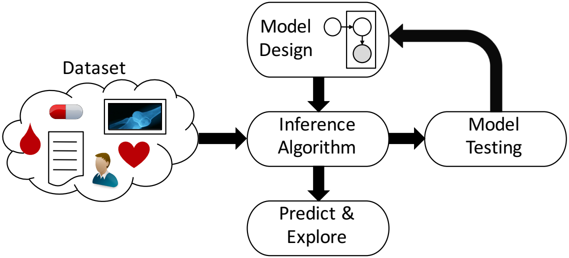
Unbiased Implicit Variational Inference

Michalis K. Titsias & Francisco J. R. Ruiz

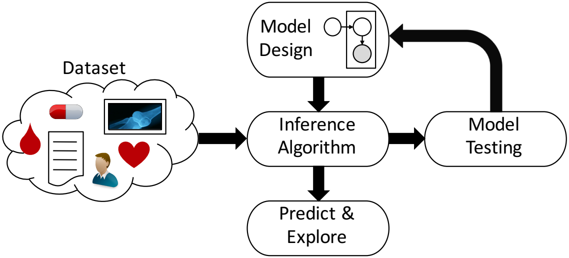
December 17th, 2018

Probabilistic Modeling Pipeline



- Posit generative process with hidden and observed variables
- Given the data, reverse the process to infer hidden variables
- Use hidden structure to make predictions, explore the dataset, etc.

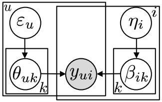
Probabilistic Modeling Pipeline

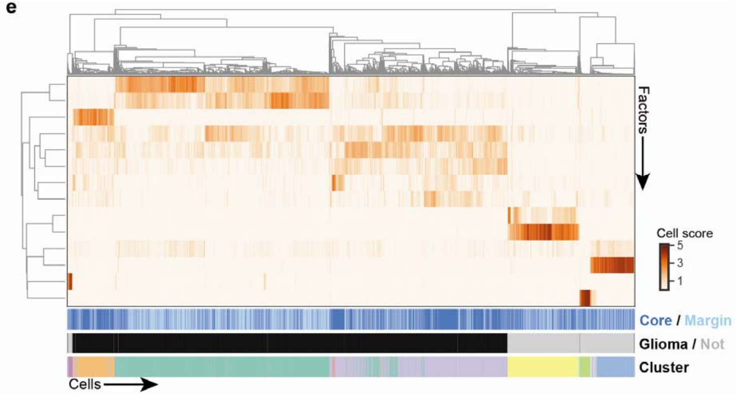


- Incorporate domain knowledge with interpretable components
- Separate assumptions from computation
- Facilitate collaboration with domain experts

Applications: Gene Signature Discovery

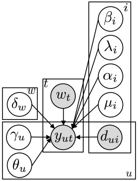
Can we identify *de novo* gene expression patterns in scRNA-seq?

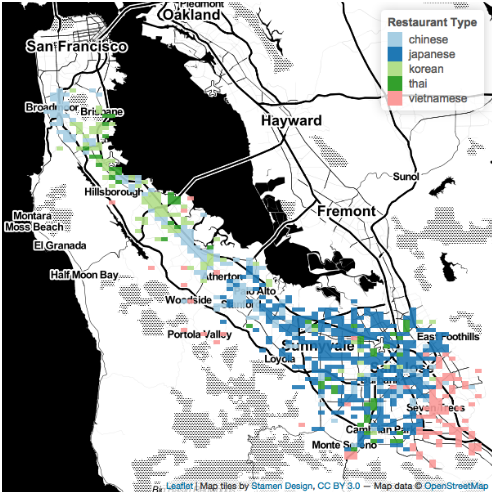




Applications: Consumer Preferences

Can we use mobile location data to find
the most promising location for a new restaurant?

