NET Institute*

www.NETinst.org

Working Paper #13-16

September 2013

Approximate Variational Inference for a Model of Social Interactions

Angelo Mele
John Hopkins University

* The Networks, Electronic Commerce, and Telecommunications (“NET”) Institute, <http://www.NETinst.org>, is a non-profit institution devoted to research on network industries, electronic commerce, telecommunications, the Internet, “virtual networks” comprised of computers that share the same technical standard or operating system, and on network issues in general.

APPROXIMATE VARIATIONAL INFERENCE FOR A MODEL OF SOCIAL INTERACTIONS

ANGELO MELE
JOHNS HOPKINS UNIVERSITY

ABSTRACT. This paper proposes approximate variational inference methods for estimation of a strategic model of social interactions. Players interact in an exogenous network and sequentially choose a binary action. The utility of an action is a function of the choices of neighbors in the network. I prove that the interaction process can be represented as a potential game and it converges to a unique stationary equilibrium distribution. However, exact inference for this model is infeasible because of a computationally intractable likelihood, which cannot be evaluated even when there are few players. To overcome this problem, I propose variational approximations for the likelihood that allow approximate inference. This technique can be applied to any discrete exponential family, and therefore it is a general tool for inference in models with a large number of players. The methodology is illustrated with several simulated datasets and compared with MCMC methods.

Keywords: *Bayesian Methods, Microeconometrics*

Date: October 1, 2013. Work in progress - Please do not cite.

I am grateful to Roger Koenker, who introduced me to variational methods for inference. I gratefully acknowledge financial support from the NET Institute, www.NETinst.org.

1. INTRODUCTION

Social networks shape our behavior and are important determinants of socioeconomic outcomes. The economics literature has framed the study of social interactions as strategic interactions among rational players, where the interactions take place in a network of relationships determining who influences whom. With the emergence of online social networks, researchers have access to enormous amounts of data to study and estimate models of social interactions.

However, while the abundance of data makes inference more reliable, the literature focuses on small datasets or networks with few hundreds of players. One of the reasons is that most models are estimated using simulation methods, which become infeasible when the number of players grows large. The aim of the present work is to propose a model of social interactions and a scalable estimation technique that allows approximate inference in large economies, when simulation methods become too burdensome or infeasible.

In our model, social interactions occur sequentially: players interact on an exogenous network and sequentially choose a binary action. An agent's payoff is a function of the choices of individuals that are connected to her through the network and a stochastic shock to his preferences. For example, we can think of a teenager that decides to smoke or not: her decision depends on her preferences for smoking but also the peer pressure of her friends. In this model, the incentives of the players can be characterized using a potential function, which simplifies the computation of equilibria. It is shown that the sequential interaction process converges to a unique stationary equilibrium, which characterizes the probability of observing a specific action profile in the population in the long run.

The likelihood is proportional to an intractable normalizing constant, that cannot be computed exactly. One possible solution is to employ MCMC methods to sample from this intractable likelihood. For example, Mele (2011*b*) proposed a Bayesian estimation strategy for a model of network formation with similar structure, using a Markov Chain Monte Carlo sampler that avoids the computation of the constant and allows inference in high-dimensional models. Such estimation employs an approximate version of the exchange algorithm (Murray et al. (2006)) and consists of a double Metropolis-Hastings step: first, a parameter is proposed; then the sampler simulates a realization from the stationary equilibrium of the model using a metropolis-hastings algorithm; the proposed parameter is accepted if the simulated and observed configuration of actions are similar.

However, the second step of the exchange algorithm may require an infeasible number of simulations to generate an *exact* sample from the stationary distribution of the model. While there are several improvements of the original algorithm that relax the requirement of exact sampling (Liang (2010); Liang et al. (2010); Mele (2011*b*)), the computational cost of the simulation is very high when the number of players is large and for models with high-dimensional parameter spaces. This computational bottleneck could imply that for some models and large datasets, the algorithm is practically infeasible.¹ In addition, the MCMC methods require additional effort to diagnose convergence, which may be extremely burdensome for high-dimensional models. The researcher has to monitor convergence for each parameter, storing the output of the sampler. This may require a huge amount of

¹See Diaconis and Chatterjee (2011); Bhamidi et al. (forthcoming) for several examples.

computational resources. Finally, even when the MCMC convergence is not an issue, it may mix slowly, therefore requiring very long runs to provide reliable approximations of the posterior. Including preference heterogeneity or unobserved heterogeneity would further increase the computational burden.

To overcome these computational problems, I propose an alternative algorithm for estimation of such models. I use a variational approximation for the log-likelihood of the model.² Variational approximations are a set of deterministic approximation techniques, frequently used in the machine learning literature. These methods are gaining popularity in statistics, political science, marketing, and network science because they allow fast and reliable inference in models where the usual MCMC is infeasible.³

The algorithm consists of approximating the log-likelihood using a more tractable distribution. The latter distribution is then optimized to obtain the best approximation. This method allows Bayesian and classical inference. It can be shown that the procedure provides an approximation that minimizes the Kullback-Leibler distance between the true likelihood and the approximated likelihood. Given the approximating distribution chosen, the error of approximation is minimized, but it is fixed. This is an important difference between variational approximations and stochastic approximations techniques based on Monte Carlo, where the error of approximation can be decreased by increasing the number of simulations. Another difference between the variational approximation and MCMC is that the former converges quite fast to the final approximation, while the latter may take an unfeasible long time to produce acceptable results. In addition, the assessment of convergence for MCMC schemes is usually very complicated, requiring multiple chains to be run in parallel, especially for high-dimensional models. The assessment of convergence for variational approximation in contrast is very simple and can be done by monitoring a single scalar. The main limitation of variational mean-field approximations is that they underestimate the correlation among variables. As a consequence the approximated likelihood shows less variability and correlation among variables.

