ST451 - Lent term
Bayesian Machine Learning

Kostas Kalogeropoulos

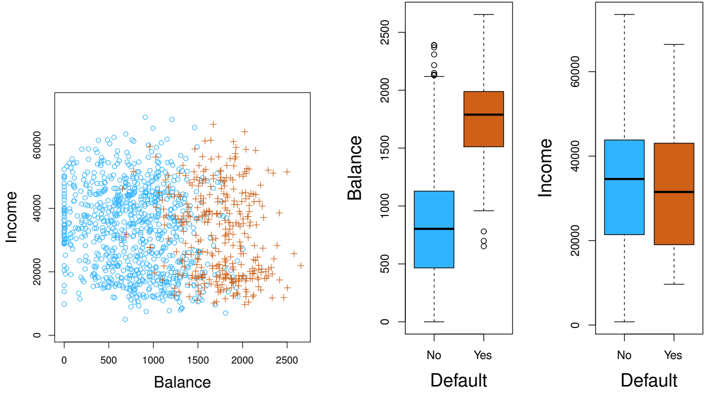
Variational Bayes / Approximation

Summary of last lecture

- Classification Problem: Categorical y, mixed X.
- Generative models: Specify $\pi(y)$ with 'prior' probabilities, then $\pi(X|y)$ for each category of y, e.g. LDA
- Discriminative models: Logistic regression, maximum likelihood via Newton-Raphson.
- Bayesian Logistic Regression: Use of Laplace approximation similar results with MLE.
- Prediction Assesment: Accuracy, Area under the ROC curve and log score rule.

Motivating Example 1

'Default' dataset consist of three variables: annual income, credit card balance and whether or not the person has defaulted in his/her credit card.



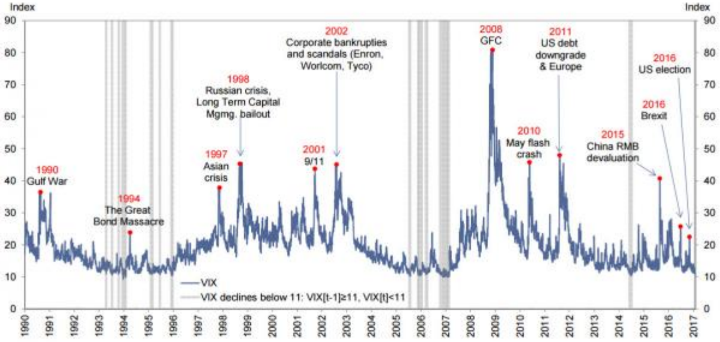
The aim is to build a model to predict whether a person will default based on annual income and monthly credit card balance.

Motivating Example 2

Volatility Index (VIX) provided by Chicago Board of Exchange (CBOE). Derived from the S&P 500 index options. Represents market's expectation of its future 30-day volatility. A measure of market risk.

Exhibit 3: VIX levels 1990-present

Shaded events represent VIX declining below 11, i.e. VIX[t-1]≥11, VIX[t]<11. Daily data from 1/2/1990– 1/27/2017.

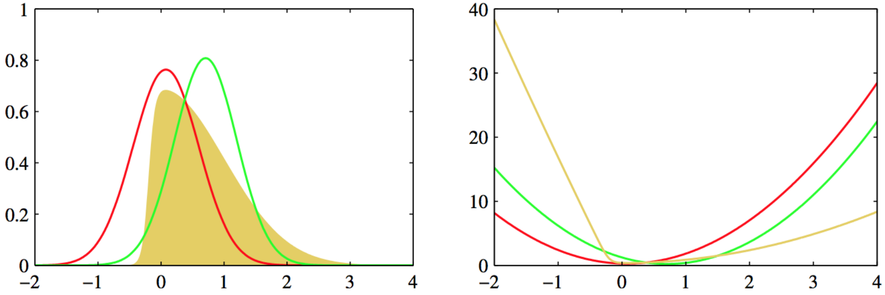


Source: Chicago Board Options Exchange (CBOE). Goldman Sachs Global Investment Research.

Variational vs Laplace Approximation

In both of these examples (and many others), the posterior and the exact distribution of MLEs are typically intractable.

Last week we used the Laplace approximation. This week we will look into the Variational approximation. Below we see these approximations in terms of the pdfs (left) and negative log scale (right).



