Yale University, Department of Statistics

Variational inference via score matching

August 6, 2019

Contents

1 Introduction

Score matching is a technique to perform nonparametric density estimation. Score matching seeks a density q to minimize a Fisher divergence, defined via

$$\mathcal{L}(q) = \mathbb{E}_p \|\nabla_\theta \log q(\theta) - \nabla_\theta \log p(\theta)\|_2^2 \tag{1}$$

We propose to use this objective to optimize a posterior.

2 Score-matching for flexible posterior inference

Suppose we observe data x and latent variable z. We know the conditional likelihood $p(x|z)$ and the prior density $p(z)$. We seek $p(z|x)$. Suppose we assume that $p(z|x) \propto e^{f(z|x)}$, i.e., the posterior has an exponential family form. Suppose further that $f(z|x) = f_x(z) = \gamma^\top \phi(z)$, where $\gamma \in \mathbb{R}^K$ and $\phi : \mathbb{R}^d \to \mathbb{R}^K$. Then just as in Lemma **??** the objective simplifies to:

$$\int_z p(z|x) \left(\frac{1}{2} \gamma' A(z) \gamma + \gamma' k(z) \right) \, dz \tag{2}$$

which has optimal solution

$$\hat{\gamma} = \bar{A}^{-1}\bar{k} \tag{3}$$

where

$$\bar{A} = \int p(z,x)A(z)\ dz$$
$$\bar{k} = \int p(z,x)k(z)\ dz$$

where $\bar{A}, \bar{k}$ are defined below. But first note a key fact here: the optimal solution requires integrals only over $p(z,x)$, not over $p(z|x)$, since the form in (3) allows the normalizing constant $p(x)$ to drop out. Nevertheless, we probably can't actually compute these integrals over the joint density. If we could, computing $p(x)$ wouldn't be a problem in the first place. We have considered importance-sampling. One issue is that often $p(x|z)$ is extremely small if n is large. Another is that, e.g., in the Gaussian family case, we still need $p(x)$. See Example 1.

Example 1 (Score-matching with $\mathbb{E}_p$ for posterior inference when q has Gaussian form)**.** *Suppose q has form $N(\mu_x, \Sigma_x)$ where $\mu_x \in \mathbb{R}^d, \Sigma_x \in \mathbb{R}^{d\times d}$. Let our prior $p(z)$ be the $N(0, I_d)$ density. Note that we could have $d \gg n$. We have $q(z) \propto e^{g(z)}$ where*

$$g(z) = \gamma'\phi(z)$$

where $K = 2d$ and

$$\gamma = \left(\Sigma^{-1}, \Sigma^{-1}\mu\right)^T$$
$$\phi(z) = \left(-\frac{1}{2}zz', z\right)^T$$

So (for $d = 1$, to keep it simple):

$$\frac{\partial\phi(z)}{\partial z} = (-z, 1)^T$$
$$\frac{\partial^2\phi(z)}{\partial z^2} = (1, 0)^T$$

We have $A(z) \in \mathbb{R}^{2d\times 2d}$. Let e.g. z^2 indicate $(z_1^2, \ldots, z_d^2)$. We have

$$A(z) = \begin{pmatrix} diag(z^2) & diag(-z) \\ diag(-z) & diag(1) \end{pmatrix} \text{ and}$$
$$k(z) = (rep(-1, d), rep(0, d))^T$$

Each submatrix is in $\mathbb{R}^{d\times d}$. We parameterize γ via B; it is some multi-layer non-linear function with many parameters B. Our objective, written as a sum, is

$$\mathbb{E}_{p(z|x)}\left(\sum_{j\le d}\gamma_{1j}^2 z_j^2 - \sum_{j\le d}\gamma_{2j}\gamma_{1j}z_j + \sum_{j\le d}\gamma_{2j}^2 - \sum_{j\le d}\gamma_{1j}\right)$$

Note that as noted in **?** *and Lin* et al. *(2016), there are closed-form solutions of* μ, Σ. *To see it, note that if* $q(z)$ *is* $N(\mu, \Sigma)$,

$$\log q(z) \propto \frac{-(z-\mu)'\Sigma^{-1}(z-\mu)}{2}$$

So the score-matching objective is

$$\frac{1}{2}\|\nabla_z \log q(z)\|_2^2 + \Delta_z \log q(z) = \frac{1}{2}\|\Sigma^{-1}(z-\mu)\|_2^2 + tr(\Sigma^{-1})$$