The preliminary results of this paper suggest that the variational approach can be quite effective in estimation of social interaction models, providing a reliable alternative estimation method when the MCMC is infeasible or too burdensome. Alternative deterministic approximation techniques are exploited in the ERGM literature by Chandrasekhar and Jackson (2012) and Amir et al. (2012). Their focus is on consistency of the estimators. Moreover, the large-deviations argument of Diaconis and Chatterjee (2011) essentially provides an asymptotic mean-field variational approximation to the ERGM normalizing constant.

The methodology presented in this paper is applicable to any discrete exponential family and has potential applications in many different contexts. For example, it can be used to estimate autologistic models in spatial econometrics. Or Exponential random graph models in social network analysis. More generally the variational approximation methods can be

²See Wainwright and Jordan (2008) and Jaakkola (2000) for an introduction to variational methods for approximation and inference.

³Ormerod and Wand (2010) provide a concise introduction and several examples familiar to statisticians. Grimmer (2011) provides an application to estimation of topic models in political science. Braun and McAuliffe (2010) consider variational approximations for discrete choice models. Airolidi et al. (2008) and Schweinberger and Handcock (2013) use variational methods to estimate network formation models.

used when the MCMC methods are too expensive in terms of computational time required to undertake the estimation, or when MCMC becomes infeasible due to the dimensionality of the model. Variational methods provide fast and reliable estimation methods when the MCMC is unreliable or simply too expensive to work with.

This paper provides preliminary evidence that the methodology could be useful for estimation of many highly complex models in econometrics.

2. A SIMPLE MODEL OF SOCIAL INTERACTIONS

We model social interactions in a setting where players' choose binary actions. Social interactions occur on an exogenous network of social relationships, described by a symmetric adjacency matrix g . Each element $g_{ij} = g_{ji}$ indicates the existence of a social tie between individual i and j : if $g_{ij} = 1$, individuals i and j are friends; if $g_{ij} = 0$, they do not interact or influence each other. Individual i chooses a binary action y_i , to maximize her utility function. To be concrete, let's consider the example of youth smoking: here y_i represents the action of smoking ($y_i = 1$). The utility function is

$$(2.1) \quad u_i(y_i, y_{-i}, x, g) = \alpha(x_i, x_{\mathcal{N}_i})y_i + \sum_{j=1}^n \gamma(x_i, x_j)g_{ij}y_jy_i + \varepsilon(y_i)$$

where $\mathcal{N}_i = \{j \text{ s.t. } g_{ij} = 1\}$ is the set of i 's neighbors. The first component of the utility function $\alpha(x_i, x_{\mathcal{N}_i})y_i$ represent the consumption value of smoking, and it depends on the characteristics x_i of individual i , and the characteristics of his peers $x_{\mathcal{N}_i}$. This component captures both individual and contextual effects in Manski (1993)'s terminology. The second part $\sum_{j=1}^n \gamma(x_i, x_j)g_{ij}y_jy_i$ contains the endogenous social interaction effects, mediated through the network g . Let's consider two players i and j that are connected in the network, i.e. $g_{ij} = 1$. When player i chooses her binary action y_i , she receives an additional utility $\gamma(x_i, x_j)g_{ij}y_jy_i$ that depends on her characteristics x_i , but also the characteristics x_j and the action y_j of her peer. In addition, individuals receive an idiosyncratic preference shock $\varepsilon(y_i)$ that shifts the utility of a specific choice, as it is standard in the discrete choice literature. In the rest of the paper, I will use shorthand notation

$$(2.2) \quad \alpha_i := \alpha(x_i, x_{\mathcal{N}_i})$$

$$(2.3) \quad \gamma_{ij} := \gamma(x_i, x_j)$$

The smoking interaction process evolves according to a stochastic best-response dynamics, as in Nakajima (2007); Mele (2011b); Blume (1993) and Badev (2013). At period t an individual is randomly selected from the population with probability $\rho(x_i, g)$

$$(2.4) \quad i \sim \rho(x_i, g)$$

Then the preference shocks are realized $\varepsilon_t = (\varepsilon_{1t}, \varepsilon_{0t})$ and individual i updates his smoking decision y_i^t by maximizing his current utility, given the current smoking profile of his neighbors. Player i chooses $y_i^t = 1$ if and only if

$$(2.5) \quad u_i(y_i^t = 1, y_{-i}^{t-1}, x, g) + \varepsilon_{1t} \geq u_i(y_i^t = 0, y_{-i}^{t-1}, x, g) + \varepsilon_{0t}$$

Players are myopic: when they decide to update their choices they do not take into account the future costs and benefits of their action.

An additional assumption on the interaction effects γ_{ij} allows a very simple characterization of the equilibria.

ASSUMPTION 2.1. *The interaction effects are symmetric*

$$(2.6) \quad \gamma_{ij} = \gamma_{ji}$$

This assumption guarantees the existence of a potential function⁴ that characterizes the deterministic incentives of all players in each period.

PROPOSITION 2.1. *There exists a function $Q(y, x, g)$*

$$(2.7) \quad Q(y, x, g) = \sum_{i=1}^n \alpha_i y_i + \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \gamma_{ij} g_{ij} y_j y_i$$

that characterizes the deterministic incentives of the players, i.e. for any i

$$(2.8) \quad Q(y_i = 1, y_{-i}, x, g) - Q(y_i = 0, y_{-i}, x, g) = u_i(y_i = 1, y_{-i}, x, g) - u_i(y_i = 0, y_{-i}, x, g)$$

In a model without preference shocks, the Nash equilibria correspond to the local maxima of the potential function. In addition the Nash equilibria are absorbing states of the system. The role of the preference shocks is to allow the system to escape the absorbing states. As a consequence the Markov Chain describing the evolution of social interactions is ergodic.