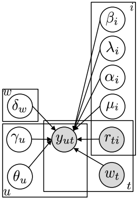


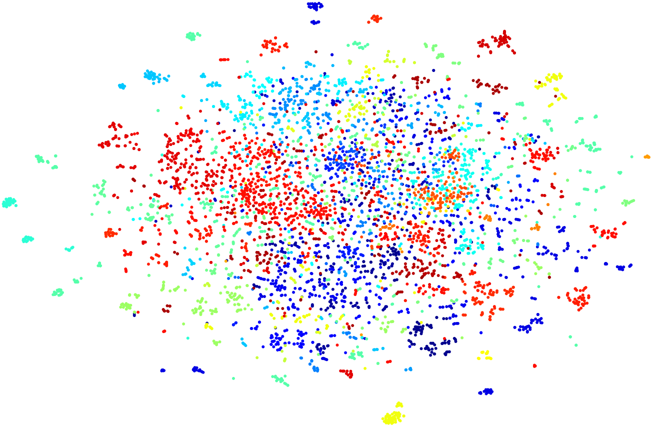


Restaurants in the Bay Area

Applications: Shopping Behavior

Can we use past shopping transactions to learn customer preferences and predict demand as a function of price?





Background: Probabilistic Modeling

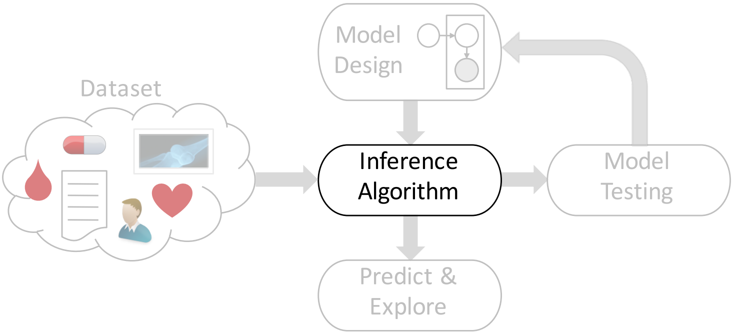
- Latent variables z
- Observations x
- Probabilistic model $p(x, z)$
- Posterior $p(z \mid x) = \frac{p(x,z)}{\int p(x,z)dz}$

Background: Probabilistic Modeling

$$p(z \mid x) = \frac{p(x, z)}{\int p(x, z) dz}$$

- The posterior allows us to explore the data and make predictions
- Approximating the posterior is the central challenge of Bayesian inference

Inference



Dataset
Model
Design
Inference
Algorithm
Model
Testing
Predict &
Explore

Background: Variational Inference

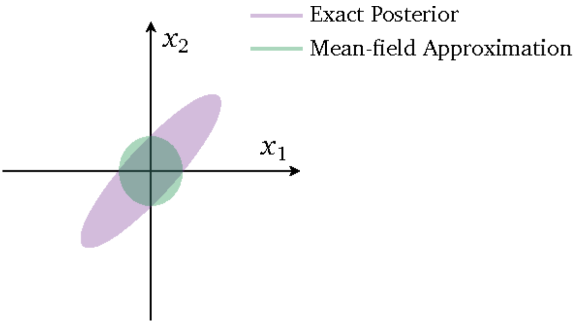
- Variational inference approximates the posterior
- Find simpler distribution $q_\theta(z) \approx p(z \mid x)$
- Use KL divergence to measure similarity between $q_\theta(z)$ and $p(z \mid x)$
- Minimize KL divergence w.r.t. variational parameters θ

Background: Mean-Field Variational Inference

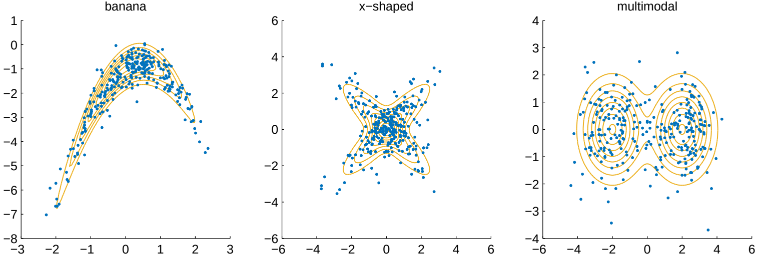
- Classical VI: Mean-field variational distribution:

$$q_\theta(z) = \prod_n q_{\theta_n}(z_n)$$

- Simple, but might not be accurate

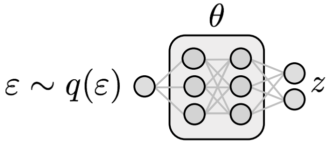


Our Goal: More Expressive Variational Distributions



Blue dots: samples from $q_\theta(z)$

Variational Inference with Implicit Distributions



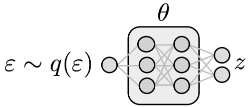
- Easy to draw samples from $q_\theta(z)$:

$$\text{sample } \varepsilon \sim q(\varepsilon); \qquad \text{set } z = f_\theta(\varepsilon)$$

- Cannot evaluate the density $q_\theta(z)$

- Flexible distribution due to the non-linear transformation

VI with Implicit Distributions is Hard



- The VI objective is the ELBO (equivalent to minimizing KL),

$$\mathcal{L}(\theta) = \mathbb{E}_{q_\theta(z)}\Big[\underbrace{\log p(x,z)}_{\text{model}} - \underbrace{\log q_\theta(z)}_{\text{entropy}}\Big]$$

- Gradient of the objective $\nabla_\theta \mathcal{L}(\theta)$ (reparameterization)

$$\nabla_\theta \mathcal{L}(\theta) = \mathbb{E}_{q(\varepsilon)}\Big[\nabla_z \left(\log p(x,z) - \log q_\theta(z)\right)\Big|_{z=f_\theta(\varepsilon)} \times \nabla_\theta f_\theta(\varepsilon)\Big]$$

- Monte Carlo estimates require $\nabla_z \log q_\theta(z)$ (not available)

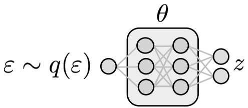
Unbiased Implicit Variational Inference

- We describe how to obtain an unbiased Monte Carlo estimator
- We avoid density ratio estimation
- Key ideas:
  1. Semi-implicit construction of $q_\theta(z)$
  2. Gradient of the entropy component as an expectation,

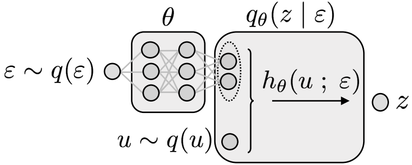
$$\nabla_z \log q_\theta(z) = \mathbb{E}_{\mathrm{distrib}(\cdot)}\left[\mathrm{function}(z, \cdot)\right]$$

UIVI Step 1: Semi-Implicit Distribution

- Implicit distribution:

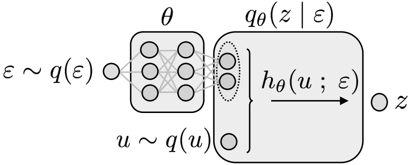


- (Semi-)implicit distribution:



UIVI Step 1: Semi-Implicit Distribution

- (Semi-)implicit distribution

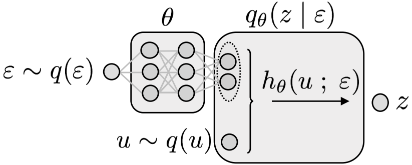


- **Example:** The conditional $q_\theta(z \,|\, \varepsilon)$ is a Gaussian,

$$q_\theta(z \,|\, \varepsilon) = \mathcal{N}\left(z \,|\, \mu_\theta(\varepsilon), \Sigma_\theta(\varepsilon)\right)$$

UIVI Step 1: Semi-Implicit Distribution

- (Semi-)implicit distribution



- The distribution $q_\theta(z)$ is still **implicit**,
  - Easy to sample,

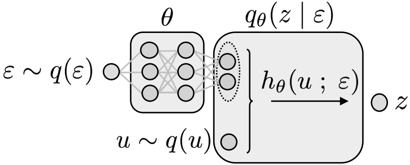
$$\begin{aligned}&\text{sample } \varepsilon \sim q(\varepsilon),\\ &\text{obtain } \mu_\theta(\varepsilon) \text{ and } \Sigma_\theta(\varepsilon)\\ &\text{sample } z \sim \mathcal{N}\left(z \mid \mu_\theta(\varepsilon), \Sigma_\theta(\varepsilon)\right)\end{aligned}$$