We can directly obtain:

$$\hat{\mu} = \mathbb{E}_{p(z|x)} z$$
$$\hat{\Sigma}_i = \mathbb{E}_{p(z|x)}(z-\hat{\mu})(z-\hat{\mu})'$$

While this is estimable for density estimation, it requires the posterior for us. Why does this happen when we can still not require $p(x)$ *if we compute the exponential family parameter? Notice that e.g. in the simple* $d = 1$ *case, the quadratic we'd obtain for score-matching is:*

$$\gamma' \mathbb{E}_{p(z|x)} \begin{pmatrix} z^2 & z \\ z & 1 \end{pmatrix} \gamma - \gamma' \mathbb{E}_{p(z|x)} \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

Optimizing this allows us to remove the $p(x)$, *but notice that then, we optimize in* γ $\gamma'\bar{A}\gamma' - \gamma'\bar{k}$, *the usual thing, where*

$$\bar{k} = \begin{pmatrix} p(x) \\ 0 \end{pmatrix}$$

Now again we can approximate $p(x)$ *via importance sampling, but we do have to approximate it either way.*

Lemma 2.1 (Score-matching simplifies to something simple using integration by parts)**.** *Let* $\hat{q} = \operatorname{argmin}_q \mathcal{L}(q)$ *where* $\mathcal{L}(q)$ *is as above. Then*

$$\begin{aligned}\hat{q} &= \operatorname*{argmin}_q \int_z p(z|x) \sum_{i \le d} \frac{1}{2}\left(\frac{\partial \log q(z)}{\partial z_i}\right)^2 + \frac{\partial^2 \log q(z)}{\partial z_i^2}\, dz \\ &= \operatorname*{argmin}_q \int_z p(z|x)\left(\frac{1}{2}\|\nabla \log q\|_2^2 + \triangle \log q\right)\, dz\end{aligned}$$

Proof. I just write this for one dimension. For an extension to higher dimensions, see Hyvarinen (2005). I sometimes write $p'(z)$ instead of $\frac{\partial p(z)}{\partial z}$.

$$\int p(z)\left(\frac{\partial \log p(z)}{\partial z} - \frac{\partial \log q(z)}{\partial z}\right)^2 dz \stackrel{c}{=} \int p(z)\left(\frac{\partial \log q(z)}{\partial z}\right)^2 - 2\int p'(z)\frac{\partial \log q(z)}{\partial z}$$

And using integration by parts,

$$\int p'(z)\frac{\partial \log q(z)}{\partial z} \stackrel{c}{=} -\int p(z)\frac{\partial^2 q(z)}{\partial z^2}$$

□

The score-matching objective simplifies into a nice quadratic for exponential families. First, let

$$q(z) = \frac{e^{g(z)}}{\int e^{g(z)} dz}$$

Suppose $g : \mathbb{R}^d \to \mathbb{R}$, with

$$g(z) = \gamma' \phi(z)$$

where $\phi : \mathbb{R}^d \to \mathbb{R}^K$ and $\gamma \in \mathbb{R}^K$. Write $\phi(z) = (\phi_1(z), \ldots, \phi_K(z))'$. Estimating q means estimating γ. Define

$$\frac{\partial \phi(z)}{\partial z_i} = \left(\frac{\partial \phi_1(z)}{\partial z_i}, \ldots, \frac{\partial \phi_K(z)}{\partial z_i} \right)^T$$
$$\frac{\partial^2 \phi(z)}{\partial z_i^2} = \left(\frac{\partial^2 \phi_1(z)}{\partial z_i^2}, \ldots, \frac{\partial^2 \phi_K(z)}{\partial z_i^2} \right)^T$$

For a non-bounded density (the second line is for an exponential family), the objective is, as in Janofsky (2015) and Hyvarinen (2007):

$$\begin{aligned} h_\gamma(z) &= \sum_{i \leq d} \frac{1}{2} \left(\frac{\partial g(z)}{\partial z_i} \right)^2 + \frac{\partial^2 g(z)}{\partial z_i^2} \\ &= \sum_{i \leq d} \frac{1}{2} \left(\gamma' \frac{\partial \phi(z)}{\partial z_i} \right)^2 + \gamma' \frac{\partial^2 \phi(z)}{\partial z_i^2} \end{aligned}$$