I study the long run evolution of social interactions described by this model. I make two assumptions on the stochastic parts of the model.

ASSUMPTION 2.2. *All players have a positive probability of being selected for the update of their strategy*

$$(2.9) \quad \rho(x_i, g) > 0 \text{ for all } i$$

ASSUMPTION 2.3. *The preference shocks have Extreme value distribution Type 2 and are i.i.d. across agents and over time.*

These assumptions on the stochastic processes driving the evolution of the smoking profile are necessary to obtain a unique stationary equilibrium in closed form.

THEOREM 2.1. *The smoking profile sequence generated by the model converges to a unique stationary distribution π which has the form*

$$(2.10) \quad \pi(y, x, g) = \frac{\exp [Q(y, x, g)]}{\sum_{\omega \in \mathcal{Y}} \exp [Q(\omega, x, g)]}$$

The proof is similar to Mele (2011b) and therefore is omitted for brevity. The stationary distribution is the likelihood of observing a smoking profile y in the long run. Under the assumption that the smoking profile observed in the data is a draw from this stationary distribution, we can estimate the utility functions using $\pi(y, x, g)$ as likelihood. However, the normalizing constant

$$(2.11) \quad c(\mathcal{Y}, x, g) \equiv \sum_{\omega \in \mathcal{Y}} \exp [Q(\omega, x, g)]$$

⁴See Monderer and Shapley (1996) for a detailed analysis of potential functions and potential games. This characterization is exploited in Mele (2011b) to estimate a model of strategic network formation.

sums over all possible smoking profiles $\omega \in \mathcal{Y}$. In a model with 100 players, the number of possible smoking profiles $|\mathcal{Y}|$ is $2^{100} \simeq 1.26 \cdot 10^{30}$ and therefore the evaluation of the likelihood is infeasible. This computational problem is a severe challenge for the estimation of these models.

3. VARIATIONAL INFERENCE METHODS

Exact inference is impossible, because the evaluation of the likelihood is infeasible even for models with few players. Therefore I propose a variational algorithm that approximates the normalization constant $c(\mathcal{Y}, x, g)$, using a mean-field approach. The variational approach consists of transforming the problem of computing a sum (or integral) into an optimization problem, that can be solved iteratively.

To introduce the variational approach, I make an additional assumption on the functional form of $\alpha(x_i)$ and $\gamma(x_i, x_j)$.

ASSUMPTION 3.1. *The functions $\alpha(x_i, \theta_\alpha)$ and $\gamma(x_i, x_j, \theta_\gamma)$ are respectively linear in θ_α and θ_γ*

As a consequence the potential function is a linear function of parameters θ and the stationary distribution belongs to the discrete exponential family with normalizing constant (Lehman (1983); Chandrasekhar and Jackson (2012); Mele (2011a))

$$(3.1) \quad \pi(y, x, g; \theta) = \frac{\exp [\theta^T t(y, x, g)]}{\sum_{\omega \in \mathcal{Y}} \exp [\theta^T t(\omega, x, g)]}$$

with log-likelihood

$$(3.2) \quad \ell(y, x, g; \theta) = \ln \pi(y, x, g; \theta) = \theta^T t(y, x, g) - \kappa(\mathcal{Y}, x, g; \theta)$$

where $\kappa(\mathcal{Y}, x, g; \theta)$ is called the *log-partition function* and it's defined as

$$(3.3) \quad \kappa(\mathcal{Y}, x, g; \theta) = \ln c(\mathcal{Y}, x, g, \theta) = \ln \sum_{\omega \in \mathcal{Y}} \exp [\theta^T t(\omega, x, g)]$$

The variational approach consists of restating the problem of computing $\kappa(\mathcal{Y}, x, g; \theta)$ as a maximization problem. In practice we consider an approximating distribution for the likelihood, and we find the best approximating distribution by minimizing the Kullback-Leibler distance between the likelihood and the approximation.

A simple application of the Jensen's inequality provides a lower bound to the log constant (3.3). Let $q(y)$ be an arbitrary distribution of smoking profiles.

$$(3.4) \quad \kappa(\mathcal{Y}, x, g; \theta) = \ln \sum_{\omega \in \mathcal{Y}} \exp [\theta^T t(\omega, x, g)]$$

$$(3.5) \quad = \ln \sum_{\omega \in \mathcal{Y}} q(\omega) \frac{\exp [\theta^T t(\omega, x, g)]}{q(\omega)}$$

$$(3.6) \quad \geq \sum_{\omega \in \mathcal{Y}} q(\omega) \ln \left\{ \frac{\exp [\theta^T t(\omega, x, g)]}{q(\omega)} \right\}$$

$$(3.7) \quad = \sum_{\omega \in \mathcal{Y}} q(\omega) \theta^T t(\omega, x, g) - \sum_{\omega \in \mathcal{Y}} q(\omega) \ln q(\omega)$$

$$(3.8) \quad = \theta^T \mathbb{E}_q [t(\omega, x, g)] + \mathcal{H}(q) \equiv \mathcal{L}(q)$$

where $\mathcal{H}(q) = -\sum_{\omega \in \mathcal{Y}} q(\omega) \ln q(\omega)$ is the entropy of distribution q . We can find the best lower bound by maximizing (3.8) with respect to the distribution q .

$$(3.9) \quad \kappa^*(\mathcal{Y}, x, g; \theta) = \max_{q(\theta) \in \mathcal{Q}} \mathcal{L}(q) = \max_{q(\theta) \in \mathcal{Q}} \left\{ \theta^T \mathbb{E}_q [t(\omega, x, g)] + \mathcal{H}(q) \right\}$$

