Variational Bayes under Model Misspecification

Yixin Wang

COLUMBIA UNIVERSITY

October 28, 2019

(Joint work with David Blei)

Approximate Bayesian inference

- Two steps:

  1. Posit a probability model, i.e. the **joint distribution** of **latent** variables $\boldsymbol{\theta}$ and **observed** variables $\boldsymbol{x}$,

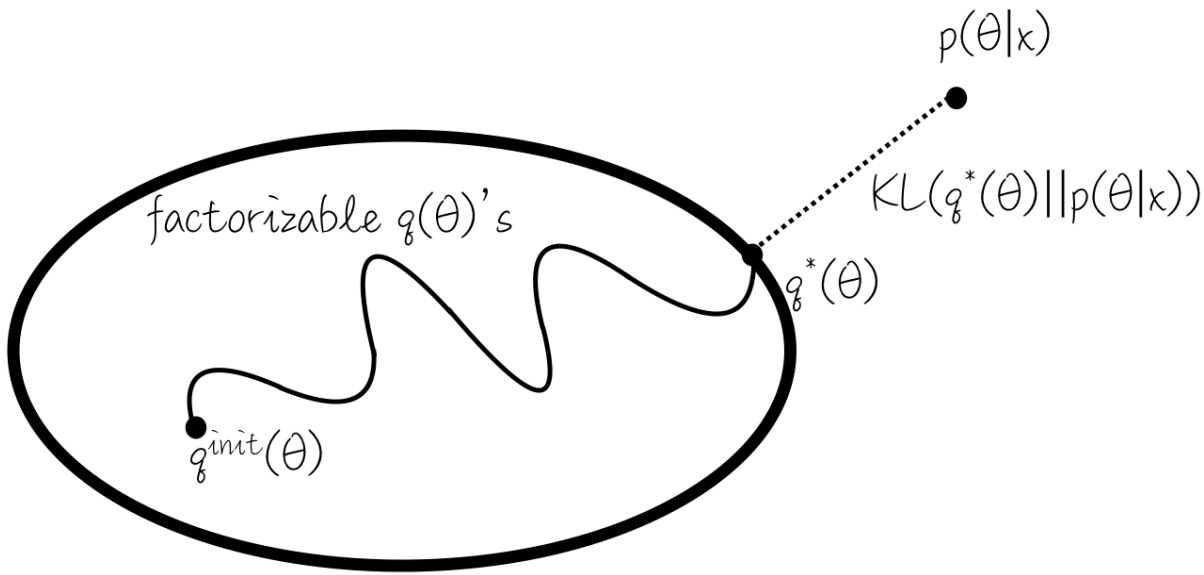
  $$p(\boldsymbol{\theta}, \boldsymbol{x}).$$

  2. Infer the unknown, through the **posterior**, the conditional distribution of the latent variables $\boldsymbol{\theta}$ given the observed $\boldsymbol{x}$,

  $$p(\boldsymbol{\theta} \mid \boldsymbol{x}) = \frac{p(\boldsymbol{\theta}, \boldsymbol{x})}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{\theta}) p(\boldsymbol{x} \mid \boldsymbol{\theta})}{\int p(\boldsymbol{\theta}, \boldsymbol{x}) \, \mathrm{d}\boldsymbol{\theta}}.$$

- The **integral** denominator is often intractable. We appeal to **approximate posterior inference** – Markov chain Monte Carlo (MCMC) or variational Bayes (VB) (Wainwright & Jordan, 2008; Blei et al, 2017).

Variational Bayes (VB)

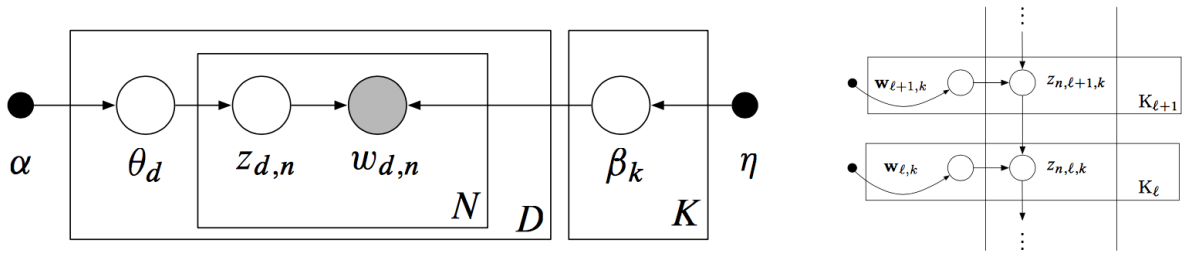


- Variational Bayes (VB) solves posterior inference with **optimization**.
  - Posit a variational family of distributions over the latent variables $\boldsymbol{\theta} = \theta_{1:K}$,

$$q(\boldsymbol{\theta}) = \prod_{k=1}^{K} q_{\theta_k}(\theta_k).$$

  - Find the member $q^*(\boldsymbol{\theta})$ in this family that is closest (in KL) to the exact posterior $p(\boldsymbol{\theta} \mid \boldsymbol{x})$.

Where is variational Bayes used?



- Mixture models (Bishop, 2006; Murphy, 2012)
- Generalized linear mixed models (McCulloch & Neuhaus, 2001)
- Stochastic block models (Wang & Wong, 1987; Snijders & Nowicki, 1997)
- Mixed membership models (Blei et al., 2003; Pritchard et al., 2000)
- and many others...

The lack-of-theory of variational Bayes

- Ghahramani & Beal, 2000.

Of course, vis-a-vis MCMC, the main disadvantage of variational approximations is that they are not guaranteed to find the exact posterior in the limit. However, with a straightforward application of sampling, it is possible to take the result of

- Jaakkola, 2001.

One of the main open problems in the use of variational approximation methods is characterizing their accuracy. We would like to obtain performance guarantees for specific classes of graphical models (upper/lower bounds that can be obtained from several variational formulations provide such guarantees only for specific instantiations of the inference problem and would not serve as *a priori* guarantees). Another open problem concerns

- Blei & Lafferty, 2006.

between the approximate and true posterior is small. For many problems this optimization problem is computationally manageable, while standard methods, such as Markov Chain Monte Carlo, are impractical. The tradeoff is that variational methods do not come with the same theoretical guarantees as simulation methods. See [13] for a modern review of variational methods for statistical inference.

