Inference on the Effects of Observed Features in Latent Space Models for Networks

Zachary Jones Matthew Denny Bruce Desmarais Hanna Wallach

August 30, 2016

Abstract

The latent space model (LSM) for network data is a generative probabilistic model that combines a generalized linear model with a latent spatial embedding of the network. It has been used to decrease error in the estimation of and inference regarding the effects of observed covariates. In applications of the LSM, it is assumed that the latent spatial embedding can control for unmeasured confounding structure that is related to the values of edges in the network. As far as we know, there has been no research that considers the LSM's performance in adjusting for unmeasured structure to reduce estimation and inferential errors. We investigate the LSM's performance via a Monte Carlo study. In the presence of an unmeasured covariate that can be appropriately modeled using a latent space, estimation and inferential error remain high under even moderate confounding. However, the prediction error of the LSM when unmeasured network structure is present is substantially lower in most cases. We conclude that the LSM is most appropriately used for exploratory or predictive tasks.[1]

1 Introduction

Inferential analysis of political network data has grown increasingly sophisticated in recent years. Political networks scholars are well versed in the risks associated with ignoring unmodeled network structure. Dependencies such as reciprocity, transitivity, and homophily—if not accounted for—can lead to biased estimates and errors in hypothesis testing, much in the way that omitted variable bias can affect results in conventional regression models (**???**). A number of statistical modeling frameworks have been proposed to account for confounding structure in network data that cannot be modeled with observed covariates. These include the exponential random graph model (ERGM) (e.g., **???**), the latent space model (LSM) (e.g., **???**), and the stochastic actor oriented model (SAOM) (e.g., **??**).

Despite their growing popularity, few studies exist that investigate the performance of these models in adjusting for confounding network structure. The approach to adjusting for dependencies in the two other models commonly used for network data—ERGM and SAOM—is quite similar to adjusting for confounding covariates in regression modeling. The researcher specifies a set of dependencies that (s)he hypothesizes to be important in the generative model for the network. These dependencies are then explicitly included in a model that simultaneously represents the effects of

[1] This work was supported by National Science Foundation grants DGE-1144860, SES-1558661, SES-1619644, and CISE-1320219. Any opinions, findings, and conclusions or recommendations are those of the authors and do not necessarily reflect those of the sponsor.

observed covariates (**?**). The LSM takes a different approach, which involves the incorporation of latent variables to model network structure. The LSM has the advantage over ERGM and SAOM in that researchers need not develop a set of hypothesized dependencies in order to model network structure that is not reflected in observed covariates. However, this advantage hinges upon the capacity for the LSM to discover unmodeled structure that could otherwise be attributed to the observed covariates (i.e., inferring confounding structure). In the current study, we focus on the LSM, examining its performance in reducing estimation and inferential errors regarding the effects of observed covariates, via adjustment for confounding network structure.

1.1 Central Problem

Latent variable inference, generally conceived, presents the possibility of representing unmeasured data in statistical models. The LSM, introduced by **?**, is used to estimate the effect of covariates in the presence of latent network structure. Here the distance function $|z_i - z_j|$ represents latent network structure as homophily with respect to latent variables. The distance function is additively combined with a regression on observed dyadic covariates, x_{ij}, to form a linear predictor for tie prediction. As with a GLM, a link function, g^{-1}, maps this linear predictor to the appropriate edge distribution.

$$\mathbb{E}(y_{ij}|x_{ij}) = g^{-1}(\alpha + \beta x_{ij} - |z_i - z_j|)$$

However, it is unclear that the introduction of a latent space decreases the expected error for the parameter(s) of interest when aspects of the network structure (e.g., homophily) that are unmeasured, are correlated with measured covariates. There are two reasons that using the LSM may lead to increased error. First, the latent configurations inferred may result in a representation of the network wherein a node's position in the latent space is spuriously correlated with the observed covariates, leading to reduced efficiency. Second, if the unobserved (i.e., latent) network structure is truly correlated with the observed covariates, the unobserved structure that can be correlated with the observed variable may be attributed to the observed variable, while the latent space parameters are used to model other sources of variation.

