The Shrinkage-Delinkage Trade-off

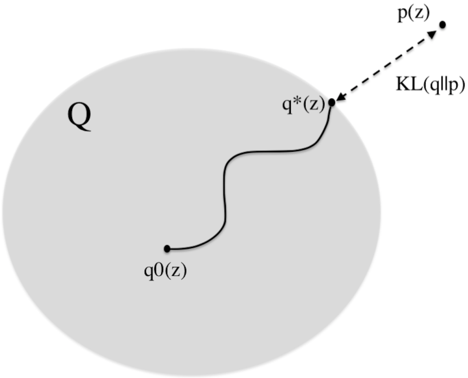
An analysis of factorized Gaussian approximations
for variational inference

Charles Margossian
& Lawrence Saul

Flatiron Institute, Center for
Computational Mathematics
New York, NY

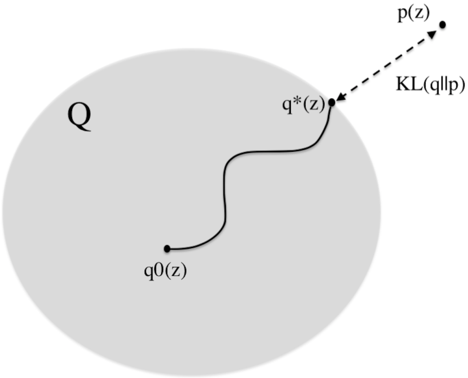
Image credit: Gabriele Veldkamp and Markus Maurer, CC

Variational inference



$$q^* = \underset{q \in \mathcal{Q}}{\operatorname{argmin}} \ \mathrm{KL}(q \,||\, p)$$

Variational inference



$$q^* = \underset{q \in \mathcal{Q}}{\operatorname{argmin}} \operatorname{KL}(q \,||\, p)$$

Usually $\operatorname{KL}(q \,||\, p) \neq 0$... so what?

Factorized variational inference (F-VI)

$$q(\mathbf{z}) = \prod_{i=1}^{n} q(z_i).$$

Factorized variational inference (F-VI)

$$q(\mathbf{z}) = \prod_{i=1}^{n} q(z_i).$$

Applications

- **Statistical Physics:** mean-field approximation of Gibbs distributions.
- **Bayesian Statistics:** Learn the mean, variance, and quantile of interpretable variables.
- **Machine Learning**: deep generative models such as VAEs.

Fact: F-VI cannot estimate the correlations between different elements of $\mathbf{z}$.

Fact: F-VI cannot estimate the correlations between different elements of $\mathbf{z}$.

Common wisdom:

- $q(z_i) \neq p(z_i)$
- F-VI tends to underestimate the "uncertainty" of $p(\mathbf{z})$.

Fact: F-VI cannot estimate the correlations between different elements of $\mathbf{z}$.

Common wisdom:

- $q(z_i) \neq p(z_i)$
- F-VI tends to underestimate the "uncertainty" of $p(\mathbf{z})$.

Which notion of uncertainty should we use?

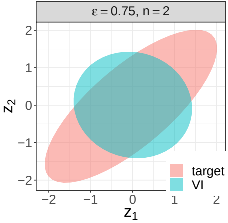
- Marginal variance, $\mathrm{Var}(z_i)$
- Entropy, $\mathcal{H}(p) = -\mathbb{E} \log p(\mathbf{z})$
- Frequentist intervals of Bayes estimators (Wang and Titterington, 2005)

$$p(\mathbf{z}) = \mathrm{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \text{ with } \mathbf{z} \in \mathbb{R}^n \text{ and } \mathrm{corr}_p(z_1, z_2) = \varepsilon.$$

$$p(\mathbf{z}) = \mathrm{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \text{ with } \mathbf{z} \in \mathbb{R}^n \text{ and } \mathrm{corr}_p(z_1, z_2) = \varepsilon.$$
$$q(\mathbf{z}) = \mathrm{Normal}(\boldsymbol{\nu}, \boldsymbol{\Psi}), \text{ where } \boldsymbol{\Psi} \text{ is diagonal.}$$

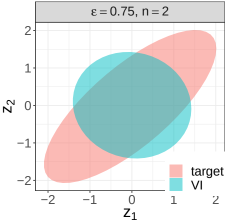
$p(\mathbf{z}) = \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\mathbf{z} \in \mathbb{R}^n$ and $\text{corr}_p(z_1, z_2) = \varepsilon$.