This unconstrained optimization problem is usually not simple and does not have a closed form solution for most problems. However, the relationship in (3.8) above is true for any distribution q , therefore we can restrict our search to those distribution q that guarantee explicit solution, solving a constrained optimization problem.⁵

⁵Using the properties of the exponential families we can rewrite the log-likelihood in *mean parameterization form* (Wainwright and Jordan (2008)). For distributions in the exponential family, the vector of moments $\mu(\theta)$ is given by

$$\mu(\theta) = \mathbb{E}_\pi [t(\omega, x, g)] = \nabla \kappa(\mathcal{Y}, x, g; \theta)$$

We define the set of all possible mean parameters as

$$(3.10) \quad \mathcal{M} \equiv \{\mu : \exists p \text{ s.t. } \mathbb{E}_p [t(\omega, x, g)] = \mu\}$$

which is called the *marginal polytope* in the literature. Let $\mathcal{M}^0$ denote the interior of $\mathcal{M}$. We can write the conjugate function of the log-partition as

$$\kappa^*(\mu(\theta)) = \sup_{\theta \in \Theta} \{ \theta^T \mu(\theta) - \kappa(\theta) \}$$

where we have suppressed the dependence of the log-partition function from $\mathcal{Y}, x, g$ for notational convenience. It can be shown that the function $\kappa^*(\mu(\theta))$ is the negative entropy of the model with parameter $\theta(\mu) = (\nabla \kappa)^{-1}(\mu)$ (Wainwright and Jordan (2008), thm 3.4, p.67). By convex duality we have the log partition function as

$$(3.11) \quad \kappa(\theta) = \sup_{\mu \in \mathcal{M}^0} \{ \theta^T \mu - \kappa^*(\mu) \}$$

The equation (3.11) is a variational representation of the log-partition function $\kappa(\theta)$.

A variational approximation for the model is an approximation of the optimization in (3.11). Since the optimization problems is not tractable in most cases, we restrict the set $\mathcal{M}^0$ and provide an approximation of the negative entropy $\kappa^*(\mu)$. In practice the variational approximation problem can be stated as

$$(3.12) \quad \kappa_{approx}(\theta) = \sup_{\mu \in \mathcal{M}_{approx}^0} \{ \theta^T \mu - \kappa_{approx}^*(\mu) \}$$

Different restrictions on $\mathcal{M}^0$ and $\kappa_{approx}(\theta)$ lead to different variational approximations.

In addition we can show that the maximization problem can be interpreted as minimizing the Kullback-Leibler distance between the likelihood $\pi(y, x, g; \theta)$ and the approximating distribution $q(y)$. Indeed the Kullback-Leibler divergence between $q(y)$ and $\pi(y, x, g; \theta)$ is

$$\begin{aligned}
KL(q|\pi) &= \sum_{\omega \in \mathcal{Y}} q(\omega) \ln \left[\frac{q(\omega)}{\pi(\omega, x, g; \theta)} \right] \\
&= \sum_{\omega \in \mathcal{Y}} q(\omega) \ln [q(\omega)] - \sum_{\omega \in \mathcal{Y}} q(\omega) \ln [\pi(\omega, x, g; \theta)] \\
&= -\mathcal{H}(q) - \theta^T \mathbb{E}_q [t(\omega, x, g)] + \mathbb{E}_q [\kappa(\mathcal{Y}, x, g; \theta)] \\
&= -\mathcal{L}(q) + \text{constant}
\end{aligned}$$

Therefore maximizing $\mathcal{L}(q)$ is equivalent to minimizing $KL(q|\pi)$.

3.1. Mean-Field Methods. The mean-field variational approximation consists of choosing a distribution q that is completely factorized, i.e.

$$(3.13) \quad q(y) = \prod_{i=1}^n q_i(y_i) = \prod_{i=1}^n \mu_i$$

where we have defined $\mu_i = \mathbb{E}_q(y_i) = \Pr(y_i = 1)$. Straightforward algebra shows that

$$(3.14) \quad \mathcal{H}(q) = \sum_{i=1}^n \mathcal{H}_i(\mu_i) = - \sum_{i=1}^n [\mu_i \ln \mu_i + (1 - \mu_i) \ln(1 - \mu_i)]$$

and the expected value of the sufficient statistics under the completely factorized q is

$$\begin{aligned}
\theta^T \mathbb{E}_q [t(\omega, x, g)] &= \mathbb{E}_q \theta^T [t(\omega, x, g)] \\
&= \mathbb{E}_q \sum_i \alpha(x_i, x_{N_i}; \theta_\alpha) y_i + (1/2) \sum_i \sum_j \gamma(x_i, x_j; \theta_\gamma) g_{ij} y_i y_j \\
&= \sum_i \alpha(x_i, x_{N_i}; \theta_\alpha) \mathbb{E}_q [y_i] + (1/2) \sum_i \sum_j \gamma(x_i, x_j; \theta_\gamma) g_{ij} \mathbb{E}_q [y_i y_j] \\
&= \sum_i \alpha(x_i, x_{N_i}; \theta_\alpha) \mu_i + (1/2) \sum_i \sum_j \gamma(x_i, x_j; \theta_\gamma) g_{ij} \mu_i \mu_j
\end{aligned}$$

Therefore the lower bound for $\kappa(\mathcal{Y}, x, g; \theta)$ in the space of completely factorized $q(y)$ can be found by solving the following optimization problem

$$\begin{aligned}
(3.15) \quad \kappa(\mathcal{Y}, x, g; \theta) &\geq \kappa_{MF}(\mathcal{Y}, x, g; \mu(\theta)) \\
&= \max_{\{\mu_1, \dots, \mu_n\} \in [0, 1]^n} \left\{ \sum_{i=1}^n \alpha_i(\theta_\alpha) \mu_i + (1/2) \sum_{i=1}^n \sum_{j=1}^n \gamma_{ij}(\theta_\gamma) g_{ij} \mu_i \mu_j \right. \\
(3.16) \quad &\left. - \sum_{i=1}^n [\mu_i \ln \mu_i + (1 - \mu_i) \ln(1 - \mu_i)] \right\}
\end{aligned}$$