2 Applications and Development of the Latent Space Model

The LSM has seen use in a variety of fields in which network data is common, particularly the social sciences. The apparent appeal of the LSM appears to be driven primarily by the LSM's usefulness in modeling transitivity (i.e., clustering) and homophily, which are ubiquitous in social networks. In political science the LSM and variations on the form developed in **?** have been used to estimate the effect of democracy on the probability of a militarized interstate dispute (**?**), the amount of portfolio investment between states (**?**), and the effect of multimember districts on the probability of collaboration between state legislators in the United States (**?**). Variations of the LSM developed for networks measured over time have been applied to the study of international trade, wherein the effects of various features of trading partners are estimated (**?**). In ecology the LSM has been used to study the sociality of elephants (**?**) and orcas (**?**), birds (**?**), to discover ecological communities (**??**), and to study food webs (**?**). In epidemeology it has been used to identify clusters of infected persons for later isolation (**?**) and to study patterns of interaction amongst physicians (**?**). In marketing and business research it has been used to study inter-group trust (**?**), optimal bundling and pricing of goods and brands for retailers (**?**). It has been used to describe topic-specific patterns of interaction in e-mail communication networks (**?**). Lastly, in neuroscience it has been proposed as a method for modelling fMRI data (**?**).

Although the LSM is arguably most useful as an exploratory or predictive model (see **?** for a discussion of the differences between predictive and explanatory modeling), it has been applied in some cases to reduce estimation and/or inferential error with respect to the effects of observed covariates. For example **?**, in a high profile example, argue that the the LSM improves inference about the effects of democracy, international trade, and participation in international organizations on the probability of inter-state conflict.

> "The history of international disputes, and consequently the extant data on militarized interstate disputes, is replete with ... dependencies. We formally incorporate and estimate the extent of these ... dependencies in our model of the Kantian peace in order to more precisely determine the effects of the Kantian tripod on international conflict" (**?**, p. 585)

Likewise, similar claims are made in **????**.

> "For the most part, however, most dyadic research in international relations ignores the essential features of dyads in that they fail to satisfy the assumption of independence or, by construction, have missing data but ignore its effects: both of these bias the results in a fundamental way" (**?**, p. 2).

> "This approach combines a network analysis with a standard-looking regression to permit us to access the importance of our explanatory factors without having them biased by the interdependencies in the network we are studying" (**?**, p. 15).

> "The presence of ... dependence implies misspecification and a high likelihood of bias in most current applications. Building on the latent space framework we model the world trading system without assuming a particular network structure or the sufficiency of particular network statistics" (**?**, p. 20).

> "... the latent space model allows for the assessment of distance between two unconnected actors while simultaneously controlling for the interdependence inherent in network data. This interdependence in latent space positions allows the model to control for common network effects like reciprocity or transitivity that would ordinarily bias results" (**?**, p. 336).

The above examples are drawn from political science, but we see a similar logic for using the LSM articulated in recent work in ecology by **?**, p. 989.

> "…can create artificially exaggerated synchrony rates regardless of motivation for signaling, which can be modeled with a term known as "transitivity" in the social network literature. A latent space model (Krivitsky et al. 2009) was used to examine the propensities for synchrony between helper males and the primary male while accounting for the variability deriving from the transitivity."

Though this list of quotations is by no means complete, the last example to which we point comes from the business literature, and uses the bilinear form of the latent space model (**?**, p. 7).

> "This model controls for higher order team dynamics using the bilinear component $z_i'z_j$ and also allows both trustor ($x_{tor,i}$) and trusted ($x_{ted,j}$) characteristics to be investigated along with the dyadic covariates ($x_{d,i,j}$)."