$q(\mathbf{z}) = \text{Normal}(\boldsymbol{\nu}, \boldsymbol{\Psi})$, where $\boldsymbol{\Psi}$ is diagonal.



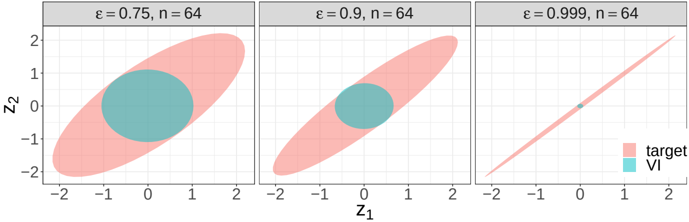
$n = 2$ example (e.g MacKay, 2003; Bishop, 2006; Turner and Sahani, 2011; Blei et al., 2017)

$p(\mathbf{z}) = \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\mathbf{z} \in \mathbb{R}^n$ and $\text{corr}_p(z_1, z_2) = \varepsilon$.
$q(\mathbf{z}) = \text{Normal}(\boldsymbol{\nu}, \boldsymbol{\Psi})$, where $\boldsymbol{\Psi}$ is diagonal.



$n = 2$ example (e.g MacKay, 2003; Bishop, 2006; Turner and Sahani, 2011; Blei et al., 2017)

$n = 64$



Plan

1. For FG-VI applied to Gaussian target, show

$$\begin{aligned}\mathrm{Var}_q(z_i) &\leq \mathrm{Var}_p(z_i)\\ \mathcal{H}(q) &\leq \mathcal{H}(p)\end{aligned}$$

2. Relationship between variance shrinkage and entropy gap... *or why the 2-D projections can be misleading*

3. Non-Gaussian targets

Factorized Gaussian Variational Inference (FG-VI)

$$p(\mathbf{z}) = \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$
$$q(\mathbf{z}) = \text{Normal}(\boldsymbol{\nu}, \boldsymbol{\Psi}), \text{ where } \boldsymbol{\Psi} \text{ is diagonal.}$$

Proposition

$KL(q||p)$ is minimized by

$$\begin{aligned} \boldsymbol{\nu} &= \boldsymbol{\mu} \\ \Psi_{ii} &= \frac{1}{\Sigma_{ii}^{-1}}. \end{aligned}$$

In general, $\Psi_{ii} \neq \Sigma_{ii}$.

Theorem

When FG-VI targets a Gaussian, we underestimate uncertainty in two ways,

1. **Variance shrinkage:**

$$\Psi_{ii} \leq \Sigma_{ii}, \quad \forall i.$$

2. **Entropy gap:**

$$\mathcal{H}(q) \leq \mathcal{H}(p).$$

Theorem

When FG-VI targets a Gaussian, we underestimate uncertainty in two ways,

1. **Variance shrinkage:**

$$\Psi_{ii} \leq \Sigma_{ii}, \quad \forall i.$$

2. **Entropy gap:**

$$\mathcal{H}(q) \leq \mathcal{H}(p).$$

Proof of (1) is intriguingly simple but not obvious.

Proof of (2):

$$\begin{aligned}
\mathcal{H}(p) - \mathcal{H}(q) &= -\frac{1}{2}\log|\mathbf{\Psi}| - \left(-\frac{1}{2}\log|\mathbf{\Sigma}|\right) \\
&= \mathrm{KL}(q||p) \\
&\geq 0.
\end{aligned}$$

How does the entropy gap relate to the variance shrinkage?

How does the entropy gap relate to the variance shrinkage?

Correlation matrix:

$$C_{ij} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}}, \quad C_{ii} = 1.$$

Shrinkage matrix:

$$S_{ii} = \frac{\Sigma_{ii}}{\Psi_{ii}} = \Sigma_{ii}\Sigma_{ii}^{-1}$$

How does the entropy gap relate to the variance shrinkage?