The problem is concave in each μ_i but it is not globally concave in general. The first order conditions with respect to each μ_i provide a system of equations

$$(3.17) \quad \mu_i = \frac{\exp \left[\alpha_i(\theta_\alpha) + \sum_{j=1}^n \gamma_{ij}(\theta_\gamma) g_{ij} \mu_j \right]}{1 + \exp \left[\alpha_i(\theta_\alpha) + \sum_{j=1}^n \gamma_{ij}(\theta_\gamma) g_{ij} \mu_j \right]}; \quad i = 1, \dots, n$$

It is standard to use a naive coordinate ascent method, where each μ_i is updated iteratively using (3.17) above. At iteration s the coordinate i is updated according to

$$(3.18) \quad \mu_i^{(s+1)} = \frac{\exp \left[\alpha_i(\theta_\alpha) + \sum_{j=1}^n \gamma_{ij}(\theta_\gamma) g_{ij} \mu_j^{(s)} \right]}{1 + \exp \left[\alpha_i(\theta_\alpha) + \sum_{j=1}^n \gamma_{ij}(\theta_\gamma) g_{ij} \mu_j^{(s)} \right]}; \quad i = 1, \dots, n$$

This method is guaranteed to converge to a local maximum of the optimization problem (3.15). The approximation of the normalizing constant allows approximate Maximum Likelihood estimation and approximate Bayesian inference.

There is an additional variational approximation that we can exploit. We can restrict $\mu_i = \mu$ for all i , which in turn implies (let Δ be the network density) a simpler mean-field variational optimization problem

$$\kappa_{MF}(\mu(\theta)) = \max_{\mu \in [0,1]} \left\{ \alpha n \mu + \frac{\gamma}{2} n \Delta \mu^2 - n [\mu \ln \mu + (1 - \mu) \ln(1 - \mu)] \right\}$$

and the mean-field updates are simplified to the following equation

$$(3.19) \quad \mu = \frac{\exp [\alpha + \gamma \Delta \mu]}{1 + \exp [\alpha + \gamma \Delta \mu]}$$

The optimization is now in one variable only and the mean-field equation has always at least one solution. This approximation works surprisingly well in our examples. In the ERGM literature, Diaconis and Chatterjee (2011) show that if the number of players grows large $n \rightarrow \infty$ the mean-field approximation becomes exact $\lim_{n \rightarrow \infty} \kappa_{MF}(\mu(\theta)) = \kappa(\theta)$, therefore this approximation could lead to the development of consistent estimators. We will compare both approximations in the numerical experiments.

3.2. Approximate Maximum Likelihood Estimation. The variational approximation allows evaluation of the approximated log-likelihood for estimation. Instead of maximizing the exact likelihood we maximize the approximated log-likelihood estimated with mean-field variational approximations. So the inference problem becomes

$$(3.20) \quad \theta_{mf} = \arg \max_{\theta \in \Theta} \ell_{MF}(q(\theta)) = \arg \max_{\theta \in \Theta} \{ \theta^T t(y, x, g) - \kappa_{MF}(\mathcal{Y}, x, g; \theta) \}$$

The mean-field approximation allows evaluation of the approximated log-likelihood on a grid of values for the parameters. The idea is similar to the MCMC-MLE of Geyer and Thompson (1992), which approximates the ratio of normalizing constants using an MCMC sample and then uses the approximated likelihood to perform inference. The MCMC-MLE procedure is known to work best when the MCMC simulation is started at parameter values

“close” to the maximum likelihood estimate and it is shown to converge almost surely. However, for models with a large number of players, the simulations are very expensive and may become infeasible or unreliable. For example, Diaconis and Chatterjee (2011) show that the variance of the MCMC-MLE goes to infinity for large samples. The variational approximation provides a computationally efficient alternative when the MCMC-MLE is infeasible.

3.3. Approximate Bayesian Estimation. Assuming a prior distribution $p(\theta)$, it is possible to write the posterior as

$$(3.21) \quad p(\theta|y, x, g) = \frac{\pi(y, x, g; \theta)p(\theta)}{\int_{\Theta} \pi(y, x, g; \eta)p(\eta)d\eta}$$

which is a doubly intractable constant distribution (Murray et al. (2006))

$$(3.22) \quad p(\theta|y, x, g) = \frac{\frac{\exp[Q(y, x, g; \theta)]}{c(\mathcal{Y}, x, g)}p(\theta)}{\int_{\Theta} \frac{\exp[Q(y, x, g; \eta)]}{c(\mathcal{Y}, x, g)}p(\eta)d\eta} = \frac{\frac{\exp[Q(y, x, g; \theta)]}{c(\mathcal{Y}, x, g)}p(\theta)}{p(y, x, g)}$$

For these models the usual Metropolis-Hastings scheme fails. The standard metropolis Hastings algorithm consists of the following steps: at iteration s , when current parameter is $\theta_s = \theta$

(1) Propose a new parameter θ' from the proposal distribution $q_{\theta}(\cdot|\theta)$

$$(3.23) \quad \theta' \sim q_{\theta}(\cdot|\theta)$$

(2) Update θ_{s+1} according to

$$(3.24) \quad \theta_{s+1} = \begin{cases} \theta' & \text{with prob. } \alpha(\theta, \theta') \\ \theta & \text{with prob. } 1 - \alpha(\theta, \theta') \end{cases},$$

where

$$\begin{aligned} \alpha(\theta, \theta') &= \min \left\{ 1, \frac{\pi(y, x, g; \theta') p(\theta') q_{\theta}(\theta'|\theta)}{\pi(y, x, g; \theta) p(\theta) q_{\theta}(\theta|\theta')} \right\} \\ &= \min \left\{ 1, \frac{\exp[Q(y, x, g; \theta')] c(\mathcal{Y}, x, g; \theta) p(\theta') q_{\theta}(\theta'|\theta)}{\exp[Q(y, x, g; \theta)] c(\mathcal{Y}, x, g; \theta') p(\theta) q_{\theta}(\theta|\theta')} \right\} \end{aligned}$$

However, the computation of $\alpha(\theta, \theta')$ is not feasible because the ratio $c(\mathcal{Y}, x, g; \theta)/c(\mathcal{Y}, x, g; \theta')$ cannot be evaluated.