- Sung, Ghahramani, & Bang, 2008.

monitoring convergence. On the other hand, the variational method, which is called *Variational Bayes (VB)* for Bayesian inferences, can require much less computation and comes with an easy to evaluate convergence criterion, but does not have the same asymptotic guarantees as MCMC.

- Braun & McAuliffe, 2010.

it takes to generate an adequate number of MCMC draws. This advantage comes at the cost of a biased approximation, in contrast to the consistency guarantees that accompany MCMC. We

- Blei, Kucukelbir, & McAuliffe, 2017.

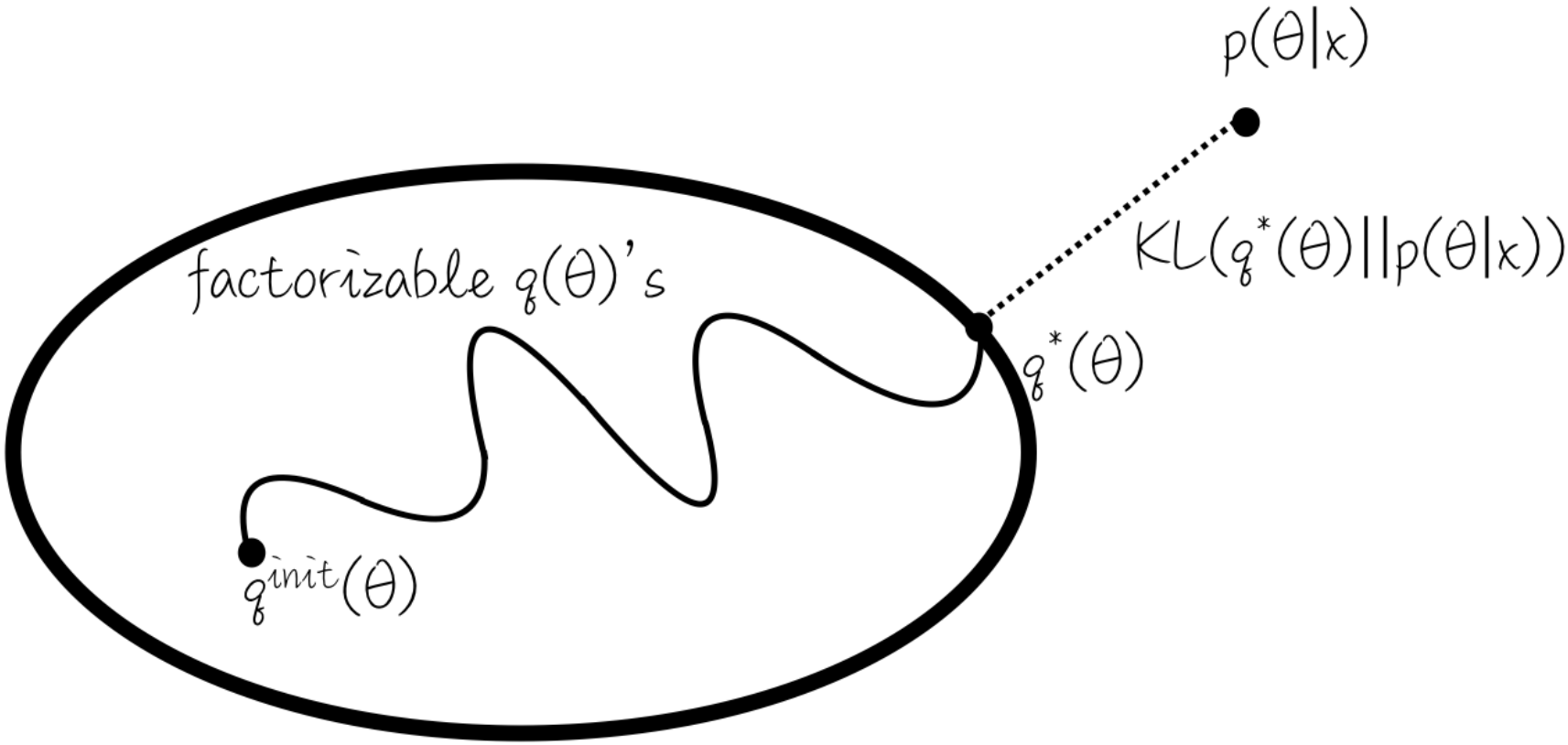
when should she use variational inference? We will offer some guidance. MCMC methods tend to be more computationally intensive than variational inference but they also provide guarantees of producing (asymptotically) exact samples from the target density (Robert and Casella, 2004). Variational inference does not enjoy such guarantees—it can only find a density close to the target—but tends to be faster than MCMC. Because it rests on

This talk: Theoretical guarantees of VB

Wang, Y., & Blei, D. M. (2017).
Frequentist consistency of variational Bayes.
arXiv preprint arXiv:1705.03439.
To appear in *Journal of the American Statistical Association*.

- The VB posterior is consistent and asymptotically normal.
- The VB estimate is consistent and asymptotically normal.

Variational Bayes (VB)