Correlation matrix:

$$C_{ij} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii}\Sigma_{jj}}}, \quad C_{ii} = 1.$$

Shrinkage matrix:

$$S_{ii} = \frac{\Sigma_{ii}}{\Psi_{ii}} = \Sigma_{ii}\Sigma_{ii}^{-1}$$

Theorem

(shrinkage-delinkage trade-off)

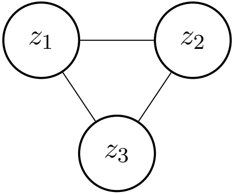
$$\mathcal{H}(p) - \mathcal{H}(q) = \underbrace{\frac{1}{2}\log|\mathbf{S}|}_{\geq 0} - \underbrace{\frac{1}{2}\log|\mathbf{C}|^{-1}}_{\geq 0}.$$

- Two competing forces: shrinkage and delinkage.

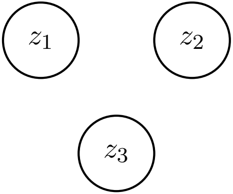
Theorem

(shrinkage-delinkage trade-off)

$$\mathcal{H}(p) - \mathcal{H}(q) = \underbrace{\frac{1}{2}\log|\mathbf{S}|}_{\geq 0} - \underbrace{\frac{1}{2}\log|\mathbf{C}|^{-1}}_{\geq 0}.$$



Linked graphical model,
$p(\mathbf{z}) \neq \prod_i p(z_i)$

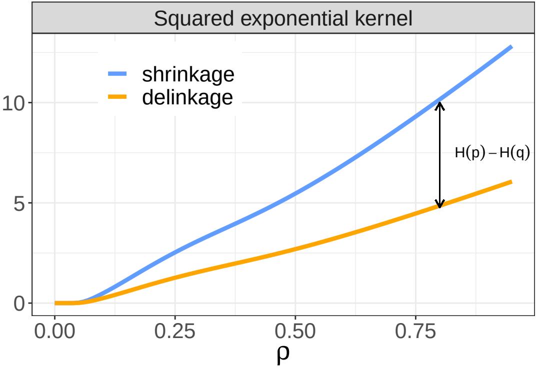


Delinked graphical model,
$q(\mathbf{z}) = \prod_i q(z_i)$

$n = 10$

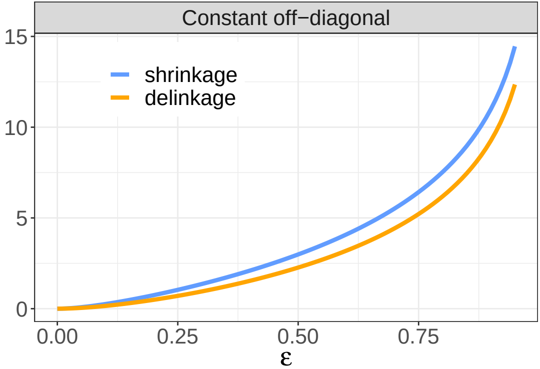
Example: squared exponential kernel

$$\Sigma_{ij} = \exp(-(x_i - x_j)^2/\rho^2)$$



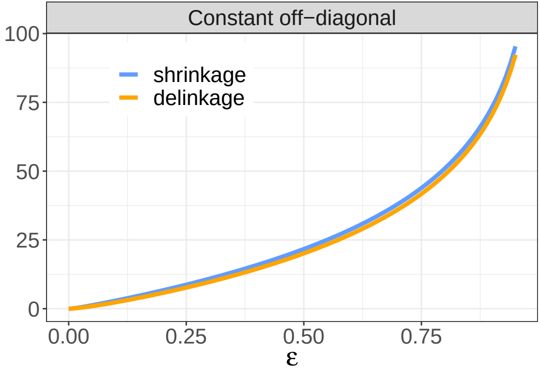
$n = 10$

Example: covariance with constant off-diagonal terms, ε.



$n = 64$

Example: covariance with constant off-diagonal terms, ε.