The *approximate bayesian estimation* technique consists of substituting the exact log-likelihood with the approximated log-likelihood computed using the variational mean-field approximation. Therefore we consider the approximated posterior

$$(3.25) \quad \log p(\theta|y, x, g) \approx \log \tilde{p}(\theta|y, x, g) = \ell_{MF}(q(\theta)) + \log p(\theta) - \log p(y)$$

The unnormalized approximated posterior can be easily evaluated (and interpolated) on a grid of parameter values, to find the posterior mode.

An alternative approach consists of using an approximate MCMC scheme that I call *variational Metropolis-Hastings*. The variational Metropolis-Hastings consists of estimating the

ratio $\frac{c(\mathcal{Y}, x, g; \theta)}{c(\mathcal{Y}, x, g; \theta')}$ with the mean-field approximated ratio $\frac{\exp[\kappa_{MF}(\mathcal{Y}, x, g; \mu(\theta))]}{\exp[\kappa_{MF}(\mathcal{Y}, x, g; \mu(\theta'))]}$ and substitute it in the acceptance ratio $\alpha(\theta, \theta')$ to obtain an approximated version $\tilde{\alpha}(\theta, \theta')$

$$\tilde{\alpha}(\theta, \theta') = \min \left\{ 1, \frac{\exp[Q(y, x, g; \theta')]}{\exp[Q(y, x, g; \theta)]} \frac{\exp[\kappa_{MF}(\mathcal{Y}, x, g; \mu(\theta))]}{\exp[\kappa_{MF}(\mathcal{Y}, x, g; \mu(\theta'))]} \frac{p(\theta')}{p(\theta)} \frac{q_\theta(\theta'|\theta)}{q_\theta(\theta|\theta')} \right\}$$

While this procedure is very appealing for his simplicity, it is unstable in many applications. In particular, it is not clear if the variational MH algorithm provides a proper markov chain. In recent work de Freitas et al. (2001) make some progress, showing an adaptive procedure where the mean-field approximation is used as proposal in the MCMC scheme and the posterior is estimated using an approximated MH scheme.

An additional alternative consists of the *Variational Bayes*, where the variational approximation is applied directly to the posterior (Bishop (2006)). We will not exploit this approximation here.

3.4. MCMC methods. As shown above the metropolis-hastings algorithm cannot be implemented in this case. The reason lies on the fact that the posterior distribution is a doubly-intractable distribution: both likelihood and posterior contain an intractable normalizing constant, which cannot be computed or evaluated in reasonable time.

The literature has relied on several algorithms to perform the intractable estimation. The exchange algorithm, first developed by Murray et al. (2006), allows estimation of the posterior distribution by using an auxiliary variable. The algorithm consists of proposing a new parameter at each iteration, as in the traditional metropolis scheme; given the proposed parameter, another metropolis-hastings sampler generates a (approximate) draw from the likelihood without evaluating it; the parameter is accepted or rejected according to a modified metropolis-hastings ration that includes likelihood ratios computed at the original data, simulated draw from the likelihood, current and proposed parameter. The scheme satisfies the detailed balance condition and therefore it is a proper MCMC sampler, that generates approximate samples from the posterior. Mele (2011b) used this scheme to estimate a network formation model.

4. NUMERICAL EXPERIMENTS

In this section we compare the variational estimates with the output of the exchange algorithm for several simulated datasets.

4.1. A Toy model example. We consider a toy model of social interactions with two parameters to explain the theory and perform some computational experiments. The utility function is given by

$$u_i(y_i, y_{-i}, x, g) = \alpha y_i + \gamma \sum_{j=1}^n g_{ij} y_j y_i$$

and therefore the potential function is

$$(4.1) \quad Q(y, x, g) = \alpha \sum_{i=1}^n y_i + \frac{\gamma}{2} \sum_{i=1}^n \sum_{j=1}^n g_{ij} y_j y_i$$

The variational mean-field approximation is given by

$$\kappa_{MF}(\mathcal{Y}, x, g; \mu(\theta)) = \max_{\{\mu_1, \dots, \mu_n\} \in [0,1]^n} \left\{ \alpha \sum_{i=1}^n \mu_i + \frac{\gamma}{2} \sum_{i=1}^n \sum_{j=1}^n g_{ij} \mu_i \mu_j - \sum_{i=1}^n [\mu_i \ln \mu_i + (1 - \mu_i) \ln(1 - \mu_i)] \right\}$$

Using the recursion (3.18) at parameter θ we obtain an approximated likelihood and estimates for the mean parameters μ .