If the LSM does reduce inferential/estimation error under certain conditions, it will serve as a valuable general model for explanatory analysis of network data, especially since its use does not require the researcher to specify a set of network dependencies from theory. However, as far as we know, there has been no research into the performance of the LSM in adjusting for confounding network structure. Additionally, the predictive performance of the LSM has only seen limited evaluation (**?**), despite having been used for model selection (**????**). Hence, finding the conditions under which the LSM reduces prediction error also may affect future use of the LSM.

The LSM has seen a substantial amount of subsequent development and extension. The latent space has been represented as a k-dimensional Euclildean space and by latent factors (**??**). Additional structure has been introduced by adding random effects, which, for example, may involve sender or receiver specific effects for directed networks which capture differential activity rates amongst nodes (**?**). Within this framework **?** models dynamic network data by treating the latent space as a stochastic process. **?** enable the LSM to model clustering that is not representable as homophily (i.e., stochastic equivalence) by combining the LSM with latent cluster models. **?** shows that the LSM and latent cluster models are special cases of an "eigenmodel." That is, an eigendecomposition of a symmetric sociomatrix can be used to represent both latent space and cluster models, but not vice-versa. Most of these developments have represented new forms for the latent space and/or the further development of the LSM as a Bayesian hierarchical model. In the next section we propose an approach to specifying the Bayesian prior in order to improve inference regarding covariate parameters.

3 Bayesian Inference and Priors in the Latent Space Model

Maximum likelihood estimation (MLE) is problematic in the LSM since the value of the likelihood function is invariant to rotation of the latent space (**?**) (i.e., only the distances between points matter; not their absolute positions), which means that the ML estimate is unidentified. As such, inference in the LSM is conducted using a Bayesian approach. Latent positions and other parameters are sampled using a Metropolis-Hastings algorithm. After sampling, latent space positions are rotated to match a common set of positions, via Procrustes transformation (**?**), in order to eliminate variance due to rotation of the latent space. **?** propose the use of diffuse independent normal priors for the latent positions and regression parameters. They also recognize the fact that isolates' (i.e., nodes with no ties) finite positions are only identified through the prior in the LSM. In one application they actually just exclude the isolates. Excluding isolates seems fine if the inferential purpose is to estimate positions in the latent space, but for representative explanatory analyses excluding isolates would amount to selecting on the dependent variable. Unlike in many practical applications of Bayesian inference, it is not possible to use an improper "flat" prior (**?**) with the LSM, as an informative prior is necessary to assure finite latent positions. Since it is necessary to use an informative prior with the LSM, we ask whether it is possible to specify the prior in a way that improves inference regarding covariate parameters. We propose to calibrate the prior to be on a scale comparable to the linear predictor estimated from a GLM. This scaling should discourage the latent space from replicating the linear predictor, encouraging the discovery of structure that remains in the data after identifying the effects of measured covariates.

To derive our scaling rule, we follow the thought experiment of a one-dimensional latent space $\mathbf{z}$ with a dyadic covariate $\mathbf{x}$. If we assume that the $\mathbf{z}$ are independent and distributed $\mathcal{N}(0, \sigma^2)$, it is straightforward to derive the distribution of $|z_i - z_j|$.[2] Since it is a special case of a "normal difference distribution", we know that $z_i - z_j \sim \mathcal{N}(0, 2\sigma^2)$ (**?**). The normality of $z_i - z_j$ implies that $|z_i - z_j|$ has a folded normal distribution (**?**). Following from this result, we know that the variance of $|z_i - z_j|$ is

$$\sigma^2(2 - 4/\pi).$$

Let θ^2 be the empirical variance of $\mathbf{x}$. Then the variance in the linear predictor is $\beta^2\theta^2$. Given an estimate $\hat{\beta}$, we can set the variance of the latent space prior to a multiple of $\hat{\beta}^2\theta^2/(2 - 4/\pi)$ in order to tune the prior to avoid replicating the linear predictor via the distances between latent coordinates.

3.1 Simulation Study

We use a simulation study to evaluate the LSM's performance at inferring covariate parameters in the situation in which we have some observed covariates as well as omitted network structure which can be represented using a Euclidean latent space. We are particularly interested in whether the LSM can adjust for confounding when the confounding variable is unmeasured. In the simulation design, we study how the level of confounding affects the performance of the LSM relative to the GLM. We evaluate performance in terms of bias, inferential error, and prediction.

