \mathcal{N}(x_{ij}; m_{ij}, s_{ij})$$

in which we let $\mathbf{M}, \mathbf{S}$ contain elementwise means and variances respectively for a matrix-valued variable $\mathbf{X}$; in matrix factorization we will let $\mathbf{X} = \begin{bmatrix} \mathbf{U} \\ \mathbf{V} \end{bmatrix} \in \mathbb{R}^{(n+m)\times k}$ be the stacked matrix of latent traits. We will focus on the restrictive case where each row is described by the same diagonal covariance $\mathbf{\Sigma} \in \mathbb{R}^{k\times k}$, so that the density can be written

$$q^*(\mathbf{X}) = \mathcal{N}(\mathbf{X}; \mathbf{M}, \mathbf{\Sigma}) \propto \exp\left\{-\frac{1}{2}\mathrm{Tr}\left[(\mathbf{X}-\mathbf{M})\mathbf{\Sigma}^{-1}(\mathbf{X}-\mathbf{M})^T\right]\right\};$$

below we will restrict this further to the case where $\mathbf{\Sigma} = c\mathbf{I}$ is isotropic. We hope to relax both of these assumptions in future work.

We consider the symmetrized approximate posterior

$$\tilde{q}(\mathbf{X}) = \int_{\mathbf{T}\in \mathrm{O}(k)} \mathcal{N}(\mathbf{X}\mathbf{T}^T; \mathbf{M}, \mathbf{\Sigma}) dV(\mathbf{T}) = \int_{\mathbf{T}\in \mathrm{O}(k)} \mathcal{N}(\mathbf{X}; \mathbf{M}\mathbf{T}, \mathbf{T}^T\mathbf{\Sigma}\mathbf{T}) dV(\mathbf{T})$$

which is a continuous mixture of Gaussians with means and (co)variances corresponding to rotations within the column space of $\mathbf{M}$. For symmetrized inference we must compute the divergence

$$\begin{aligned} KL[q^*_\theta \| \tilde{q}_\theta] &= \mathbb{E}_{q^*}\left[\log \frac{q^*_\theta(\mathbf{X})}{\tilde{q}_\theta(\mathbf{X})}\right] \\ &= \mathbb{E}_{q^*}\left[-\log \int_{\mathbf{T}\in \mathrm{O}(k)} \frac{\mathcal{N}(\mathbf{X}\mathbf{T}^T; \mathbf{M}, \mathbf{\Sigma})}{\mathcal{N}(\mathbf{X}; \mathbf{M}, \mathbf{\Sigma})} dV(\mathbf{T})\right]. \end{aligned} \tag{6}$$

We approximate the expectation by Monte Carlo, so that it remains only to evaluate the interior integral $r_\theta(\mathbf{X}) = -\log \int_{\mathbf{T}\in \mathrm{O}(k)} \frac{\mathcal{N}(\mathbf{X}\mathbf{T}^T; \mathbf{M}, \mathbf{\Sigma})}{\mathcal{N}(\mathbf{X}; \mathbf{M}, \mathbf{\Sigma})} dV(\mathbf{T})$. Examining this more closely,

$$\begin{aligned} r_\theta(\mathbf{X}) &= -\log \int_{\mathbf{T}\in \mathrm{O}(k)} e^{-\frac{1}{2}\mathrm{Tr}\left[(\mathbf{X}\mathbf{T}^T-\mathbf{M})\mathbf{\Sigma}^{-1}(\mathbf{X}\mathbf{T}^T-\mathbf{M})^T\right] + \frac{1}{2}\mathrm{Tr}\left[(\mathbf{X}-\mathbf{M})\mathbf{\Sigma}^{-1}(\mathbf{X}-\mathbf{M})^T\right]} dV(\mathbf{T}) \\ &= -\log \int_{\mathbf{T}\in \mathrm{O}(k)} e^{-\frac{1}{2}\left(\mathrm{Tr}\left[\mathbf{X}^T\mathbf{X}\left(\mathbf{T}^T\mathbf{\Sigma}^{-1}\mathbf{T}-\mathbf{\Sigma}^{-1}\right)\right]\right)} e^{\mathrm{Tr}\left[\mathbf{\Sigma}^{-1}\frac{\mathbf{M}^T\mathbf{X}+\mathbf{X}^T\mathbf{M}}{2}(\mathbf{T}^T-\mathbf{I})\right]} dV(\mathbf{T}) \\ &= -\log \int_{\mathbf{T}\in \mathrm{O}(k)} f(\mathbf{T}) g(\mathbf{T}) dV(\mathbf{T}) \end{aligned}$$

we find that it is effectively an inner product of two functions f, g over the orthogonal group. The first factor, $f(\mathbf{T}) = \exp\left\{-\frac{1}{2}\left(\mathrm{Tr}\left[\mathbf{X}^T\mathbf{X}\left(\mathbf{T}^T\boldsymbol{\Sigma}^{-1}\mathbf{T} - \boldsymbol{\Sigma}^{-1}\right)\right]\right)\right\}$, measures the discrepancy between the precision matrix $\boldsymbol{\Sigma}^{-1}$ and its transformed counterpart $\mathbf{T}^T\boldsymbol{\Sigma}^{-1}\mathbf{T}$, while the second factor, $g(\mathbf{T}) = \exp\left\{\mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}\frac{\mathbf{M}^T\mathbf{X}+\mathbf{X}^T\mathbf{M}}{2}(\mathbf{T}^T - \mathbf{I})\right]\right\}$, measures the extent to which a given transformation $\mathbf{T}$ aligns the observed value $\mathbf{X}$ with the mean $\mathbf{M}$. Note that f vanishes in the case of isotropic covariance $\boldsymbol{\Sigma} = c\mathbf{I}$, since this implies $\mathbf{T}^T\boldsymbol{\Sigma}^{-1}\mathbf{T} = \boldsymbol{\Sigma}^{-1}$. In this case we have