All the computations (excluding the estimation with the exchange algorithm) are performed using the statistical software R on a Dell Precision T3500 with 2 Intel Xeon CPUs at 3.47GHZ and 24GB of RAM.⁶ For each graph below, I generated the data from a long MCMC simulation, using a Metropolis-Hastings sampler. All the datasets were generated with parameters $\alpha = -1.0$ and $\gamma = 5/n$.⁷ For each example I evaluated the log-likelihood and posterior on a grid of values for the parameters.⁸ The estimates of posteriors are obtained under four different assumptions for the priors. I used independent priors $\mathcal{N}(0, 10^2) \mathcal{N}(0, 10^2)$, $\mathcal{N}(-1, 5^2) \mathcal{N}(5, 5^2)$, $\mathcal{N}(-1, 1^2) \mathcal{N}(5, 1^2)$ and $\mathcal{U}[-5, 5] \mathcal{U}[0, 10]$.

4.2. Bayesian Estimation with Exchange algorithm. The estimates using the exchange algorithm are shown in Figure 4.1. The prior for the figure is $\mathcal{N}(0, 10^2) \mathcal{N}(0, 10^2)$. The code written in R was extremely slow and inefficient, therefore the code used in estimation is written in Fortran 90 for optimal speed. The results show that for $n = 100$ (left graph) we are able to obtain a good approximation of the posterior, and the posterior mode is quite close to the parameter vector that generated the data. The approximation deteriorates when $n = 1000$ (right graph). To obtain a better approximation we could run additional simulations: however, the computational time required to obtain the posterior for $n = 100$ was of 3 minutes; and for $n = 1000$ was of 25 hours, with an optimized Fortran code.

I ran several examples and the results are similar. The exchange algorithm time-consuming. To run a simulation with $n = 1000$, when the parameters are simulated for 10000 iterations and at each iteration the interaction process is simulated 1000 log 1000 times,⁹ it takes about 3011.41s if $p = 0.10$ and 5930.25s if $p = 0.50$.

⁶This leaves enormous room for improvement in the speed of the algorithms, because introducing loops in R is extremely inefficient from a computational point of view. Codes in Fortran or C++ would be faster by several magnitudes.

⁷The re-scaling by factor n guarantees that the two sufficient statistics are of the same order when the number of players grows large.

⁸Finding the modes using standard algorithms, e.g. Nelder-Mead, takes few seconds.

⁹This number of simulations is suggested by Diaconis and Chatterjee (2011) using a convergence rate argument for the Metropolis-Hasting chain that samples from the likelihood

FIGURE 4.1. Posterior estimated using the exchange algorithm

The figures represent the level curves of the posterior distributions estimated using the variational mean field algorithm. The **red triangle** is the true parameter that generated the data, the **blue dot** is the estimated posterior mode. In this example $n = 100$ (left) and $n = 1000$; and the data are generated from an Erdos-Renyi graph with $p = 0.10$. Priors are independent normals.

4.3. Approximate Bayesian Estimation. The variational approach consists of approximating the likelihood using the variational mean-field algorithm and then compute the approximate posterior. The first example is a network with $n = 100$ players. The network is generated as a random Erdos-Renyi network with probability of a link $p = 0.10$ and the prior is $\mathcal{N}(0, 10^2) \mathcal{N}(0, 10^2)$.¹⁰

In Figure 4.2 we show the estimated posteriors: for independent normal priors. The figure 4.2A represent the posterior estimated using the exchange algorithm; figure 4.2B is the variational approximation in (3.18), while the figure 4.2C uses the simpler mean-field approximation in (3.19).

The results are encouraging: the posterior modes are reasonably close to the “true” parameters even when we adopt a very flat prior. The variational approximation is close to the results obtained with the exchange algorithm, while the computational burden of the mean-field algorithm is quite small: few seconds to recover the posterior in 4.2C and few minutes for the one in 4.2B. Since the two variational approximations are quite close in the numerical examples, I report only the simpler mean-field results in the remaining of the paper.

In Figure 4.3 we show the estimated posteriors when $n = 500$ and the network is a random graph with probability of a link $p = 0.10$. The estimates obtained with the exchange algorithm and the ones with the variational approach are again very similar. Notice that the variability of the variational posterior is smaller than the one estimated by the exchange algorithm. This is a known property of the variational mean-field approximation: it leads to underestimation of variability.

The results for models with $n = 100$ and Erdos-Renyi network generated with $p = .50$ are as encouraging. These are shown in Figure 4.4.

¹⁰This guarantees that each individual has an average of 10 neighbors.

FIGURE 4.2. Level curves of posterior

The figures represent the level curves of the posterior distributions estimated using the exchange algorithm and the two variational mean-field algorithms. The **red triangle** is the true parameter that generated the data, the **blue dot** is the estimated posterior mode. In this example $n = 100$ and the data are generated from an Erdos-Renyi graph with $p = 0.10$. Priors are independent normals.

FIGURE 4.3. Level curves of posterior

The figures represent the level curves of the posterior distributions estimated using the exchange algorithm and the simpler variational mean field algorithm. The **red triangle** is the true parameter that generated the data, the **blue dot** is the estimated posterior mode. In this example $n = 500$ and the data are generated from an Erdos-Renyi graph with $p = 0.10$. Priors are independent normals.

FIGURE 4.4. Level curves of posterior

The figures represent the level curves of the posterior distributions estimated using the exchange algorithm and the variational mean field algorithm. The **red triangle** is the true parameter that generated the data, the **blue dot** is the estimated posterior mode. In this example $n = 100$ and the data are generated from an Erdos-Renyi graph with $p = 0.50$. Priors are independent normals.

The main message seems to be that the variational approximation works for approximate bayesian estimation in models of moderate size. Ideally we want to compare these posteriors to the posterior estimated using the exchange algorithm. However, the exchange algorithm was too expensive to be run in R. We are developing Fortran routines for both MCMC methods and variational methods, to compare precision and speed of computation.

In terms of speed of computation, we obtain the posterior evaluated on the grid in few seconds/minutes. The grid we used is 121×441 and it is evaluated recursively: this code is extremely inefficient in R because it involves for loops.