Now here are the definitions of A, k. Note $A \in \mathbb{R}^{K \times K}, k \in \mathbb{R}^K$. For a non-bounded density:

$$k(z) = \sum_{i \leq d} \frac{\partial^2 \phi(z)}{\partial z_i^2}$$
$$A(z) = \sum_{i \leq d} \frac{\partial \phi(z)}{\partial z_i} \frac{\partial \phi(z)}{\partial z_i}'$$

3 Bounded density

And for a bounded density:

$$k_1(z) = \sum_{i \leq d} 2(2z_i - 1) z_i (1 - z_i) \frac{\partial \phi(z)}{\partial z_i}$$
$$k_2(z) = \sum_{i \leq d} z_i^2 (1 - z_i)^2 \frac{\partial^2 \phi(z)}{\partial z_i^2}$$
$$k(z) = k_1(z) - k_2(z)$$
$$A(z) = \sum_{i \leq d} \frac{\partial \phi(z)}{\partial z_i} \frac{\partial \phi(z)}{\partial z_i}' z_i^2 (1 - z_i)^2$$

For a bounded density (the second line is for an exponential family), the objective is:

$$h_\gamma(z) = \sum_{i \le d} \frac{1}{2}\left(\frac{\partial g(z)}{\partial z_i} z_i(1-z_i)\right)^2 - 2(2z_i - 1)z_i(1-z_i)\frac{\partial g(z)}{\partial z_i} + z_i^2(1-z_i)^2\frac{\partial^2 g(z)}{\partial z_i^2}$$
$$= \sum_{i \le d} \left(\gamma'\frac{\partial \phi(z)}{\partial z_i} z_i(1-z_i)\right)^2 - 2(2z_i - 1)z_i(1-z_i)\gamma'\frac{\partial \phi(z)}{\partial z_i} + z_i^2(1-z_i)^2\gamma'\frac{\partial^2 \phi(z)}{\partial z_i^2}$$

In one dimension, and for $z \in [0,1]$, this simplifies to the following. Let $\phi : [0,1] \to \mathbb{R}^K$.

$$h_\gamma(z) = \frac{1}{2}\left(\frac{\partial g(z)}{\partial z} z(1-z)\right)^2 - 2(2z-1)z(1-z)\frac{\partial g(z)}{\partial z} + z(1-z)\frac{\partial^2 g(z)}{\partial z^2}$$
$$= \frac{1}{2}z^2(1-z)^2\gamma' A(z)\gamma - 2(2z-1)z(1-z)\gamma' k_1(z) + z^2(1-z)^2\gamma' k_2(z)$$

where

$$k_1(z) = \left(\frac{\partial \phi_1(z)}{\partial z}, \ldots, \frac{\partial \phi_K(z)}{\partial z}\right)'$$
$$k_2(z) = \left(\frac{\partial^2 \phi_1(z)}{\partial z^2}, \ldots, \frac{\partial^2 \phi_K(z)}{\partial z^2}\right)'$$
$$k(z) = k_1(z) - k_2(z)$$
$$A(z) = k_1(z)k_1(z)'$$

3.1 Score-matching with $\mathbb{E}_q$

The following lemma shows that the score-matching objective (1) decomposes into two components: a component matching q to p, and a component restricting q to be reasonably smooth. Thus the score-matching divergence decomposes in a similar manner as the KL divergence.

Lemma 3.1. *Suppose* $q(\theta) = \exp(f(\theta) - \Psi)$ *where* Ψ *is the normalizing constant. Let* $\mathcal{L}(q) = \mathbb{E}_q \|\nabla_\theta \log p - \nabla_\theta \log q\|_2^2$. *Then*

$$\mathcal{L}(q) =^c \frac{1}{2}\mathbb{E}_q \|\nabla_\theta \log p(\theta, x)\|_2^2 + \mathbb{E}_q \Delta \log p(\theta, x) + \mathbb{E}_q \|\nabla_\theta f(\theta)\|_2^2$$

Proof. [Only in one dimension for now.]

$$\mathcal{L}(q) = \int q(\theta)\left(\frac{\partial}{\partial \theta}\log p\right)^2 - 2\int q(\theta)\frac{q'(\theta)}{q(\theta)}\frac{p'(\theta)}{p(\theta)} + \int q(\theta)\left(\frac{\partial}{\partial \theta}\log q(\theta)\right)^2$$