  - The variational distribution $q_\theta(z)$ is not tractable,

$$q_\theta(z) = \int q(\varepsilon) q_\theta(z \mid \varepsilon) d\varepsilon$$

UIVI Step 1: Semi-Implicit Distribution

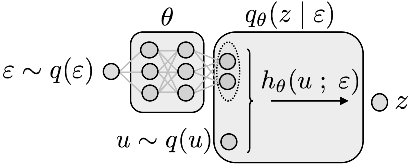
- (Semi-)implicit distribution



- **Assumptions** on the conditional $q_\theta(z \mid \varepsilon)$:
  - Reparameterizable
  - Tractable gradient $\nabla_z \log q_\theta(z \mid \varepsilon)$
    Note: this is different from $\nabla_z \log q_\theta(z)$ (still intractable)

UIVI Step 1: Semi-Implicit Distribution

- (Semi-)implicit distribution



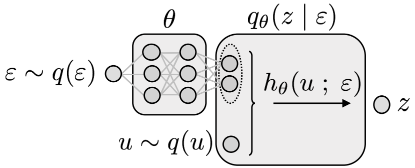
- The Gaussian meets both assumptions:
  - Reparameterizable,

$$u \sim \mathcal{N}(u \mid 0, I), \qquad z = h_\theta(u\,;\,\varepsilon) = \mu_\theta(\varepsilon) + \Sigma_\theta(\varepsilon)^{1/2} u$$

  - Tractable gradient,

$$\nabla_z \log q_\theta(z \mid \varepsilon) = -\Sigma_\theta(\varepsilon)^{-1}(z - \mu_\theta(\varepsilon))$$

UIVI Step 2: Gradient as Expectation



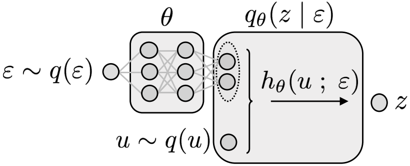
- Goal: Estimate the gradient of the entropy component, $\nabla_z \log q_\theta(z)$
- Rewrite as an expectation,

$$\nabla_z \log q_\theta(z) = \mathbb{E}_{q_\theta(\varepsilon' \mid z)} \left[\nabla_z \log q_\theta(z \mid \varepsilon')\right]$$

- Form Monte Carlo estimate,

$$\nabla_z \log q_\theta(z) \approx \nabla_z \log q_\theta(z \mid \varepsilon'), \qquad \varepsilon' \sim q_\theta(\varepsilon' \mid z)$$

UIVI: Full Algorithm

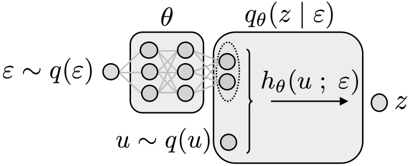


- The gradient of the ELBO is

$$\nabla_\theta \mathcal{L}(\theta) = \mathbb{E}_{q(\varepsilon)q(u)}\left[\nabla_z \left(\log p(x,z) - \log q_\theta(z)\right)\Big|_{z=h_\theta(u\,;\,\varepsilon)} \times \nabla_\theta h_\theta(u\,;\,\varepsilon)\right]$$

- Estimate the gradient based on samples:
  1. Sample $\varepsilon \sim q(\varepsilon)$, $u \sim q(u)$ (standard Gaussians)
  2. Set $z = h_\theta(\varepsilon\,;\,u) = \mu_\theta(\varepsilon) + \Sigma_\theta(\varepsilon)^{1/2} u$
  3. Evaluate $\nabla_z \log p(x,z)$ and $\nabla_\theta h_\theta(u\,;\,\varepsilon)$
  4. Sample $\varepsilon' \sim q_\theta(\varepsilon' \mid z)$
  5. Approximate $\nabla_z \log q_\theta(z) \approx \nabla_z \log q_\theta(z \mid \varepsilon')$

