Symmetrized Variational Inference

Dave Moore, UC Berkeley

Berkeley
Artificial Intelligence Research Laboratory

Overview

- Widely used probabilistic models (matrix factorization, VAEs) contain *parameter symmetries* that cause approximate inference to underfit.
- We model this effect by fitting an explicitly *symmetrized* approximate posterior.
- Initial results show that this improves predictions and avoids underfitting.

Approximate Inference -> Implicit Regularization

Intuition: larger models should better fit / capture structure in data.

Observation: in practice, variational autoencoders refuse to use extra hidden units ("component collapse").

Theory: can show that variational matrix factorization (linear analogue of VAEs) ignores extra hidden units, shrinks small singular values to zero. (Nakajima et al., 2013)

This **unwanted (implicit) regularization** is caused by approximate inference – it is not present in the true Bayesian posterior!

Symmetrized Posteriors

Classic VI: fit approximate posterior q by minimizing KL[q | p], equivalent to maximizing an evidence lower bound (ELBO)

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

Given base posterior q*, we define the *symmetrized posterior* $\tilde{q}$ as a uniform mixture under transformations from group **G:**

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

Sampling interpretation: first draw **z*** ~ q*, then apply a (uniformly chosen) random transformation to sample **z = Tz***.

The symmetrized posterior $\tilde{q}$ matches symmetries of the true posterior; yields a tighter evidence bound:

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

To apply: need to compute/approximate $KL[\,q^*|\tilde{q}]$ for specific symmetry group. Can do this for Gaussian q* under orthogonal group O(k), matching matrix factorization/VAE symmetries.

Orthogonally symmetrized Gaussians

Symmetrizing the column space of an elementwise Gaussian matrix over the orthogonal group yields a continuous mixture of Gaussians:

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

which decomposes as

$$= -\mathbb{E}_{q^*}\left[\log\int_{\mathbf{T}\in O(k)} \exp\left\{-\frac{1}{2}\mathrm{Tr}\left[\mathbf{X}^T\mathbf{X}\overbrace{(\mathbf{T}^T\mathbf{\Sigma}^{-1}\mathbf{T}-\mathbf{\Sigma}^{-1})}^{\text{vanishes for isotropic }\Sigma}\right] + \frac{1}{2}\mathrm{Tr}\left[\mathbf{\Sigma}^{-1}(\mathbf{M}^T\mathbf{X}+\mathbf{X}^T\mathbf{M})(\mathbf{T}^T-\mathbf{I})\right]\right\} dV(\mathbf{T})\right]$$

Taking $\mathbf{A} = \frac{1}{2}\mathbf{\Sigma}^{-1}(\mathbf{M}^T\mathbf{X}+\mathbf{X}^T\mathbf{M})$, this simplifies to

$$= \mathbb{E}_{q^*}\left[-\log\int_{\mathbf{T}\in O(k)} \mathrm{etr}\left[\mathbf{A}\mathbf{T}^T - \mathbf{A}\right]\right]$$

$$= \mathbb{E}_{q^*}\left[\mathrm{Tr}\left[\mathbf{A}\right] - \log {}_0F_1\left[\frac{k}{2};\frac{1}{4}\mathbf{A}\mathbf{A}^T\right]\right]$$