$$\begin{aligned} r_\theta(\mathbf{X}) &= -\log \int_{\mathbf{T}\in\mathrm{O}(k)} g(\mathbf{T}) dV(\mathbf{T}) \\ &= -\log \int_{\mathbf{T}\in\mathrm{O}(k)} \exp\left\{\mathrm{Tr}\left[\frac{1}{2c}(\mathbf{M}^T\mathbf{X} + \mathbf{X}^T\mathbf{M})(\mathbf{T}^T - \mathbf{I})\right]\right\} dV(\mathbf{T}), \end{aligned}$$

and letting $\mathbf{A} = \frac{1}{2c}(\mathbf{M}^T\mathbf{X} + \mathbf{X}^T\mathbf{M})$, this simplifies to

$$\begin{aligned} &= -\log \int_{\mathbf{T}\in\mathrm{O}(k)} \exp\left\{\mathrm{Tr}\left[\mathbf{A}\mathbf{T}^T\right] - \mathrm{Tr}\left[\mathbf{A}\right]\right\} dV(\mathbf{T}) \\ &= \mathrm{Tr}\left[\mathbf{A}\right] - \log {}_0F_1\left[\frac{k}{2}; \frac{1}{4}\mathbf{A}\mathbf{A}^T\right] \end{aligned}$$

where the hypergeometric function ${}_0F_1\left[\frac{k}{2}; \frac{1}{4}\mathbf{A}\mathbf{A}^T\right] = \int_{\mathbf{T}\in\mathrm{O}(k)} \exp\left\{\mathrm{Tr}\left[\mathbf{A}\mathbf{T}^T\right]\right\} dV(\mathbf{T})$ is a form of matrix-argument Bessel function (Herz, 1955), also arising in the analysis of the non-central Wishart distribution (Muirhead, 1982). Notably it depends on $\mathbf{A}$ only through its singular values $(\sigma_i)_{i=1}^k$. This conforms with the intuition, stated above, that the effect of inference with an orthogonal mixture of Gaussians should be to encourage nonzero singular values in $\mathbf{M}$ (and, by extension, $\mathbf{A}$).

Butler and Wood (2003) derive a Laplace approximation to ${}_0F_1$, given by

$${}_0F_1\left[\frac{k}{2}; \frac{1}{4}\mathbf{A}\mathbf{A}^T\right] \approx \frac{\prod_{i=1}^k (1-\hat{y}_i^2)^{k/2} e^{\sigma_i \hat{y}_i}}{\sqrt{\prod_{i=1}^k \prod_{j=1}^k (1-\hat{y}_i^2\hat{y}_j^2)}} \tag{7}$$

for $\hat{y}_i = 2\sigma_i / (k\sqrt{4\sigma_i^2/k^2 + 1} + 1)$; in their evaluations this approximation demonstrates "very high accuracy in a variety of settings". It can be implemented stably in the log domain as a differentiable function of the singular values σ.

We use use the Laplace approximation (7) to implement the divergence (6), with automatic gradients computed using TensorFlow. Code is given in Listing 1. Although TensorFlow does not implement gradients for the SVD operator $\mathbf{A} = \mathbf{U}\mathrm{diag}(\sigma)\mathbf{V}^T$, we exploit the fact that $\mathbf{U}\mathrm{diag}(\sigma) = \mathbf{A}\mathbf{V}$, allowing us to approximate gradients $\frac{\partial\sigma(\mathbf{A})}{\partial\mathbf{A}}$ by differentiating through the column norms of $\mathbf{A}\mathbf{V}$ holding fixed the singular vectors $\mathbf{V}$. Note that $\mathbf{A}$ is a $k \times k$ matrix, where the trait dimension $k \ll n, m$ does not depend on the data size, so computing the SVD on each gradient update is relatively cheap. In practice we do not observe significant slowdowns from inference under the symmetrized $\tilde{q}$ relative to the base posterior q^*.

```
def gaussian_orthog_stochastic_kl(X, M, c):
    # X, M: 2D Tensors of matching dimensions
    # c: scalar (isotropic) variance
    # returns: stochastic estimate of KL[ q* | \tilde{q} ] for
    #   q* ~ N(X; M, cI) under orthogonal symmetry

    tmp = tf.matmul(tf.transpose(X), M)
    A = (tmp + tf.transpose(tmp) ) / (2*c)
    svs = differentiable_singular_vals(A)
    r =  tf.trace(A) - log_bessel(svs, k)
    return r

def differentiable_singular_vals(A):
    # returns singular vals of A with approximate gradients.
    d, u, v = tf.svd(A)
    ud = tf.matmul(A, tf.stop_gradient(v))
    return tf.sqrt(tf.reduce_sum(tf.square(ud), 0))

def log_bessel(svs, n):
    def r(u):
          return u/(tf.sqrt(tf.square(u) + 1.0) + 1.0)
    ys = r(2.0*svs / n)
    y2 = tf.square(ys)
    y2r = tf.reshape(y2, (n, 1))
    y2pairs = tf.matmul( y2r, tf.transpose(y2r))
    log_denom = .5 * tf.reduce_sum(tf.log(1-y2pairs))
    log_num = tf.reduce_sum(svs*ys + n/2.0 * tf.log((1-y2)))
    return log_num - log_denom
```

Listing 1: TensorFlow implementation of orthogonally symmetrized Gaussian log density.