Theorem

Suppose $\mathbf{\Sigma}$ *has constant off-diagonal terms,* $\varepsilon > 0$. *Then*

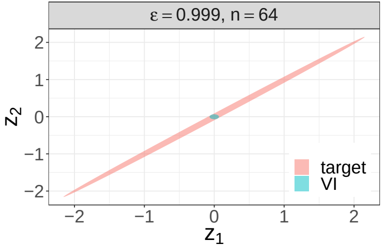
- ✓ **Vanishing entropy gap:**

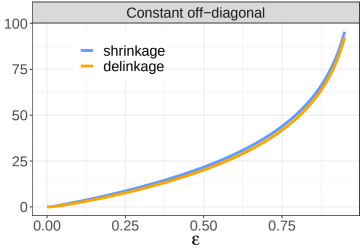
$$\lim_{n\to\infty} \frac{1}{n}(\mathcal{H}(p) - \mathcal{H}(q)) = 0$$

- ✗ **Arbitrarily bad variance shrinkage:**

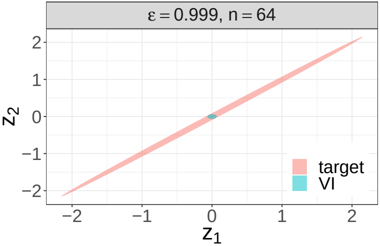
$$\lim_{n\to\infty} S_{ii} = \Sigma_{ii}/\Psi_{ii} = \frac{1}{1-\varepsilon}.$$

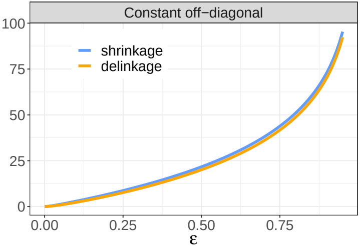
How do we reconcile these two pictures?





How do we reconcile these two pictures?





- Need to reason about the limit $n \to \infty$.
- What happens to the volume of the sphere and the ellipsoid in higher dimensions?

For Σ with constant off-diagonal, ε.

Minimize KL$(q \mid\mid p)$

✓ Vanishing entropy gap

✗ Variance shrinkage

Minimize KL$(p \mid\mid q)$

For Σ with constant off-diagonal, ε.

Minimize KL$(q \mid\mid p)$

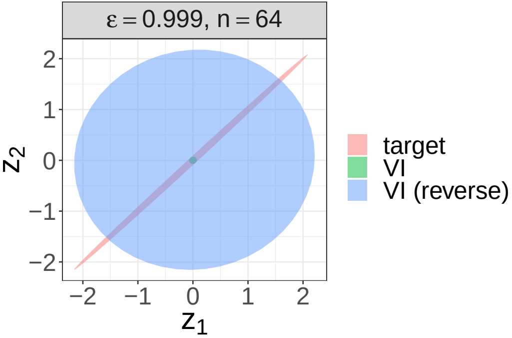
✓ Vanishing entropy gap

✗ Variance shrinkage

Minimize KL$(p \mid\mid q)$

✗ Large entropy gap

✓ No variance shrinkage



Factorized variational inference (F-VI)

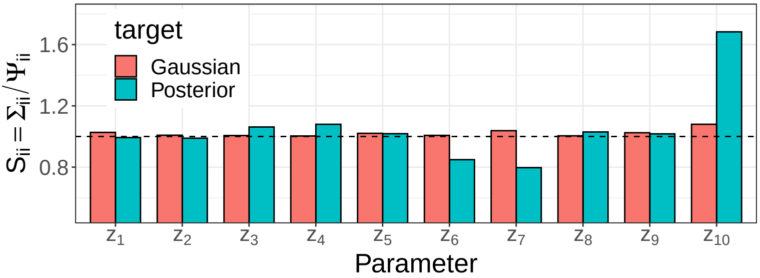
$$q(\mathbf{z}) = \prod_{i=1}^{n} q(z_i).$$

Applications

- **Statistical Physics:** mean-field approximation of Gibbs distributions.
- **Bayesian Statistics:** Learn the mean, variance, and quantile of interpretable variables.
- **Machine Learning**: deep generative models such as VAEs.

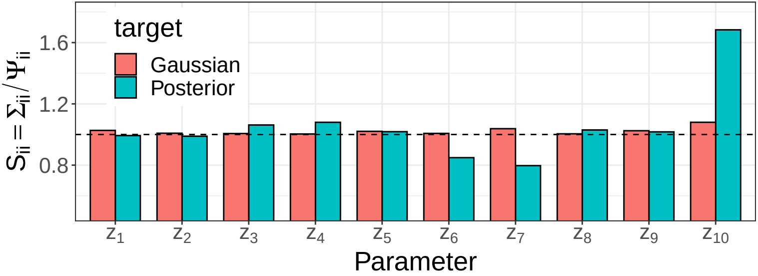
Non-Gaussian models

8 schools model (non-centered parameterization)