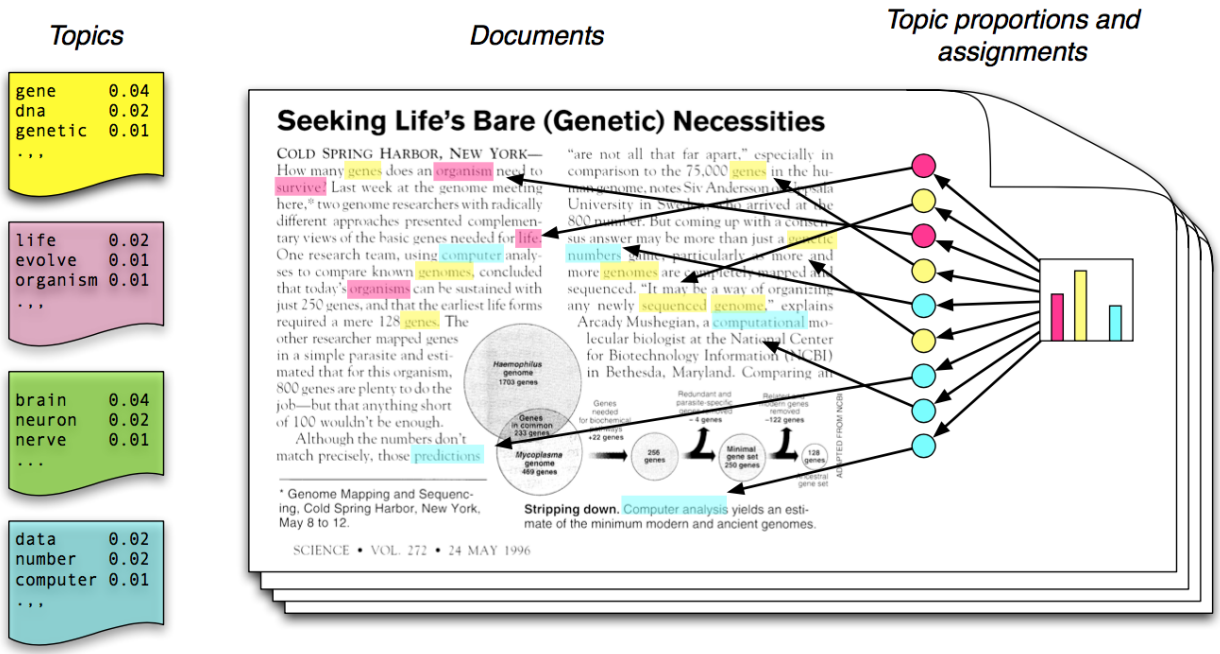
The VB posterior: $q^*(\theta)$

A curious experiment on topic models

Latent variables of interest $\boldsymbol{\theta}$: topics

Observed variables $\boldsymbol{x}$: word occurrences in documents

A curious experiment on topic models

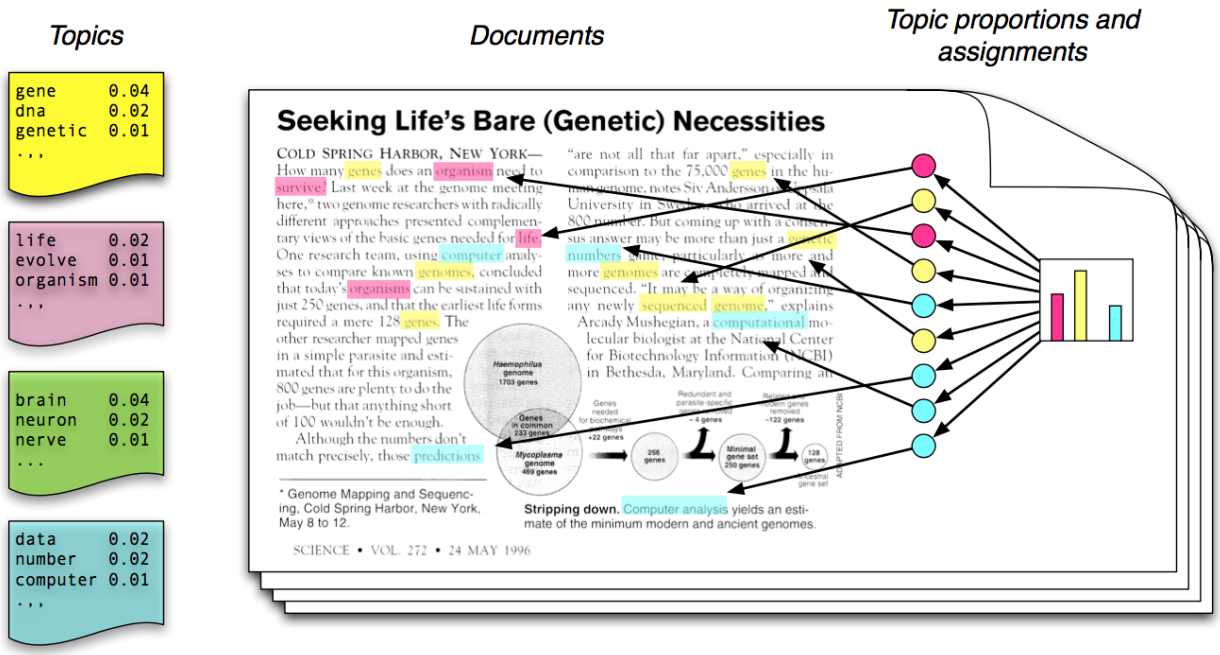


- Topics: $\theta_k \overset{iid}{\sim} Dirichlet(\alpha_\theta), k = 1, ..., K.$ (Latent variable of interest)
- Per-doc topic proportions: $\beta_d \overset{iid}{\sim} Dirichlet(\alpha_\beta), d = 1, ..., D.$
- Per-word topic assignment: $z_{di} \overset{iid}{\sim} Multinomial(\beta_d), i = 1, ..., N_d,$
- Word occurrences: $x_{di} \sim Multinomial(\theta_{z_{di}})$. (Observed variable)

$$\Rightarrow p(\boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{z}, \boldsymbol{x})$$

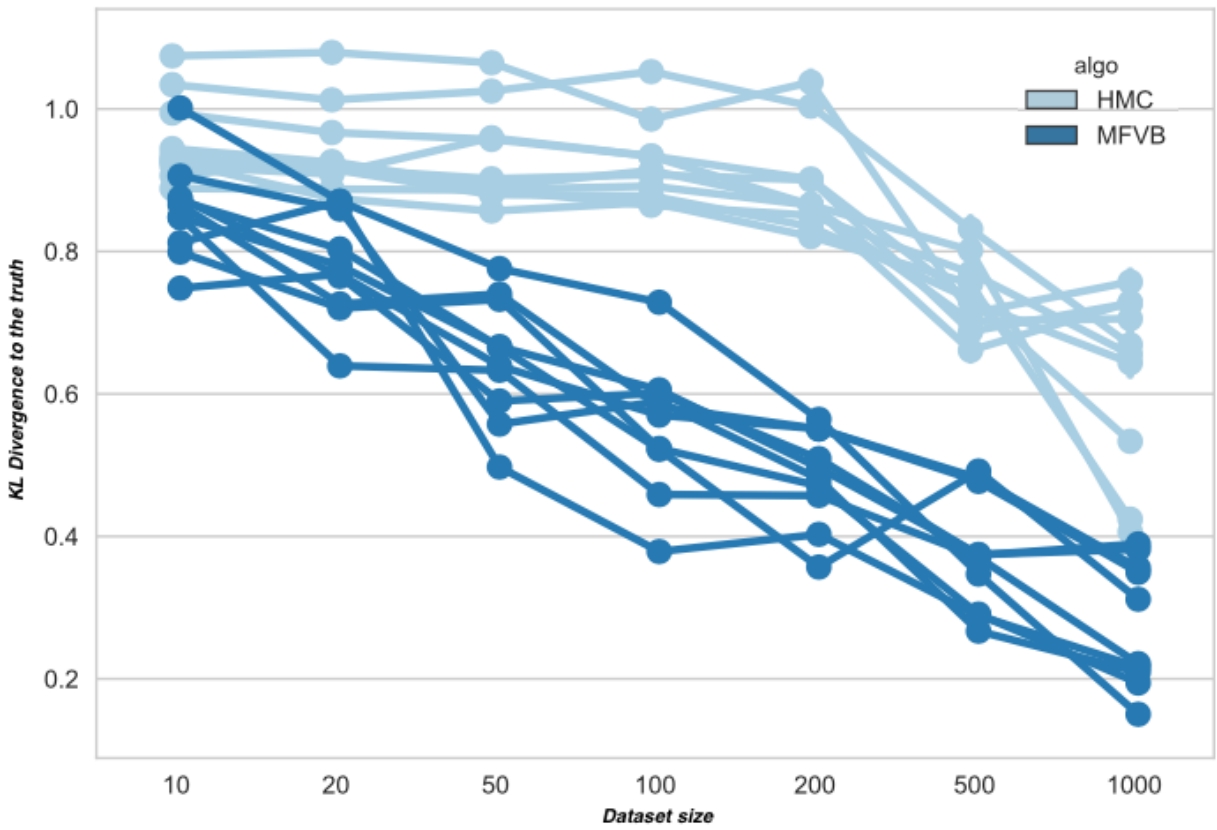
[Latent Dirichlet Allocation (LDA), Blei+ JMLR 2003]