To understand and have an idea about comparisons with the exchange algorithm, it takes about 30 minutes to generate the data from the stationary distribution of the model using a Gibbs sampler. This means that *for each* parameter proposed by the exchange algorithm, the code written in R should run for about 30 minutes before proposing a new parameter. Therefore the variational method provides a much faster alternative for these examples, while maintaining a similar level of precision in the approximation of the posterior.

5. DISCUSSION AND CONCLUSION

Estimation of peer effects and endogenous social interactions is very important for policy. I proposed a strategic model of social interaction, which can be estimated using only one observation of the interactions. The likelihood of the model is intractable and the estimation is usually performed using a variation of the exchange algorithm. However, when the number

of players grows large, the algorithm may become computationally infeasible. I propose alternative approximation methods, based on variational mean-field theory. The algorithm approximates a complicated likelihood with a simpler and more tractable distribution. The distribution is then optimized to look “close” to the true likelihood: it can be shown that the variational approximation minimizes the Kullback-Leibler distance between the likelihood and the approximating distribution.

The method is tested using few artificial datasets, with encouraging results. Using the variational algorithm, I perform approximate maximum likelihood inference and approximate bayesian inference. I compare results with the MCMC output of the exchange algorithm, showing that the mean-field variational approximation provides reliable and fast inference for the examples examined in this work.

Future research will be devoted to understand the reliability of the estimation method in more complex models and if the usual asymptotic properties of the estimators are satisfied by the mean-field method. There are alternative variational approximations based on different restrictions on the family of approximating distributions, which can be exploited in this setting. Unobserved heterogeneity can be treated using the same variational algorithms without significant additional computational burden.

The two algorithms for mean-field approximations used in this paper differ in their computational efficiency for bigger samples. The complexity of the traditional mean-field algorithm scales linearly with the number of players, while the simpler version scales only with the number of parameters. From the simpler examples shown here it seems that this computational gain does not impose a trade-off in terms of precision. While preliminary, these results are extremely encouraging, and deserve further investigation.

REFERENCES

- Airoldi, E, DM Blei, SE Fienberg and EP Xing (2008), ‘Mixed-membership stochastic block-models’, *Journal of Machine Learning Research* **9**, 1981–2014.
- Amir, Eyal, Wen Pu and Dorothy Espelage (2012), Approximating partition functions for exponential-family random graph models, NIPS Conference 2012.
- Badev, Anton (2013), Discrete games in endogenous networks: Theory and policy.
- Bhamidi, Shankar, Guy Bresler and Allan Sly (forthcoming), ‘Mixing time of exponential random graphs’, *Annals of Applied Probability*.
- Bishop, Christopher (2006), *Pattern recognition and machine learning*, Springer, New York.
- Blume, Lawrence E. (1993), ‘The statistical mechanics of strategic interaction’, *Games and Economic Behavior* **5**(3), 387–424.
- Braun, Michael and Jon McAuliffe (2010), ‘Variational inference for large-scale models of discrete choice’, *Journal of the American Statistical Association* **105**(489).
- Chandrasekhar, Arun and Matthew Jackson (2012), Tractable and consistent exponential random graph models.
- de Freitas, Nando, Pedro Højen-Sørensen, Michael Jordan and Stuart Russell (2001), Variational mcmc, in ‘Proceedings of the Seventeenth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-01)’, Morgan Kaufmann, San Francisco, CA, pp. 120–127.

- Diaconis, Persi and Sourav Chatterjee (2011), Estimating and understanding exponential random graph models.
- Geyer, Charles and Elizabeth Thompson (1992), ‘Constrained monte carlo maximum likelihood for deperdent data’, *Journal of the Royal Statistical Society, Series B (Methodological)* **54**(3), 657–699.
- Grimmer, Justin (2011), ‘An introduction to bayesian inference via variational approximations’, *Political Analysis* **19**(1), 32–47.
- Jaakkola, Tommi (2000), *Advanced mean field methods: theory and practice*, MIT PRESS, chapter Tutorial on Variational Approximation Methods.
- Lehman, E. L. (1983), *Theory of Point Estimation*, Wiley and Sons.
- Liang, Faming (2010), ‘A double metropolis-hastings sampler for spatial models with intractable normalizing constants’, *Journal of Statistical Computing and Simulation* **forthcoming**.
- Liang, Faming, Chuanhai Liu and Raymond J. Carroll (2010), *Advanced Markov Chain Monte Carlo Methods: Learning from Past Samples*, John Wiley and Sons, Ltd.
- Manski, Charles (1993), ‘Identification of endogenous social effects: The reflection problem’, *Review of Economic Studies* **60**(3), 531–542.
- Mele, Angelo (2011a), Segregation in social networks: Monte carlo maximum likelihood estimation. Working Paper.
- Mele, Angelo (2011b), A structural model of segregation in social networks. Working Paper.
- Monderer, Dov and Lloyd Shapley (1996), ‘Potential games’, *Games and Economic Behavior* **14**, 124–143.
- Murray, Iain A., Zoubin Ghahramani and David J. C. MacKay (2006), ‘Mcmc for doubly-intractable distributions’, *Uncertainty in Artificial Intelligence*.
- Nakajima, Ryo (2007), ‘Measuring peer effects on youth smoking behavior’, *Review of Economic Studies* **74**(3), 897–935.
- Ormerod, J.T. and M.P. Wand (2010), ‘Explaining variational approximations’, *The American Statistician* (64), 140–153.
- Schweinberger, Michael and Mark S. Handcock (2013), Hierarchical exponential-family random graph models with local dependence.
- Wainwright, M.J. and M.I. Jordan (2008), ‘Graphical models, exponential families, and variational inference’, *Foundations and Trends® in Machine Learning* **1**(1-2), 1–305.