Outline

Outline

1. Essentials of Variational Inference

2. Examples with mean field approximation

3. Automatic Variational Inference

Main idea

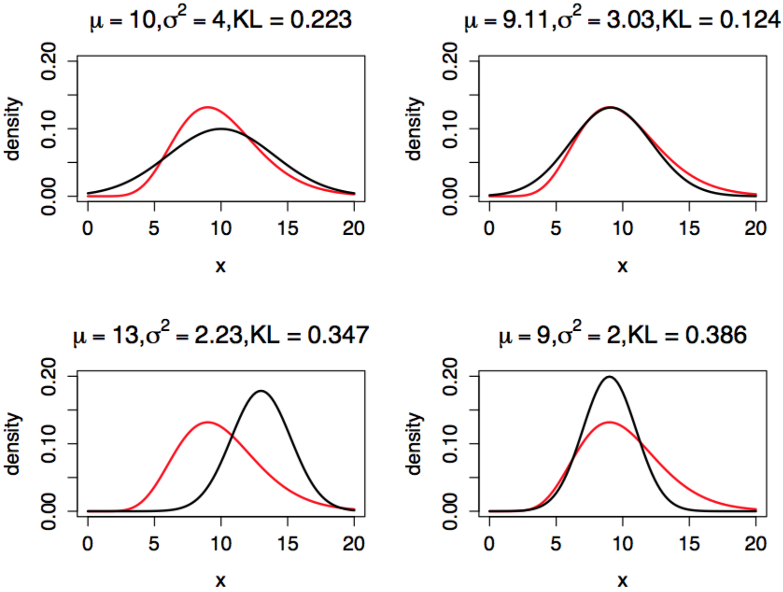
Ideally we would like to use the posterior $\pi(\theta|y)$. But it is not always available.

Laplace approximation uses a Normal distribution based on a single point (the mode).

Variational approximation usually follows the steps below

1. Consider a family of distributions $q(\theta|y, \phi)$ with parameters ϕ, e.g. Normal, Gamma etc.
2. Select ϕ such that $q(\theta|y, \phi)$ is as close as possible to $\pi(\theta|y)$.

Approximating a Gamma(0, 1) with a N(μ, σ^2)



Variational Bayes

As close as possible translates into minimising the KL divergence

$$\mathrm{KL}(q||\pi) = \int q(\theta|y, \phi) \log \frac{q(\theta|y, \phi)}{\pi(\theta|y)} d\theta$$

It can be shown that $\mathrm{KL}(q||\pi) \geq 0$ and $\mathrm{KL}(q||\pi) = 0$ iff $q \overset{D}{=} \pi$.

But $\pi(\theta|y)$ is intractable so the above is not very useful. Instead we consider the evidence lower bound (ELBO)

$$\mathrm{ELBO}(\phi) = \int q(\theta|y, \phi) \log \frac{f(y|\theta)\pi(\theta)}{q(\theta|y, \phi)} d\theta$$

Variational Bayes (cont'd)

Note that the sum of $\text{KL}(q||\pi)$ and $\text{ELBO}(\phi)$ is equal to

$$\int q(\theta|y,\phi)\log\frac{q(\theta|y,\phi)}{\pi(\theta|y)}d\theta + \int q(\theta|y,\phi)\log\frac{f(y|\theta)\pi(\theta)}{q(\theta|y,\phi)}d\theta$$

$$\int q(\theta|y,\phi)\left\{\log\frac{q(\theta|y,\phi)}{\pi(\theta|y)} + \log\frac{f(y|\theta)\pi(\theta)}{q(\theta|y,\phi)}\right\}d\theta$$

$$\int q(\theta|y,\phi)\left\{\log\frac{f(y|\theta)\pi(\theta)}{\pi(\theta|y)}\right\}d\theta = \int q(\theta|y,\phi)\log\pi(y)d\theta$$

$$\log\pi(y)\int q(\theta|y,\phi)d\theta = \log\pi(y).$$

Notes:

1. Since $\text{KL}(q||p) \geq 0$, we get that $\log\pi(y) \geq \text{ELBO}(\phi)$. Hence, the name evidence lower bound (ELBO).
2. The sum above is independent of ϕ so minimising $\text{KL}(q||p)$ is the same as maximising $\text{ELBO}(\phi)$.