A curious experiment on topic models



What if we have **infinite** data?

A curious experiment on topic models



- VB converges to the truth faster than MCMC.
- MCMC is theoretically guaranteed.
- **Same guarantees should exist for VB.**

Main theorem (in plain English)

Setup:

- Start with a **model** with latent and observed variables.
- Simulate **data** from this model with latent variables **fixed at a true value**.
- Consider the **posterior** of the latent variables **given the simulated data**.

Main theorem (in plain English)

Variational Bernstein–von Mises Theorem

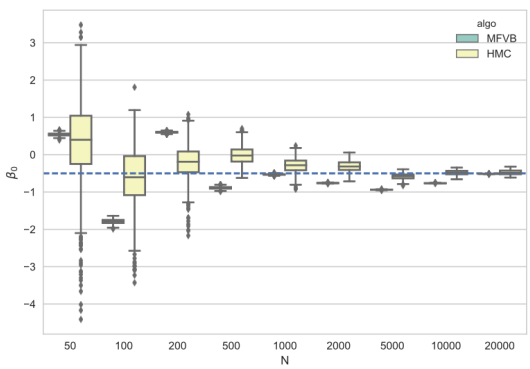
(Under standard conditions)

1. **(Consistency of the VB posterior)** The VB posterior converges in distribution to a point mass at the true value.
2. **(Asymptotic normality of the VB posterior)** The VB posterior, if re-centered and rescaled properly, converges in TV distance to a normal.
3. **(Consistency of the VB estimate)** The VB estimate, a.k.a. the mean of the VB posterior, converges to the true value.
4. **(Asymptotic normality of the VB estimate)** The VB estimate, if re-centered and rescaled properly, converges in distribution to a normal.

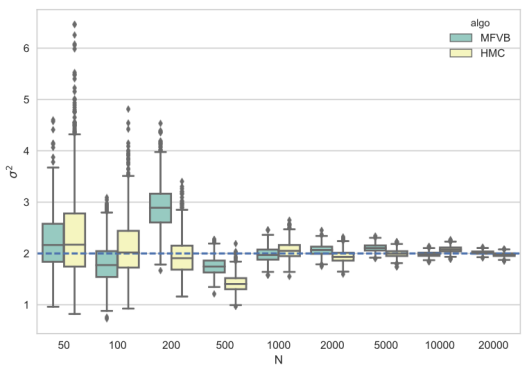
Variational BvM Theorem: Implications

- **Characterize** the VB posterior for specific models
  - Variational BvM: verify its conditions and apply
  - Prove consistency and derive asymptotic distributions for VB
  - Many existing results can verify the conditions,
    e.g. Hall et al. (2011), Bickel et al. (2013), Westling & McCormick (2015), ...
- **Use** variational Bayes with more confidence

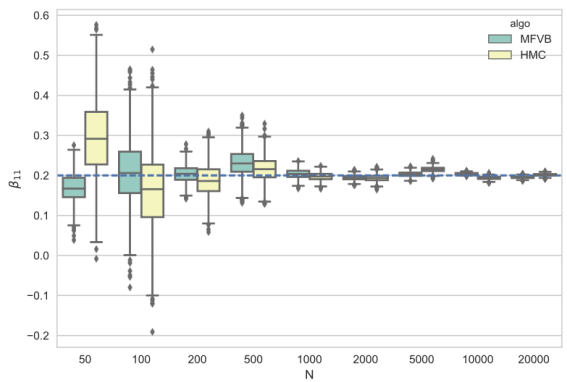
Applications: Poisson GLMM



(a) Posterior of β_0



(b) Posterior of σ^2



(c) Posterior of β_{11}

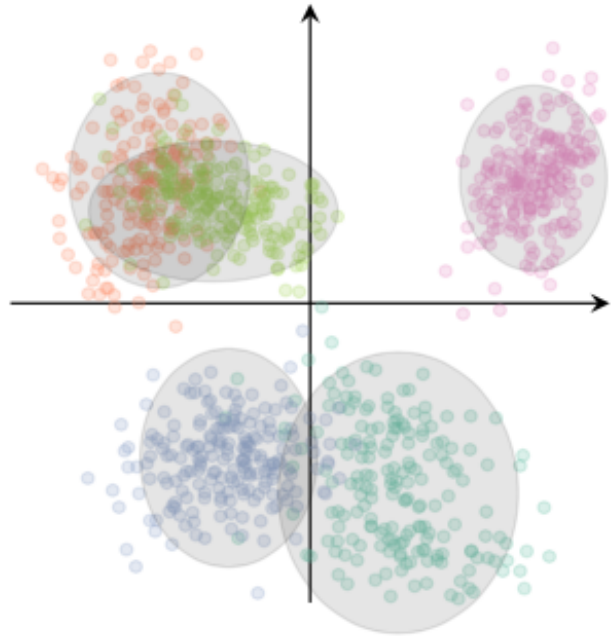
- Poisson generalized linear mixed-effects model

$$U_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2), i = 1, \ldots, m$$

$$Y_{ij} \mid X_{ij}, U_i \sim \text{Poisson}(\exp(\beta_0 + \beta_1 X_{ij} + U_i)), j = 1, \ldots, n$$

- Use Hall et al. (2011) to verify the variational BvM conditions.