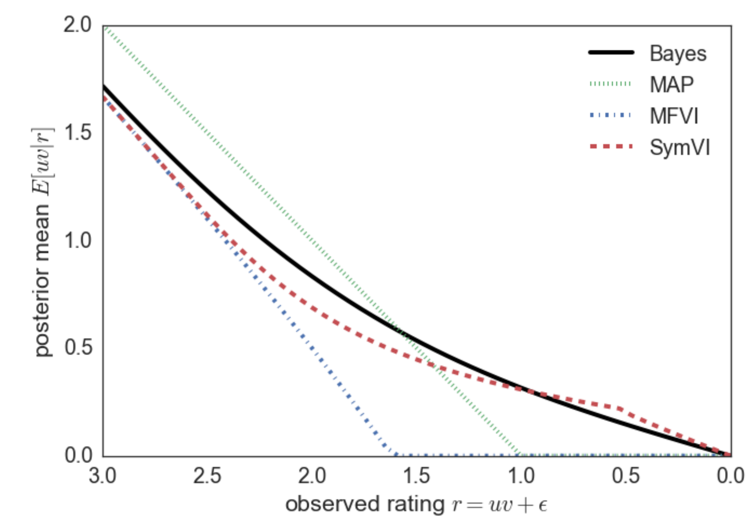
where the hypergeometric function ${}_0F_1$ depends only on singular values of **A** and can be efficiently Laplace-approximated (Butler & Wood, 2003).

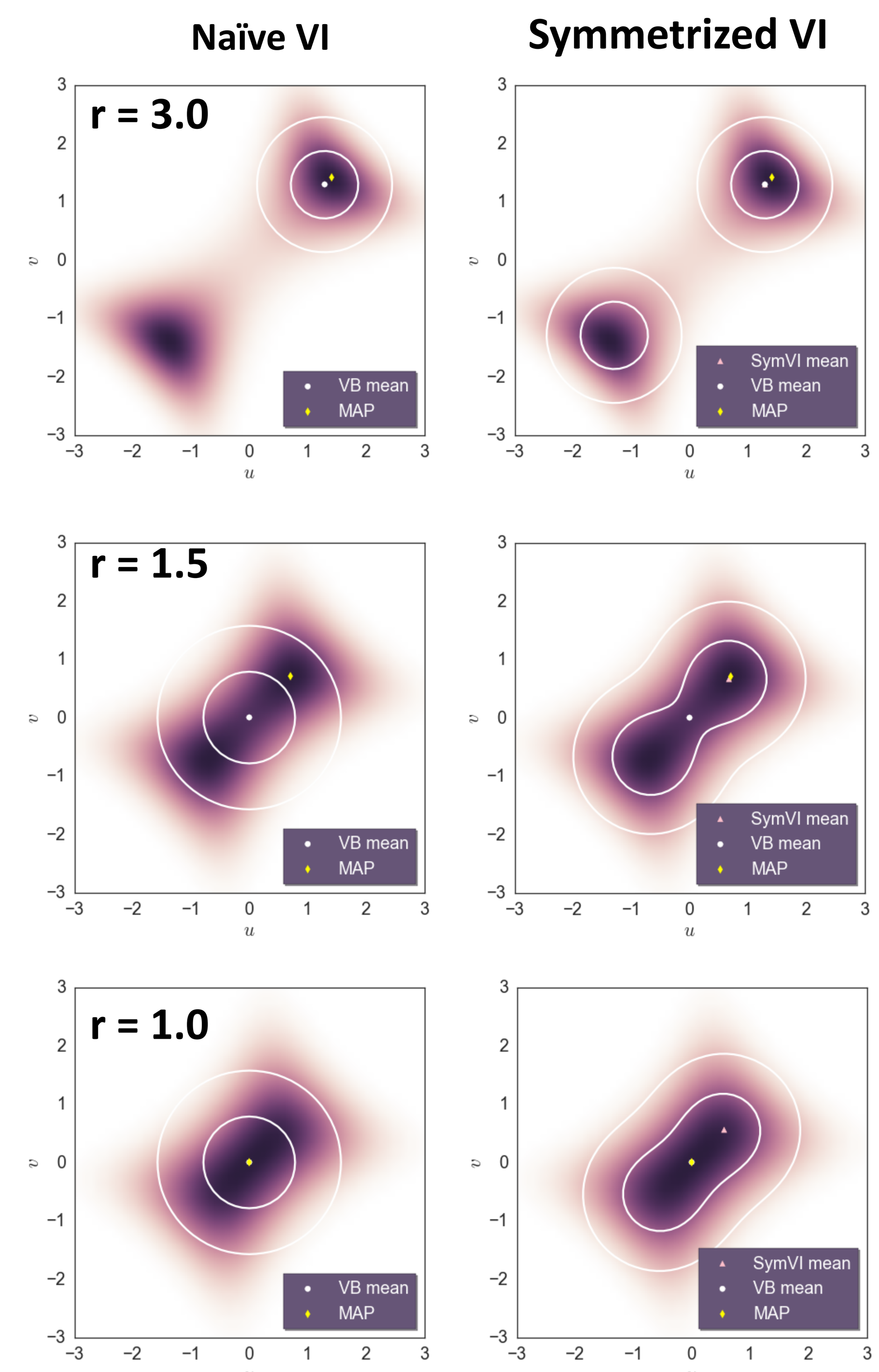
Intuitively, the symmetrized KL correction encourages nonzero singular values and low nullspace dimension in the mean matrix **M.**