Mean field approximation

Ofcourse the minimum KL divergence can still be large, depends on the choice of q.

The most widely used choice is known as the mean field approximation and assumes that q can be factorised into some components

$$q(\theta|y,\phi) = \prod_i q(\theta_i|y,\phi_i)) = \prod_i q(\theta_i)$$

This results in an algorithm that iteratively maximises ELBO(ϕ) wrt each $q(\theta_i)$ keeping $q(\theta_j)$, $j \neq i$, or else $q(\theta_{-i})$ fixed.

We refer to each $q(\theta_i)$ as VB component.

Mean field approximation (cont'd)

Let $\theta = (\theta_i, \theta_{-i})$, $q(\theta|y,\phi) = q(\theta_i)q(\theta_{-i})$, and $\pi(y,\theta) = f(y|\theta)\pi(\theta)$. Then

$$\begin{aligned}
\text{ELBO} &= \int\int q(\theta_i)q(\theta_{-i}) \log \frac{\pi(y,\theta)}{q(\theta_i)q(\theta_{-i})} d\theta_{-i} d\theta_i \\
&= \int q(\theta_i) \left\{ \int \log \pi(y,\theta) q(\theta_{-i}) d\theta_{-i} \right\} d\theta_i \\
&- \int\int q(\theta_i)q(\theta_{-i}) \log q(\theta_i) d\theta_{-i} d\theta_i \\
&- \int\int q(\theta_i)q(\theta_{-i}) \log q(\theta_{-i}) d\theta_{-i} d\theta_i
\end{aligned}$$

Note that $\int \log \pi(y,\theta) q(\theta_{-i}) d\theta_{-i} = \mathbb{E}_{q(\theta_{-i})}[\log \pi(y,\theta)]$ and define $q^*(\theta_i) = \frac{1}{z} \exp\left(\mathbb{E}_{q(\theta_{-i})}[\log \pi(y,\theta)]\right)$. Then we can write

$$\begin{aligned}
\text{ELBO} &= \int q(\theta_i) \log \frac{q^*(\theta_i)}{q(\theta_i)} d\theta_i - \int q(\theta_{-i}) \log q(\theta_{-i}) d\theta_{-i} + c \\
&= -\text{KL}(q(\theta_i)||q^*(\theta_i)) - \int q(\theta_{-i}) \log q(\theta_{-i}) d\theta_{-i} + c
\end{aligned}$$

Mean field approximation (cont'd)

Hence maximising ELBO wrt $q(\theta_i)$ is the same as minimising $\mathrm{KL}(q(\theta_i)||q^*(\theta_i)$ wrt $q(\theta_i)$.

$\mathrm{KL}(q(\theta_i)||q^*(\theta_i)$ is minimised when

$$q(\theta_i) = q^*(\theta_i) = \frac{1}{z}\exp\left(\mathbb{E}_{q(\theta_{-i})}[\log\pi(y,\theta)]\right)$$

This suggests the following algorithm

1. Initialise each $q(\theta_i)$ to the prior $\pi(\theta_i)$,
2. For each θ_i, update $q(\theta_i)$, based on $\mathbb{E}_{q(\theta_{-i})}[\log\pi(y,\theta)]$,
3. Continue until ELBO converges.

Generally well behaved algorithm when it can be derived, i.e. when we can recognise the $q^*(\theta_i)$'s.

Outline

Example: N(μ, τ^{-1})

Let $y = (y_1, \ldots, y_n)$ (y_i independent) from N(μ, τ^{-1}). Assign N$(\mu_0, (\lambda_0\tau)^{-1})$ as prior for $\theta|\tau$ and Gamma(α_0, β_0) for τ. The posterior can be derived but let's consider its variational approximation.