Applications: Mixture models



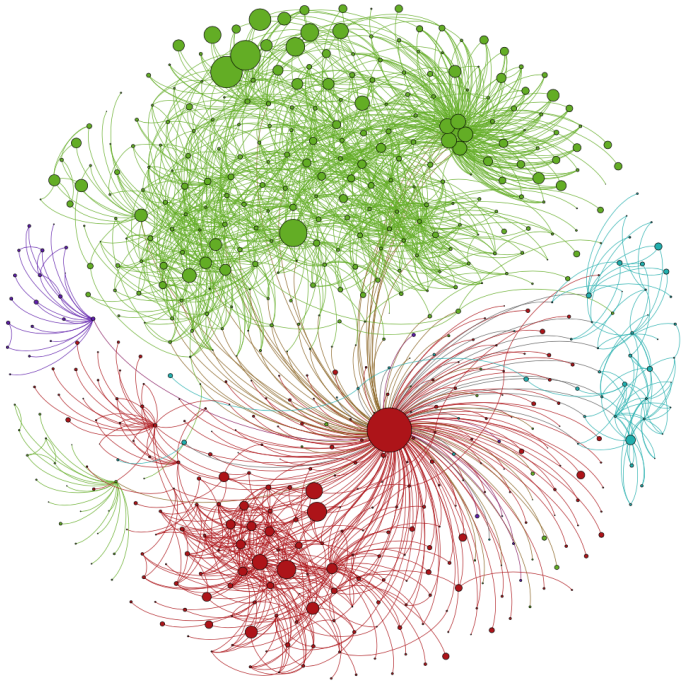
- Gaussian mixture models

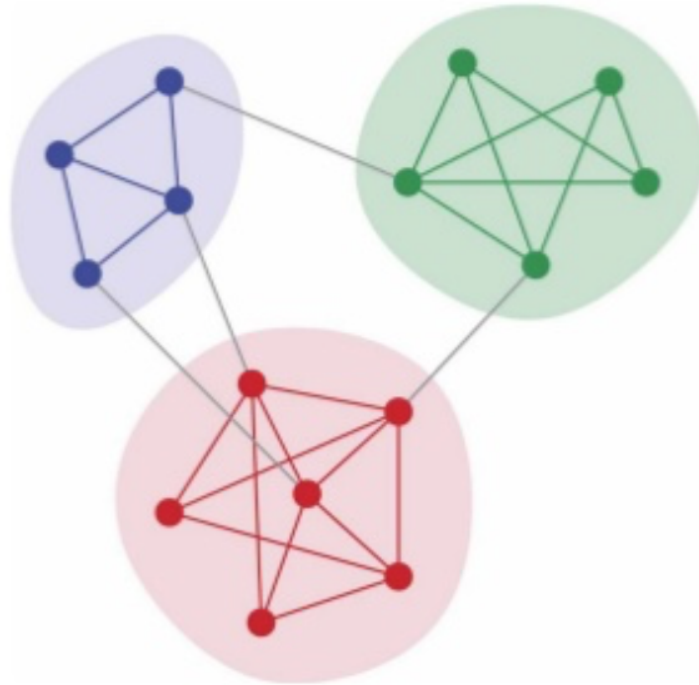
$$Z_i \overset{iid}{\sim} Multinomial(\lambda_1, ..., \lambda_K)$$

$$X_i \overset{iid}{\sim} \mathcal{N}(\mu_{Z_i}, \Sigma_{Z_i})$$

- Use Westling & McCormick (2015) to verify the conditions.

Applications: Stochastic block models





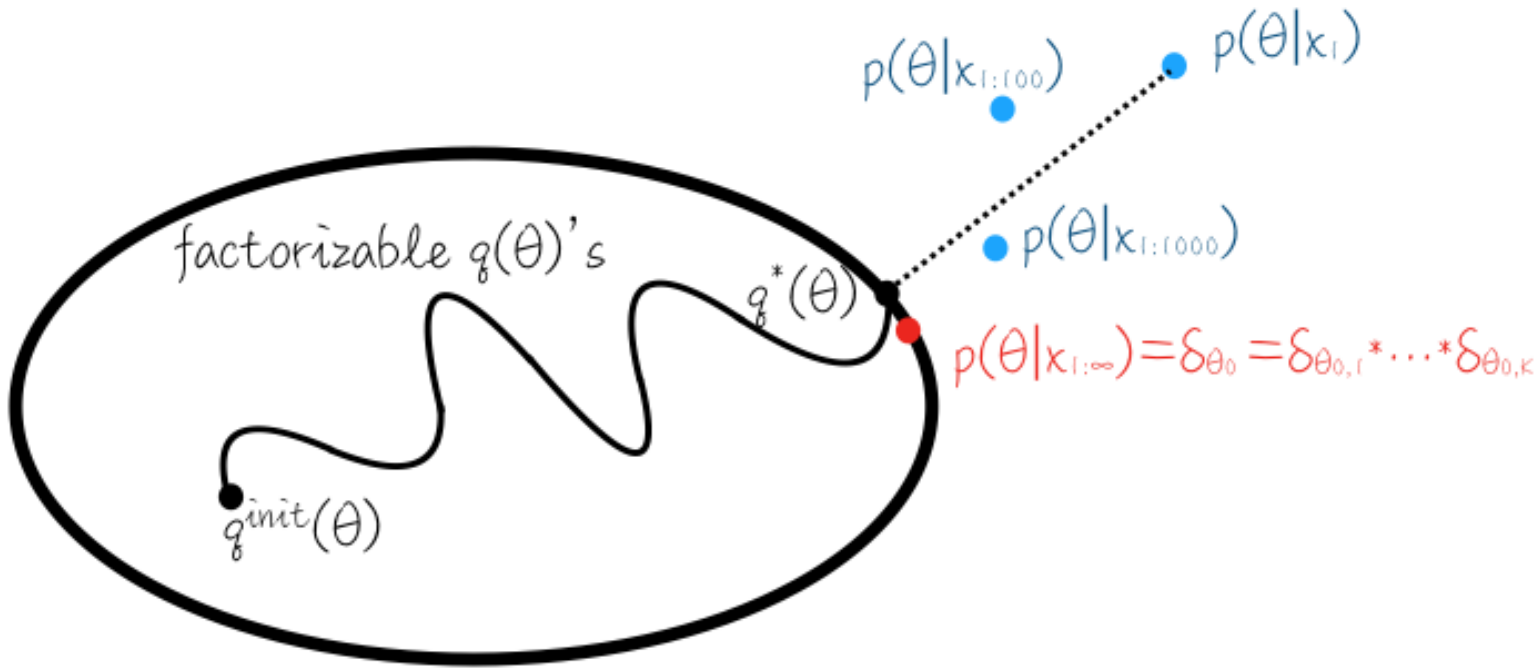
- Stochastic block models
  - $\pi \in \text{simplex}(K)$: latent class proportions with K classes
  - $H \in [0,1]^{K \times K}$: edge probabilities between two latent classes