The third term is $\mathbb{E}_q f'(\theta)^2$. And for the middle term, using integration by parts,

$$\int q(\theta)\frac{q'(\theta)}{q(\theta)}\frac{p'(\theta)}{p(\theta)} = \int q'(\theta)\frac{\partial}{\partial \theta}\log p(\theta) =^c -\int q(\theta)\frac{\partial^2}{\partial \theta^2}\log p(\theta)$$

□

The following lemma shows that when we assume $q(\theta)$ is a Gaussian density, the score-matching objective as in (1) reduces to the Laplace method as in Wang & Blei (2013). In retrospect, this is obvious by Lemma 3.3; since we assume a Gaussian form for the model, as long as we are optimizing over $\mathcal{Q}$ that includes Gaussian densities, we should obtain "the right thing." But this is nice as a sanity check.

Lemma 3.2. *Let $\theta \in \mathbb{R}^d$. Let $f(\theta) := \log p(\theta)$ and suppose for any θ_0,*

$$f(\theta) \approx f(\theta_0) + \nabla f(\theta_0)^T(\theta - \theta_0) + \frac{1}{2}(\theta - \theta_0)^T \nabla^2 f(\theta_0)(\theta - \theta_0)$$

. Let $\theta^* \in \operatorname{argmax} f(\theta)$. Let $\mathcal{Q}$ be $N(\lambda, V)$ densisties. Let $\hat{q} = \operatorname{argmin}_{q \in \mathcal{Q}} \mathbb{E}_q \|\nabla_\theta \log p - \nabla_\theta \log q\|_2^2$. Then $\hat{q}$ is the $N(\hat{\lambda}, \hat{V})$ density with

$$\hat{\lambda} = \theta^*$$
$$\hat{V} = -\nabla^2 f(\theta^*)$$

Proof. Let θ^* be a mode of f, i.e., $\nabla_\theta f(\theta^*) = 0$. Then

$$f(\theta) \approx f(\theta_0) + \frac{1}{2}(\theta - \theta^*)^T \nabla^2 f(\theta^*)(\theta - \theta^*)$$

So

$$\nabla_\theta f(\theta) \approx \nabla^2 f(\theta^*)(\theta - \theta^*)$$

Let q be the $N(\lambda, V)$ density where $\lambda \in \mathbb{R}^d$ and $V \in \mathbb{R}^{d \times d}$; λ and V are the parameters to be optimized. Now $\nabla_\theta \log q(\theta) = -V^{-1}(\theta - \lambda)$. Now $\theta = V^{1/2}\theta_0 + \lambda$, where $\theta_0 \sim N(0, I_p)$. So

$$\begin{aligned}
\mathcal{L}(q) &= \mathbb{E}_q \|\nabla_\theta \log q(\theta) - \nabla_\theta \log p(\theta)\|_2^2 \\
&= \mathbb{E}_{\theta_0 \sim N(0,I_p)} \|-V^{-1/2}\theta_0 - \nabla^2 f(\theta^*)(V^{1/2}\theta_0 + \lambda - \theta^*)\|_2^2 \\
&= \mathbb{E}_{\theta_0} \|\left(-V^{-1/2} - \nabla^2 f(\theta^*)V^{1/2}\right)\theta_0 - \nabla^2 f(\theta^*)(\lambda - \theta^*)\|_2^2 \\
&= \mathbb{E}_{\theta_0} \|\left(-V^{-1/2} - \nabla^2 f(\theta^*)V^{1/2}\right)\theta_0\|_2^2 + \|\nabla^2 f(\theta^*)(\lambda - \theta^*)\|_2^2 \\
&= \|V^{-1/2} + \nabla^2 f(\theta^*)V^{1/2}\|_F^2 + \|\nabla^2 f(\theta^*)(\lambda - \theta^*)\|_2^2
\end{aligned}$$

Optimizing in λ yields:

$$\hat{\lambda} = \theta^*$$

Let $M = \nabla^2 f(\theta^*)$. Let's do one dimension to check:

$$v^{-1/2} + mv^{1/2} = \frac{1 + mv}{\sqrt{v}}$$

And

$$\left(\frac{1 + mv}{\sqrt{v}}\right)^2 = \frac{1 + 2mv + m^2v^2}{v} = \frac{1}{v} + 2m + m^2v$$

Now

$$\frac{\partial}{\partial v}\left(\frac{1}{v}+2m+m^2v\right)=-\frac{1}{v^2}+m^2$$