Assume $q(\theta) = q(\mu)q(\tau)$

The log joint density is (c will be denoting constant from now on)

$$\begin{aligned} \log \pi(y, \theta) &= \tfrac{n}{2}\log\tau - \tfrac{\tau}{2}\sum_{i=1}^{n}(y_i - \mu)^2 + \tfrac{1}{2}\log\tau - \tfrac{\tau\lambda_0}{2}(\mu - \mu_0)^2 \\ &+ (\alpha_0 - 1)\log\tau - \beta_0\tau + c \end{aligned}$$

We will now derive the VB components $q(\mu)$ and $q(\tau)$

Example: $\mathrm{N}(\mu, \tau^{-1})$ - $q(\mu)$

For $q(\mu)$ we can focus on the terms involving μ.

$$
\begin{aligned}
\log q(\mu) &= \mathbb{E}_{q(\tau)}\left[-\tfrac{\tau}{2}\sum_{i=1}^{n}(y_i-\mu)^2 - \tfrac{\tau\lambda_0}{2}(\mu-\mu_0)^2\right] + c \\
&= -\frac{\mathbb{E}_{q(\tau)}(\tau)}{2}\left[\sum_{i=1}^{n}(y_i-\mu)^2 + \lambda_0(\mu-\mu_0)^2\right] + c.
\end{aligned}
$$

By inspection we can identify $q(\mu)$ to be the $\mathrm{N}(\mu_\phi, \tau_\phi^{-1})$, where

$$
\begin{aligned}
\mu_\phi &= \frac{\lambda_0\mu_0 + \sum_{i=1}^{n} y_i}{\lambda_0 + n} \\
\tau_\phi &= (\lambda_0 + n)\mathbb{E}(\tau)
\end{aligned}
$$

The quantity $\mathbb{E}(\tau) = \mathbb{E}_{q(\tau)}(\tau)$ will be provided by the derivation of $q(\tau)$

Example: N(μ, τ^{-1}) - $q(\tau)$

For $q(\tau)$ we take as before all the terms involving τ

$$
\begin{aligned}
\log q(\tau) &= \mathbb{E}_{q(\mu)}\left[\left(\frac{n+1}{2} + \alpha_0 - 1\right)\log\tau - \beta_0\tau - \frac{\tau}{2}\sum_{i=1}^{n}(y_i - \mu)^2\right. \\
&\quad \left. - \frac{\tau\lambda_0}{2}(\mu - \mu_0)^2\right] + c \\
&= \left(\frac{n+1}{2} + \alpha_0 - 1\right)\log\tau - \beta_0\tau \\
&\quad - \frac{\tau}{2}\mathbb{E}_{q(\mu)}\left[\sum_{i=1}^{n}(y_i - \mu)^2 + \lambda_0(\mu - \mu_0)^2\right]
\end{aligned}
$$

By inspection we can identify $q(\tau)$ to be the Gamma$(\alpha_\phi, \beta_\phi)$, where

$$
\begin{aligned}
\alpha_\phi &= \alpha_0 + \frac{n+1}{2}, \\
\beta_\phi &= \beta_0 + \frac{1}{2}\left(S_y^2 - 2\mathbb{E}(\mu)S_y + n\mathbb{E}(\mu^2)\right) + \frac{\lambda_0}{2}\left(\mu_0^2 - 2\mu_0\mathbb{E}(\mu) + \mathbb{E}(\mu^2)\right), \\
\mathbb{E}(\mu) &= \mathbb{E}_{q(\mu)}(\mu), \\
S_y &= \sum_i y_i, \quad S_y^2 = \sum_i y_i^2.
\end{aligned}
$$

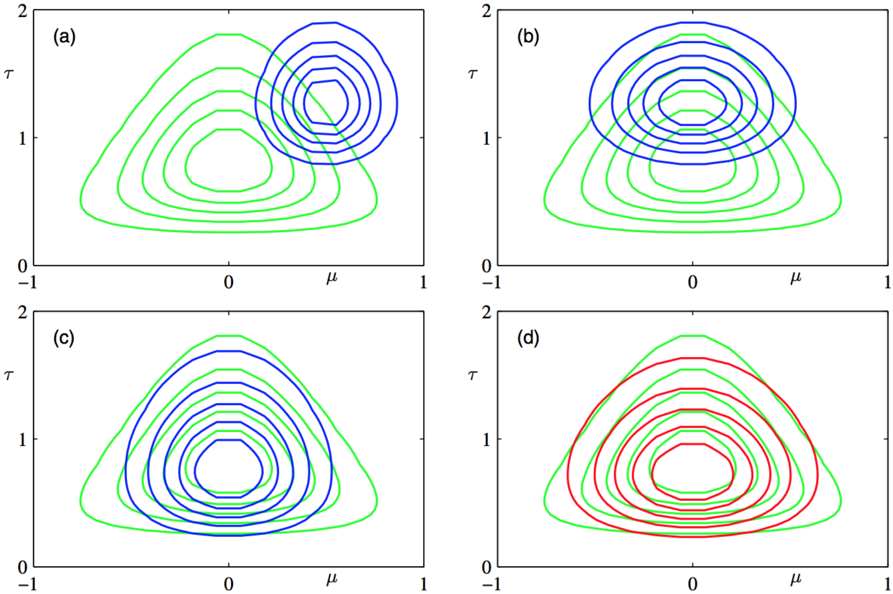
Example: N(μ, τ^{-1}) - overall algorithm