$$Z_i \mid \pi \overset{iid}{\sim} \text{Categorical}(\pi),$$

$$A_{ij} \mid Z_i, Z_j, H \overset{iid}{\sim} \text{Bernoulli}(H_{Z_i Z_j}).$$

- Use Bickel et al. (2013) to verify the conditions.

Intuition: Why is the VB posterior consistent?



- As more data comes in, the exact posterior **contracts**. It converges to a **point mass** at the truth eventually.
- Point masses are **factorizable**. So the exact posterior in the limit **sits in** the variational family.
- If the variational family **contains** the truth, the VB posterior **recovers the truth**.

$\Rightarrow$ **The VB posterior recovers the truth in the limit!**

Main technical conditions

$$q^*(\boldsymbol{\theta}) = \arg\min_{q \in \mathcal{Q}} \mathrm{KL}(q(\boldsymbol{\theta}) || p(\boldsymbol{\theta} \mid \boldsymbol{x}))$$

- **Prior mass:** priors on the latent variable $\boldsymbol{\theta}$ puts enough mass to sufficiently small balls around the true value $\boldsymbol{\theta_0}$.
- **Consistent testability:** when $\boldsymbol{\theta} \neq \boldsymbol{\theta_0}$, there exists a decision rule that always rejects the null hypothesis $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta_0}$ given infinite data.
- **Local asymptotic normality:** the probability model can be asymptotically approximated by a normal model.

These conditions ensure the consistency and asymptotic normality of the **exact** posterior (Van der Vaart, 1998).

Variational Bayes **does not** require extra technical conditions!

Main theorem: the VB posterior

1. **(Consistency of the VB posterior)** The VB posterior converges in distribution to a point mass at the true value.

- Start with a model $p(\boldsymbol{x} \mid \boldsymbol{\theta})$.
- Simulate data $\boldsymbol{x} = x_{1:n}$ iid from a density $p(\boldsymbol{x} \mid \boldsymbol{\theta} = \boldsymbol{\theta_0})$ with law $\mathbb{P}_{\boldsymbol{\theta_0}}$ for some fixed constant $\boldsymbol{\theta_0}$.
- Consider the VB posterior $q^*(\boldsymbol{\theta}) = \arg\min_{q \in \mathcal{Q}} \mathrm{KL}(q(\boldsymbol{\theta}) || p(\boldsymbol{\theta} \mid \boldsymbol{x} = x_{1:n}))$.

Under standard conditions, the VB posterior is consistent: almost surely under $\mathbb{P}_{\boldsymbol{\theta_0}}$,

$$q^*(\boldsymbol{\theta}) \xrightarrow{d} \boldsymbol{\delta}_{\boldsymbol{\theta_0}}.$$

Main theorem: the VB posterior

2. **(Asymptotic normality of the VB posterior)** The VB posterior, if re-centered and rescaled properly, converges in TV distance to a normal.

- Consider the VB posterior $q^*_{\tilde{\boldsymbol{\theta}}}(\tilde{\boldsymbol{\theta}})$ of $\tilde{\boldsymbol{\theta}} = \delta_n^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta_0})$, the re-centered and re-scaled $\boldsymbol{\theta}$. [obtained via a Jacobian density transformation from $q^*(\boldsymbol{\theta})$]

Under standard conditions, the VB posterior is asymptotically normal:

$$\left\| q^*_{\tilde{\boldsymbol{\theta}}}(\cdot) - \mathcal{N}(\cdot; \Delta_{n,\theta_0}, V^{-1}_{\theta_0,diag}) \right\|_{TV} \xrightarrow{\mathbb{P}_{\theta_0}} 0,$$

where Δ_{n,θ_0} is a zero mean normal random variable, and $V_{\theta_0,diag}$ is a diagonal matrix.

All of the constants are model-dependent.

Corollary: VB underestimates the variance

$$\left\| q^*_{\tilde{\boldsymbol{\theta}}}(\cdot) - \mathcal{N}(\cdot; \Delta_{n,\theta_0}, \mathbf{V}^{-1}_{\theta_0,\mathbf{diag}}) \right\|_{TV} \overset{\mathbb{P}_{\theta_0}}{\to} 0.$$

- Classical BvM: $\left\| p(\tilde{\boldsymbol{\theta}} \,|\, x) - \mathcal{N}(\cdot; \Delta_{n,\theta_0}, \mathbf{V}^{-1}_{\theta_0}) \right\|_{TV} \overset{\mathbb{P}_{\theta_0}}{\to} 0.$
- $\mathbf{V}^{-1}_{\theta_0}$ is a constant calculated from the LAN condition.
- $\mathbf{V}_{\theta_0,\mathbf{diag}}$ is diagonal and has the same diagonal entries as $\mathbf{V}_{\theta_0}$.
- **The limiting VB posterior underestimates the variance.**

$$\mathbb{H}(\mathcal{N}(\cdot; \Delta_{n,\theta_0}, \mathbf{V}^{-1}_{\theta_0,\mathbf{diag}})) \leq \mathbb{H}(\mathcal{N}(\cdot; \Delta_{n,\theta_0}, \mathbf{V}^{-1}_{\theta_0})),$$

where $\mathbb{H}(\cdot)$ is the entropy of the distribution.

Main theorem: the VB estimate

3. **(Consistency of the VB estimate)** The VB estimate, a.k.a. the mean of the VB posterior, converges to the true value.