Illustration: regularization from signflip symmetry

Bayesian scalar factorization:
u, v ~ N(0,1)
ε ~ N(0,1)
observed: r = uv + ε

symmetry: p(u,v|r) = p(-u,-v|r)
task: predict true "rating" uv





MAP and naïve VI predictions are pulled towards zero by the opposite-sign mode. Symmetrized predictions follow the true Bayes predictive mean.

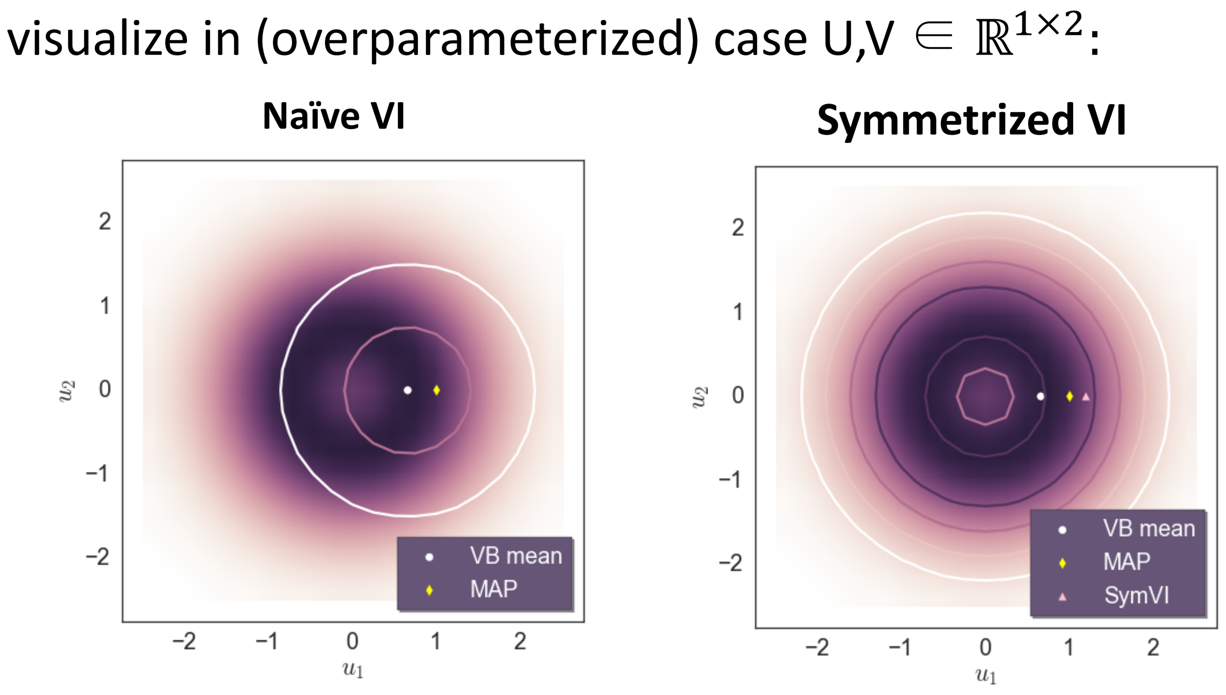
General rotation symmetry

Bayesian matrix factorization:

R = UV^T **+ ε = (UT)(VT)**^{T} **+ ε**

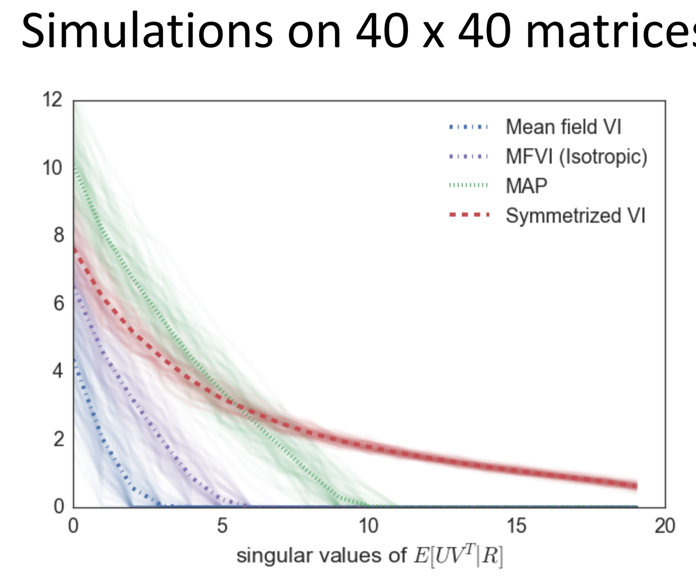
Invariant to transformation by any **T** s.t. **T(T**^{T}**) = I,** i.e., orthogonal transformations.

Can visualize in (overparameterized) case U,V $\in \mathbb{R}^{1\times 2}$:

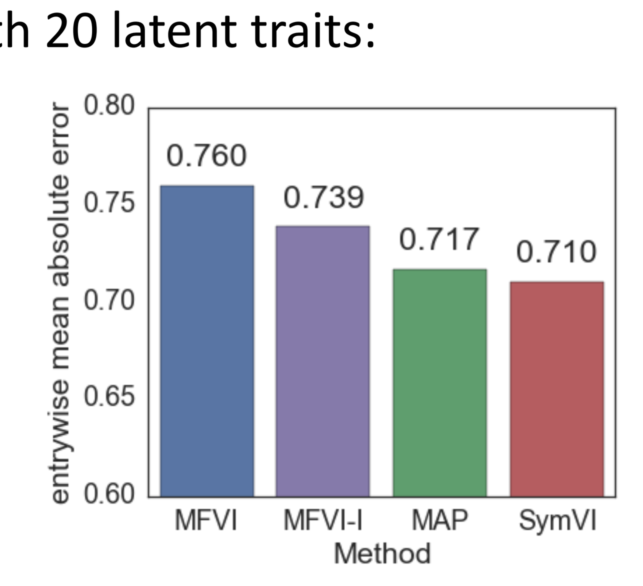


Naïve MAP/VI solutions are (again) shrunk towards zero. The symmetrized solution avoids shrinkage by implicitly modeling a continuous Gaussian mixture around the unit circle.

Simulations on 40 x 40 matrices with 20 latent traits:



Modeling posterior symmetries allows inference to use the full model capacity (all 20 traits)



Leads to improved predictive accuracy (recovering "true" noise-free ratings **UV**^{T})

Future/ongoing work:

- Extension to nonisotropic Gaussian q*.
- Other expressive posterior classes (normalizing flows, autoregressive, particle-based).
- Other symmetry groups: permutation ("label switching"), translation, scaling.
- Stochastic/minibatch inference, application to VAEs.

References

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.

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.