So overall we set $q(\mu) =$N$(\mu_\phi, \tau_\phi^{-1})$ and $q(\tau) =$Gamma$(\alpha_\phi, \beta_\phi)$.

Then we look for the q parameters $\phi = (\mu_\phi, \tau_\phi, \alpha_\phi, \beta_\phi)$ that maximise ELBO by first initialising and then iteratively updating

$$
\begin{aligned}
\mu_\phi &= \frac{\lambda_0 \mu_0 + \sum_{i=1}^n y_i}{\lambda_0 + n}, \\
\tau_\phi &= (\lambda_0 + n)\mathbb{E}(\tau), \\
\alpha_\phi &= \alpha_0 + \tfrac{n+1}{2}, \\
\beta_\phi &= \beta_0 + \tfrac{1}{2}\left(S_y^2 - 2\mathbb{E}(\mu)S_y + n\mathbb{E}(\mu^2)\right) + \tfrac{\lambda_0}{2}\left(\mu_0^2 - 2\mu_0\mathbb{E}(\mu) + \mathbb{E}(\mu^2)\right), \\
\mathbb{E}(\tau) &= \alpha_\phi / \beta_\phi, \\
\mathbb{E}(\mu) &= \mu_\phi, \\
\mathbb{E}(\mu^2) &= \frac{1}{\tau_\phi} + \mu_\phi^2.
\end{aligned}
$$

Graphical illustration of the previous algorithm



Remarks on mean field approximation

- Possible to extend this to linear regression and other exponential family models including logistic regression. In some cases some model-specific tricks are needed.
- Generally provides a good approximation to the mean but underestimates the variance (as it cannot capture posterior dependencies).
- Model selection can be done by optimising each model separately and then comparing their ELBO's. Prediction is also straightforward
- A fair amount of derivations are required and it is easy to lose the big picture. A black box would be useful.

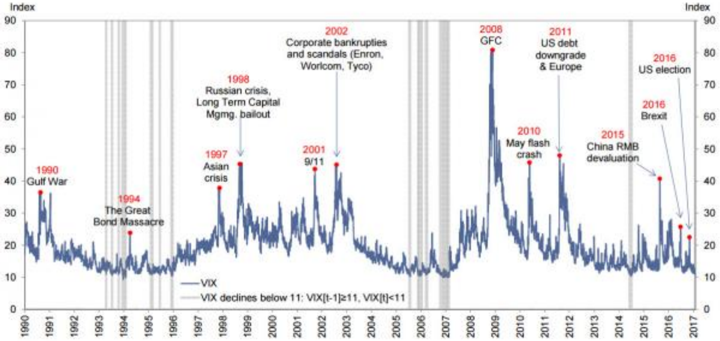
Outline

Motivating Example 2

Volatility Index (VIX) provided by Chicago Board of Exchange (CBOE). Derived from the S&P 500 index options. Represents market's expectation of its future 30-day volatility. A measure of market risk.

Exhibit 3: VIX levels 1990-present

Shaded events represent VIX declining below 11, i.e. VIX[t-1]≥11, VIX[t]<11. Daily data from 1/2/1990– 1/27/2017.



Source: Chicago Board Options Exchange (CBOE), Goldman Sachs Global Investment Research.

Modelling VIX

VIX trajectories are mean reverting and autocorrelated. A simple model that captures these stylised facts is

$$Y_t = Y_{t-1} + \kappa(\mu - Y_{t-1})\delta + \sigma\epsilon_t,$$

where Y_t is VIX at time t, and ϵ_t are independent error terms.

μ : long term mean, σ volatility of volatility, κ mean reversion speed.

A convenient option is to set $\epsilon_t \sim N(0, \delta)$ as it gives closed form posterior and distribution of MLEs.