Setting this equal to zero yields:

$$V^2=M^{-2}\Rightarrow V=\pm M^{-1}$$

which is exactly the Laplace method solution. □

Example 2 (Lemma 3.1 in the Gaussian case)**.** *Consider the objective (1). In the case where q is Gaussian, as in, Lemma 3.2, we can see that only optimizing the first two terms of the score-matching objective would lead to a terribly sharp q (with zero variance). The final term (the penalty on the smoothness of q) forces more smoothness. Let the setting be as in Lemma 3.2: q is the $N(\lambda, V)$ and the Taylor expansion for $\log p(x,\theta)$ holds. Now,*

$$\begin{aligned}\frac{1}{2}\mathbb{E}_q\|\nabla_\theta \log p(\theta,x)\|_2^2+\mathbb{E}_q\Delta\log p(\theta,x) &= \frac{1}{2}\mathbb{E}_q(\theta-\theta^*)^T\left(\nabla^2 f(\theta^*)\right)^2(\theta-\theta^*)-\mathbb{E}_q\nabla^2 f(\theta^*)\\ &=^c \frac{1}{2}tr\left(V^T\left(\nabla^2 f(\theta^*)\right)^2 V\right)+(\lambda-\theta^*)^T\left(\nabla^2 f(\theta^*)\right)^2(\lambda-\theta^*)\end{aligned}$$

which would be optimized by setting $\hat{\lambda}=\theta^*$ and $\hat{V}=0$. But we have the penalty:

$$\mathbb{E}_q\|\nabla\log q\|_2^2= tr(V^{-1})$$

Due to this penalty, we end up with a reasonable q.

3.2 Relationship between score-matching, Stein discrepancy, and OVI

Let $\psi_p(z)=\frac{\partial\log p(z)}{\partial z}=p'(z)/p(z)$ and similarly for $\psi_q(z)$. Let $\triangle p=\frac{\partial^2\log p(z)}{\partial z^2}$. Let $O^{p,q}$ be some operator depending on p,q. I will use the following form for exponential families in z:

$$p(z)=\exp\left(\gamma\phi(z)-\Psi(\gamma)\right) \tag{4}$$

The authors of Ranganath *et al.* (2016) propose to optimimze following variational objective in q:

$$\sup_{f\in\mathcal{F}}|\mathbb{E}_{q(z)}(O^p_{LS}f)(z)| \tag{5}$$

where

$$O^p_{LS}f=\psi_p f+\nabla f \tag{6}$$

Just as in the proof of Lemma 3.4, the objective (5) is

$$\sup_{f\in\mathcal{F}}|\mathbb{E}_q O^p_{LS}f| = \sup_{f\in\mathcal{F}}\mathbb{E}_{q(z)}(\psi_q(z)-\psi_p(z))f(z) \tag{7}$$

$$= \mathbb{E}_{q(z)}(\psi_q(z)-\psi_p(z))^2 \tag{8}$$

i.e., their objective is the score-matching objective with expectation over q. But if we use a specific class $\mathcal{F}$ as in the following lemma, this becomes the usual score-matching objective with expectation over p.

The following lemma shows that (5) is valid in that it is zero when $q = p$. Of course, when we express the objective as in Lemma 3.4, this becomes even more obvious.

Lemma 3.3. *Assume q is zero at the boundaries of the sample space. Then*

$$\mathbb{E}_{q(z)}\left(\psi_q f(z) + f'(z)\right) = 0$$

Proof. First, by the product rule,

$$\mathbb{E}_{p(z)}\left(\psi_q(z) f(z) + f'(z)\right) = \int_z \frac{p(z)}{q(z)} \frac{\partial\left(q(z) f(z)\right)}{\partial z}\, dz$$

Now

$$\int \frac{q(z)}{q(z)} \frac{\partial(q(z) f(z))}{\partial z}\, dz = \int \frac{\partial}{\partial z}\left(q(z) f(z)\right)\, dz = q(\infty) f(\infty) - q(-\infty) f(-\infty) = 0$$

under some assumptions on q, f. □

Lemma 3.4 (Equivalence of score-matching and Stein discrepancy under specific $\mathcal{G}$)**.** *Let $\mathcal{G}$ be a class of functions of the form $g(z) = \psi_q(z) f(z) + f'(z)$, for f in some class of functions $\mathcal{F}$. Then*

$$\sup_{g \in \mathcal{G}} |\mathbb{E}_{p(z)} g(z)| = \sup_{f \in \mathcal{F}} |\mathbb{E}_{p(z)}\left(\psi_q(z) f(z) + f'(z)\right)| = \mathbb{E}_p\left(\psi_p - \psi_q\right)^2$$