UIVI: The Reverse Conditional

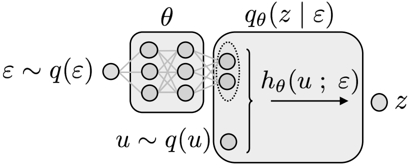


- The distribution $q_\theta(\varepsilon' \mid z)$ is the **reverse conditional**
  The conditional is $q_\theta(z \mid \varepsilon)$
- Sample from $q_\theta(\varepsilon' \mid z)$ using HMC, targeting

$$q(\varepsilon' \mid z) \propto q(\varepsilon')q_\theta(z \mid \varepsilon')$$

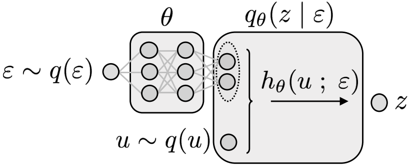
- Problem: HMC is slow... How to accelerate this?

UIVI: The Reverse Conditional



- Recall the UIVI algorithm,
  1. Sample $\varepsilon \sim q(\varepsilon)$, $u \sim q(u)$ (standard Gaussians)
  2. Set $z = h_\theta(\varepsilon\,;\,u) = \mu_\theta(\varepsilon) + \Sigma_\theta(\varepsilon)^{1/2} u$
  3. Evaluate $\nabla_z \log p(x, z)$ and $\nabla_\theta h_\theta(u\,;\,\varepsilon)$
  4. Sample $\varepsilon' \sim q_\theta(\varepsilon' \mid z)$
  5. Approximate $\nabla_z \log q_\theta(z) \approx \nabla_z \log q_\theta(z \mid \varepsilon')$
- We have that $(\varepsilon, z) \sim q_\theta(\varepsilon, z) = q(\varepsilon) q_\theta(z \mid \varepsilon) = q_\theta(z) q_\theta(\varepsilon \mid z)$
- Thus, ε is a sample from $q_\theta(\varepsilon \mid z)$
- To accelerate sampling $\varepsilon' \sim q(\varepsilon' \mid z)$, initialize HMC at ε

UIVI: The Reverse Conditional

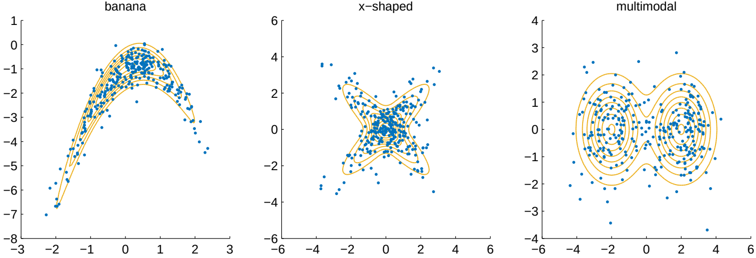


- Sample from $q_\theta(\varepsilon' \mid z)$ using HMC targeting

$$q(\varepsilon' \mid z) \propto q(\varepsilon') q_\theta(z \mid \varepsilon')$$

- Initialize HMC at stationarity (using ε)
- A few HMC iterations to reduce correlation between ε and ε'