But it is not a good choice for the spikes that we observe. A *t* distribution with low degrees of freedom is a much better option, yet intractable. No much room for the previous tricks either.

Hurdles towards Automatic Variational Inference

The procedure for variational inference can be automated. The aim is to be able to specify the likelihood and the prior and nothing else.

But even under the framework of mean field approximation, there are two main hurdles:

- Each θ_i may be given a different distribution depending also on its range, e.g. $\mathbb{R},\ \mathbb{R}^+,\ [0,1]$ etc.
- It is not always possible to derive the algorithm presented earlier. Even if it was possible its final form would depend on the model, so cannot be automated.

The recent Automatic Differentiation Variational Inference (ADVI) approach of Kucukelbir et al (2016) addresses those issues.

Transformation to the $\mathbb{R}^p$

The first step is to transform all the θ_i components to the real line using log or logit transformations where needed. Hence we transform from the parameter space Θ to $\mathbb{R}^p$ via the function $T(\cdot)$.

We can then define $\zeta := T(\theta)$ and the the joint density (likelihood times prior) can be written as

$$\pi(y,\theta) = \pi\big(y, T^{-1}(\zeta)\big)\left|\det J_{T^{-1}(\zeta)}\right|$$

Given that ζ is defined in $\mathbb{R}^p$, we can assign the Normal distribution on it. The default option is to assume p independent Normals.

$$q(\zeta|y,\phi) = \prod_{i=1}^{p} N(\zeta_i|\mu_i,\sigma_i^2)$$

Note that the corresponding $q(\theta|y)$ is not necessarily Normal.

Optimisation

Numerical optimisation can be used. It is essential to calculate the gradient of ELBO(ϕ) to obtain good performance. We can write

$$\begin{aligned}
\text{ELBO}(\phi) &= \int q(\theta|y,\phi)\log\frac{f(y|\theta)\pi(\theta)}{q(\theta|y,\phi)}d\theta \\
&= \int q(\theta|y,\phi)\log f(y|\theta)d\theta - \int q(\theta|y,\phi)\log\frac{q(\theta|y,\phi)}{\pi(\theta)}d\theta \\
&= \mathbb{E}_{q(\phi)}\left[\log f(y|\theta)\right] - \text{KL}\left[q(\theta|y,\phi)||\pi(\theta)\right]
\end{aligned}$$

The second term in the expression above can be derived analytically (Kucukelbir et al 2016), so automatic differentiation can be used.

The first term is tricky because it doesn't have a closed form. Automatic differentiation can only be used for terms inside the expectation.

Reparameterisation

To see this note the we want to calculate

$$\nabla_\phi \mathbb{E}_{q(\phi)}\left[\log f(y|\theta)\right] = \nabla_\phi \mathbb{E}_{q(\phi)}\left[f\big(y|\,T^{-1}(\zeta)\big)\left|\det J_{T^{-1}(\zeta)}\right|\right],$$

i.e. an expectation wrt ϕ.

This problem can be addressed by (further) reparameterisation, i.e. by standardising the ζ's

$$\eta_i = \frac{\zeta_i - \mu_i}{\sigma_i}, \quad i = 1, \dots, k.$$

Now we can write (for the corresponding transformation T_ϕ from θ to η)

$$\nabla_\phi \mathbb{E}_{q(\phi)}\left[\log f(y|\theta)\right] = \nabla_\phi \mathbb{E}_{q(\eta)}\left[f\big(y|\,T_\phi^{-1}(\eta)\big)\left|\det J_{T_\phi^{-1}(\eta)}\right|\right],$$

and we can see that ϕ now appears only inside the expectation.

Stochastic Gradient Descent (SGD)

The following trick allows quick and automatic estimates of the gradient of ELBO using Monte Carlo and automatic differentiation.

It can be used in state-of-the-art algorithms for big models and big data (e.g. deep networks), such as the stochastic gradient descent.

In such contexts there is no need to calculate the gradient from the entire data, but only from a small batch.

At the moment, in deep learning, this the only scalable method for Bayesian Inference.

Today's lecture - Reading

Bishop: 10.1 10.3 10.6.

Murphy: 21.1 21.2 21.3.1 21.5.

Kucukelbir A., Tran D., Ranganath R., Gelman A., Blei D.M. (2016) Automatic Differentiation Variational Inference. Available on Arxiv