Proof. By Lemma 3.3, $\mathbb{E}_{p(z)} \psi_p(z) f(z) + f'(z) = 0$. So

$$\begin{aligned}
\sup_{g \in \mathcal{G}} |\mathbb{E}_{p(z)} g(z)| &= \sup_{f \in \mathcal{F}} |\mathbb{E}_{p(z)}\left(\psi_q(z) f(z) + f'(z) - \psi_p(z) f(z) - f'(z)\right)| \\
&= \sup_{f \in \mathcal{F}} |\mathbb{E}_p(\psi_q - \psi_p) f|
\end{aligned}$$

And this is maximized when $f(z) = \psi_q(z) - \psi_p(z)$. □

Lemma 3.5. *Let $\mathcal{F}$ be a class of functions of the form (assuming the densities are positive everywhere):*

$$f(z) = \frac{p(z)}{q(z)} h(z) \tag{9}$$

for $h(z)$ in some class of functions $\mathcal{H}$. Then

$$\sup_{f \in \mathcal{F}} |\mathbb{E}_q O^p_{LS} f| = \sup_{h \in \mathcal{H}} |\mathbb{E}_p O^q_{LS} h|$$

Proof. This is clear if we use the representation in (7) and apply Lemma 3.4. That is,

$$\sup_{f\in\mathcal{F}}|\mathbb{E}_q O^p_{LS} f| = \sup_{f\in\mathcal{F}} \mathbb{E}_q(\psi_q - \psi_p) f = \sup_{h\in\mathcal{H}} \mathbb{E}_q \frac{p}{q}(\psi_q - \psi_p) h = \sup_{h\in\mathcal{H}} \mathbb{E}_p(\psi_q - \psi_p) h = \sup_{h\in\mathcal{H}} |\mathbb{E}_p O^q_{LS} h|$$

Alternatively (this is a sanity check):

$$\begin{aligned}
\sup_{f\in\mathcal{F}}|\mathbb{E}_q O^p_{LS} f| &= \sup_{f\in\mathcal{F}}|\mathbb{E}_q \left(\psi_p(z) f(z) + f'(z)\right)| \\
&= \sup_{h\in\mathcal{H}} \left| \int q(z) \frac{p'(z)}{p(z)} \frac{p(z)}{q(z)} h(z) + \int q(z) \left(\frac{q(z)p'(z) - p(z)q'(z)}{q(z)^2} h(z) + \frac{p(z)}{q(z)} h'(z) \right) \right| \\
&= \sup_{h\in\mathcal{H}} \left| \int p'(z)h(z) + \int p'(z)h(z) - \int \frac{q'(z)}{q(z)} p(z)h(z) + \int p(z)h'(z) \right| \\
&= \sup_{h\in\mathcal{H}} \left| \int 2\left(p'(z)h(z) + p(z)h'(z)\right) - \int p(z)h'(z) - \int \frac{q'(z)}{q(z)} p(z)h(z) \right| \\
&= \sup_{h\in\mathcal{H}} |\mathbb{E}_p \left(\psi_q(z)h(z) + h'(z)\right)|
\end{aligned}$$

since $p'(z)h(z) + p'(z)h(z) = \frac{\partial(p(z)h(z))}{\partial z}$, which has integral zero by Lemma 3.3. □

References

Hyvarinen, Aapo. 2005. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, **6**, 695–709.

Hyvarinen, Aapo. 2007. Some extensions of score matching. *Computational Statistics and Data Analysis*, 2499–2512.

Janofsky, Eric. 2015. *Exponential series approaches for nonparametric graphical models.* Ph.D. thesis, University of Chicago.

Lin, Lina, Drton, Mathias, & Shojaie, Ali. 2016. Estimation of high-dimensional graphical models using regularized score matching. *Electronic Journal of Statistics*, **10**(1), 806–854.

Ranganath, Rajesh, Tran, Dustin, Altosaar, Jaan, & Blei, David. 2016. Operator Variational Inference. *Pages 496–504 of:* Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., & Garnett, R. (eds), *Advances in Neural Information Processing Systems 29.* Curran Associates, Inc.

Wang, Chong, & Blei, David M. 2013. Variational inference in nonconjugate models. *Journal of Machine Learning Research*, **14**, 1005–1031.