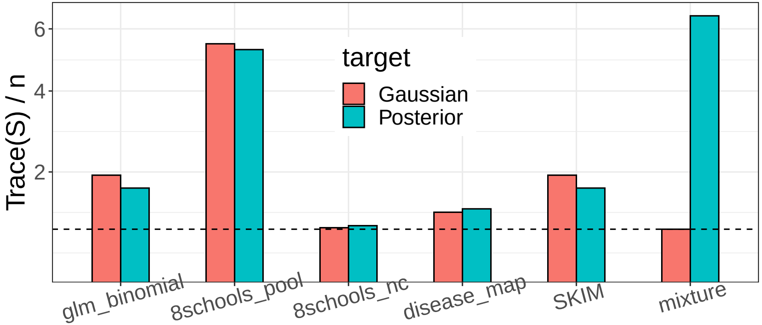
Non-Gaussian models

8 schools model (non-centered parameterization)



- The inequality $\text{Var}_q(z_i) \leq \text{Var}_p(z_i)$ is violated.
- But $\frac{1}{n}\text{trace}(\mathbf{S}) = \frac{1}{n}\sum_i S_{ii} \geq 1$.

Trace($\mathbf{S}$) for a diversity of targets.



In all examples, variance shrinkage holds *on average*.

Empirical study for entropy gap

- Requires a method to estimate the normalizing constant, such as bridge sampling (Meng and Schilling, 2002; Gronau et al., 2020); but such methods use a (skewed) Gaussian approximation.

Empirical study for entropy gap

- Requires a method to estimate the normalizing constant, such as bridge sampling (Meng and Schilling, 2002; Gronau et al., 2020); but such methods use a (skewed) Gaussian approximation.
- Can show

$$\mathcal{H}(p) - \mathcal{H}(q) \leq \frac{1}{2}(\log|\mathbf{\Sigma}| - \log|\mathbf{\Psi}|).$$

  This upper-bound is positive in all considered examples.

Empirical study for entropy gap

- Requires a method to estimate the normalizing constant, such as bridge sampling (Meng and Schilling, 2002; Gronau et al., 2020); but such methods use a (skewed) Gaussian approximation.
- Can show

$$\mathcal{H}(p) - \mathcal{H}(q) \leq \frac{1}{2}(\log|\mathbf{\Sigma}| - \log|\mathbf{\Psi}|).$$

  This upper-bound is positive in all considered examples.
- Turner and Sahani (2011) provide a counter-example where FG-VI overestimates entropy.

Contributions

- Variance shrinkage
- Entropy gap
- Shrinkage-Delinkage trade-off
- Bounds on the shrinkage and delinkage terms.
- Non-Gaussian examples

UAI paper ID: 149

Open questions

- More generally, how does the shrinkage-delinkage trade-off manifest?
- Under what conditions does F-VI underestimate entropy?
- What error do we introduce when minimizing other objective functions?

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

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

Giordano, R., Broderick, T., and Jordan, M. I. (2018). Covariances, robustness, and variational bayes. *Journal of Machine Learning Research*, 19:1 – 49.

Gronau, Q. F., Singmann, H., and Wagenmakers, E.-J. (2020). bridgesampling: An R package for estimating normalizing constants. *Journal of Statistical Software*, 92.

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

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. (2017). Automatic differentiation variational inference. *Journal of machine learning research*, 18:1 – 45.

MacKay, D. J. (2003). *Information theory, inference, and learning algorithms*.

Margossian, C. C. and Mukherjee, S. (2021). Simulating ising and potts models at critical and cold temperatures using auxiliary gaussian variables. *arXiv:2110.10801*.

Meng, X. and Schilling, S. (2002). Warp bridge sampling. *Journal of Computational and Graphical Statistics*, 11:552 – 586.

Mukherjee, R., Mukherjee, S., and Yuan, M. (2018). Global testing against sparse alternatives under Ising models. *Annals of Statistics*, 46.

Parisi, G. (1988). *Statistical Field Theory*. Addison-Wesley.

Turner, R. E. and Sahani, M. (2011). Two problems with variational expectation maximisation for time-series models. In Barber, D., Cemgil, A. T., and Chiappa, S., editors, *Bayesian Time series models*, chapter 5, pages 109–130. Cambridge University Press.