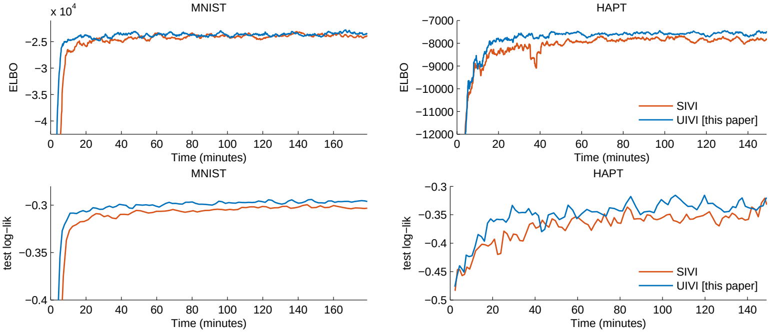
Toy Experiments



Blue dots: samples from $q_\theta(z)$

Experiments: Multinomial Logistic Regression

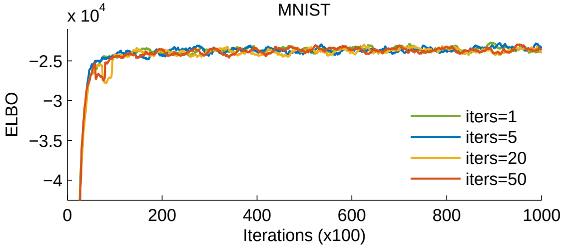
$$p(x, z) = p(z) \prod_{n=1}^{N} \frac{\exp\{x_n^\top z_{y_n} + z_{0y_n}\}}{\sum_k \exp\{x_n^\top z_k + z_{0k}\}}$$



UIVI provides better ELBO and predictive performance than SIVI

Experiments: Multinomial Logistic Regression

$$p(x, z) = p(z) \prod_{n=1}^{N} \frac{\exp\{x_n^\top z_{y_n} + z_{0y_n}\}}{\sum_k \exp\{x_n^\top z_k + z_{0k}\}}$$



Number of HMC iterations does not significantly impact results

Experiments: VAE

- Model is $p_\phi(x,z) = \prod_n p(z_n) p_\phi(x_n \,|\, z_n)$
- Amortized variational distrib. $q_\theta(z_n \,|\, x_n) = \int q(\varepsilon_n) q_\theta(z_n \,|\, \varepsilon_n, x_n) d\varepsilon_n$
- Goal: Find model parameters ϕ and variational parameters θ

method	average test log-likelihood MNIST	Fashion-MNIST
Explicit (standard VAE)	-98.29	-126.73
SIVI	-97.77	-121.53
UIVI [this paper]	$\mathbf{-94.09}$	$\mathbf{-110.72}$

UIVI provides better ELBO and predictive performance

Conclusion

- UIVI approximates the posterior with an expressive variational distribution
- The variational distribution is implicit
- UIVI directly optimizes the ELBO
- Good results on Bayesian multinomial logistic regression and VAEs

Proof of the Key Equation

- Goal: Prove that

$$\nabla_z \log q_\theta(z) = \mathbb{E}_{q_\theta(\varepsilon \mid z)} \left[\nabla_z \log q_\theta(z \mid \varepsilon)\right]$$

- Start with log-derivative identity,

$$\nabla_z \log q_\theta(z) = \frac{1}{q_\theta(z)} \nabla_z q_\theta(z)$$

- Apply the definition of $q_\theta(z)$ through a mixture,

$$\nabla_z \log q_\theta(z) = \frac{1}{q_\theta(z)} \int \nabla_z q_\theta(z \mid \varepsilon) q(\varepsilon) d\varepsilon$$

- Apply the log-derivative identity on $q_\theta(z \mid \varepsilon)$,

$$\nabla_z \log q_\theta(z) = \frac{1}{q_\theta(z)} \int q_\theta(z \mid \varepsilon) q(\varepsilon) \nabla_z \log q_\theta(z \mid \varepsilon) d\varepsilon.$$

- Apply Bayes' theorem

SIVI

- SIVI optimizes a lower bound of the ELBO,

$$\mathcal{L}_{\mathrm{SIVI}}^{(L)}(\theta) = \mathbb{E}_{\varepsilon \sim q(\varepsilon)}\left[\mathbb{E}_{z \sim q_\theta(z \mid \varepsilon)}\left[\mathbb{E}_{\varepsilon^{(1)},\ldots,\varepsilon^{(L)} \sim q(\varepsilon)}\left[\log p(x, z) - \log\left(\frac{1}{L+1}\left(q_\theta(z \mid \varepsilon) + \sum_{\ell=1}^{L} q_\theta(z \mid \varepsilon^{(\ell)})\right)\right)\right]\right]\right]$$