- Consider the VB estimate $\hat{\boldsymbol{\theta}}^* = \int \boldsymbol{\theta} \cdot q^*(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}$.

Under standard conditions, the VB estimate is consistent: under $\mathbb{P}_{\boldsymbol{\theta}_0}$,

$$\hat{\boldsymbol{\theta}}_n^* \xrightarrow{a.s.} \boldsymbol{\theta}_0.$$

Main theorem: the VB estimate

4. **(Asymptotic normality of the VB estimate)** The VB estimate, if re-centered and rescaled properly, converges in distribution to a normal.

- Consider the VB estimate $\hat{\boldsymbol{\theta}}^* = \int \boldsymbol{\theta} \cdot q^*(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}$.

Under standard conditions, the VB estimate is asymptotically normal:

$$\boldsymbol{\delta}_n^{-1}(\hat{\boldsymbol{\theta}}^* - \boldsymbol{\theta_0}) \xrightarrow{d} \Delta_{\infty,\theta_0},$$

where Δ_{∞,θ_0} is a model-dependent zero mean normal random variable, and δ_n is a model-dependent scaling constant, often taking $1/\sqrt{n}$.

Related work

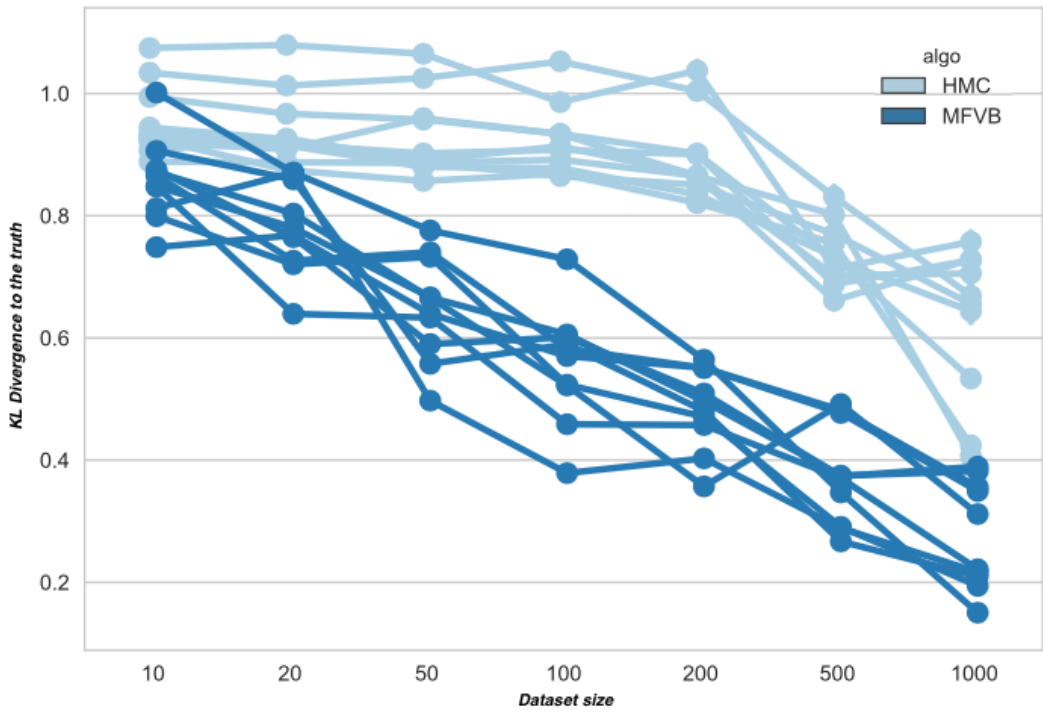
- Model-specific results on variational approximations
  - Exponential family models: Wang & Titterington (2004), You et al. (2014), Ormerod et al. (2014)
  - Poisson mixed-effects model: Hall et al. (2011)
  - Stochastic block model: Celisse et al. (2012), Bickel et al. (2013), Zhang & Zhou (2017)
  - Mixture of Gaussians: Wang & Titterington (2005, 2006), Westling & McCormick (2015)
  - Latent Gaussian Models: Sheth & Khardon (2017)
  - Latent Dirichlet Allocations: Ghorbani et al. (2018)
- **This work: General asymptotic guarantees for variational Bayes**
- Recent work on convergence rate: Alquier & Ridgway (2017), Zhang & Gao (2017), Pati et al. (2017), Yang et al. (2017), Chérief-Abdellatif et al. (2018), Fan et al. (2018), Ghorbani et al. (2018), Jaiswal et al. (2019)

More details in the paper

Wang, Y., & Blei, D. M. (2017).
Frequentist consistency of variational Bayes.
arXiv preprint arXiv:1705.03439.
To appear in *Journal of the American Statistical Association*.

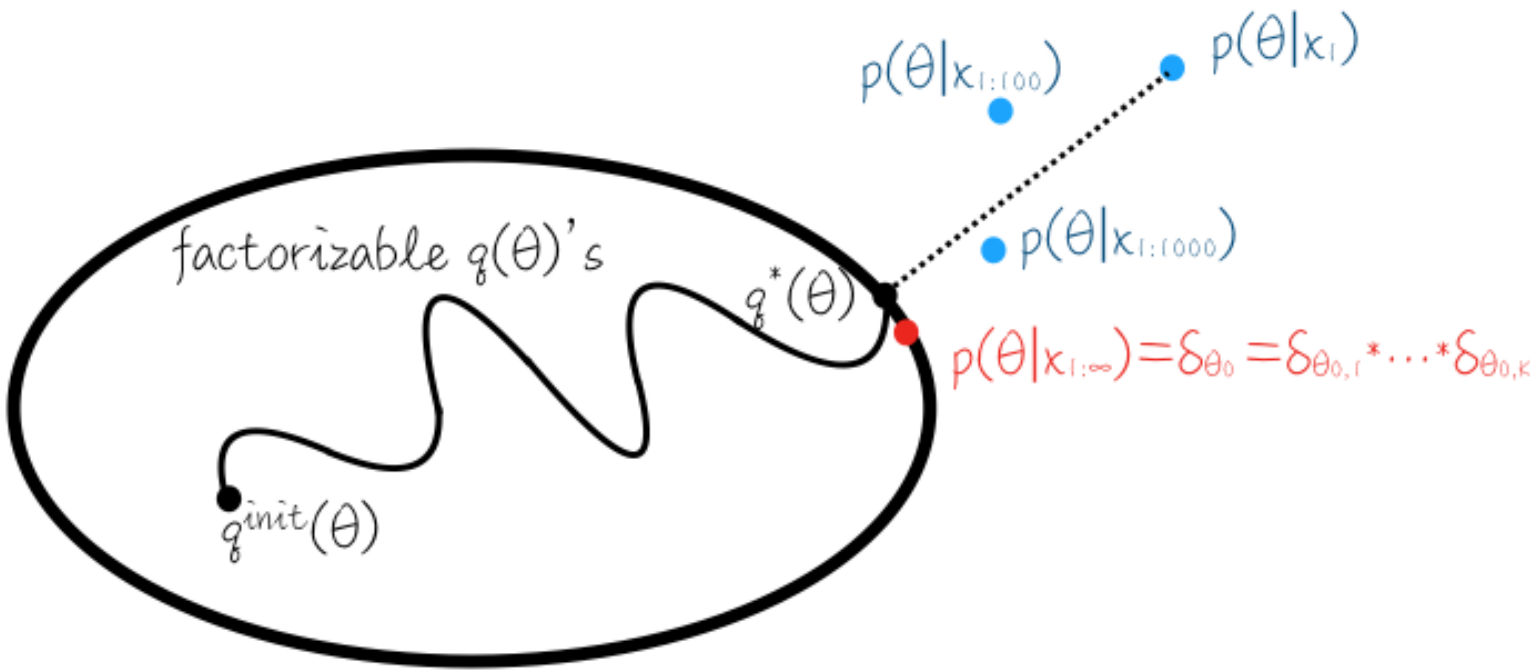
- What if # of latent variables increases with data size?
- What about a different approximating family?
- How to apply the main theorem to specific models?
- What happens in practice?

Summary



- We present **general asymptotic guarantees** for variational Bayes:
  - The VB posterior is consistent and asymptotically normal.
  - The VB estimate is consistent and asymptotically normal.

Thank you!



Wang, Y., & Blei, D. M. (2017). Frequentist consistency of variational Bayes. arXiv preprint arXiv:1705.03439. To appear in *Journal of the American Statistical Association*.

Why variational Bayes gets around $p(\boldsymbol{x})$?

$$q^*(\boldsymbol{\theta}) = \arg\min_{q \in \mathcal{Q}} \mathrm{KL}(q(\boldsymbol{\theta}) || p(\boldsymbol{\theta} \mid \boldsymbol{x}))$$

$$\begin{aligned}
&\mathrm{KL}(q(\boldsymbol{\theta}) || p(\boldsymbol{\theta} \mid \boldsymbol{x})) \\
&= \int q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta} \mid \boldsymbol{x})} \, \mathrm{d}\boldsymbol{\theta} \\
&= \int q(\boldsymbol{\theta}) \log [q(\boldsymbol{\theta}) \cdot \frac{p(\boldsymbol{x})}{p(\boldsymbol{\theta}, \boldsymbol{x})}] \, \mathrm{d}\boldsymbol{\theta} \\
&= \int q(\boldsymbol{\theta}) \log q(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} - \int q(\boldsymbol{\theta}) \log p(\boldsymbol{\theta}, \boldsymbol{x}) \, \mathrm{d}\boldsymbol{\theta} + p(\boldsymbol{x})
\end{aligned}$$

$p(\boldsymbol{x})$ does not involve $\boldsymbol{\theta}$ – irrelevant for optimization.

Main theorem: Technical assumptions

1. **(Prior mass)** The prior measure with Lebesgue-density $p(\theta)$ on Θ is continuous and positive on a neighborhood of θ_0. There exists a constant $M_p > 0$ such that $|(\log p(\theta))''| \leq M_p e^{|\theta|^2}$.

2. **(Consistent testability)** For every $\epsilon > 0$ there exists a sequence of tests ϕ_n such that

$$\int \phi_n(x) p(x \mid \theta_0) \, \mathrm{d}z \, \mathrm{d}x \to 0$$

and

$$\sup_{\theta : ||\theta - \theta_0|| \geq \epsilon} \int (1 - \phi_n(x)) p(x \mid \theta) \, \mathrm{d}z \, \mathrm{d}x \to 0.$$

3. **(Local asymptotic normality)** For every compact set $K \subset \mathbb{R}^d$, there exist random vectors Δ_{n,θ_0} bounded in probability and nonsingular matrices V_{θ_0} such that

$$\sup_{h \in K} \left| \log p(x \mid \theta + \delta_n h) - \log p(x \mid \theta) - h^\top V_{\theta_0} \Delta_{n,\theta_0} + \frac{1}{2} h^\top V_{\theta_0} h \right| \overset{P_{\theta_0}}{\to} 0,$$

where δ_n is a $d \times d$ diagonal matrix. We have $\delta_n \to 0$ as $n \to \infty$.

Proof Idea: Consistency of the VB posterior

$$q^*(\boldsymbol{\theta}) = \arg\min_{q \in \mathcal{Q}} \mathrm{KL}(q(\boldsymbol{\theta}) || p(\boldsymbol{\theta} \mid \boldsymbol{x}))$$

1. The exact posterior $p(\boldsymbol{\theta} \mid \boldsymbol{x})$ converges to the truth $\boldsymbol{\delta}_{\boldsymbol{\theta}_0}$.
   [Bernstein–von Mises theorem]
2. The functional $\mathrm{KL}(\cdot || p(\boldsymbol{\theta} \mid \boldsymbol{x}))$ Γ-converges to the functional $\mathrm{KL}(\cdot || \boldsymbol{\delta}_{\boldsymbol{\theta}_0})$.
   [Many analytic approximations]
3. The minimizer $\arg\min_{\mathcal{Q}} \mathrm{KL}(\cdot || p(\boldsymbol{\theta} \mid \boldsymbol{x}))$ converges to $\arg\min_{\mathcal{Q}} \mathrm{KL}(\cdot || \boldsymbol{\delta}_{\boldsymbol{\theta}_0})$.
   [Fundamental theorem of Γ-convergence]
4. $\arg\min_{\mathcal{Q}} \mathrm{KL}(\cdot || \boldsymbol{\delta}_{\boldsymbol{\theta}_0}) = \boldsymbol{\delta}_{\boldsymbol{\theta}_0}$.
   [$\boldsymbol{\delta}_{\boldsymbol{\theta}_0}$ lies in the factorizable family $\mathcal{Q}$]