[2] The assumption that $\mathbf{z} \sim \mathcal{N}(0, \sigma^2)$ is consistent with $\mathbf{z}$ being drawn from the prior that is conventionally used in the LSM.

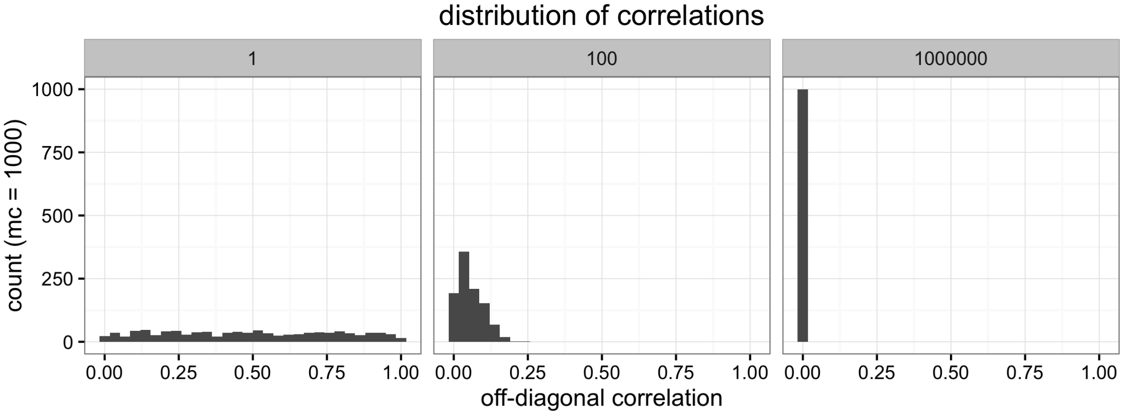

Figure 1: The distribution of the absolute value of the correlation between the observed covariate and the latent distances.

3.2 Simulation Setup

For simplicity and computational efficiency we consider a unidimensional latent space. We have no reason to suspect that our results should be unique to unidimensional LSMs. To generate an observed covariate which has a controllable collinearity with the latent network structure we follow a three-step process. First, we simulate unidimensional positions for each node, drawing from a normal distribution, and calculate the Euclidean distance $\mathbf{d}$ between each pair of positions. Second, given a target covariance matrix (Σ) among the covariates and distances, $\langle \mathbf{x}, \mathbf{d} \rangle$, we derive the conditional mean vector and covariance, assuming that $\mathbf{x}$ has a normal distribution given $\mathbf{d}$ (see (**?**, pp. 116–117) for the conditional normal derivation). Third, we simulate x as a normal random variable with the respective conditional means and covariance. Finally we standardize x to have zero mean and unit variance. To generate the covariance matrix Σ which controls the dependence between the omitted network structure and the observed covariate $\mathbf{x}$ we utilize the C-vine method of **?**.

We consider three exponential family distributions from which the adjacency matrix entries are drawn: Gaussian, binomial, and Poisson. Additionally we control the number of nodes in the network $n = 25, 50, 100$. To vary the degree of confounding attributable to the latent positions, we consider three values of the collinearity parameter η, 1, 100, and $1,000,000$, where 1 corresponds to a standard uniform distribution of the correlation (i.e., moderate collinearity) between the observed covariate and the latent distances, and $1,000,000$ to independence between the observed covariate and the latent distances. See Figure **??** for the distribution of the absolute value of the correlation generated at each value of η.

We consider the LSM with several different priors on the coefficient for the observed covariate and the latent space. We set the prior variance of β to be either 1 or 10 and alternatively use a diffuse normal prior on the latent space or scale it by $\hat{\beta}^2\theta^2/(2-4/\pi)$, where θ^2 is the empirical variance of x.

The LSM is estimated using the canonical implementation in the `latentnet` package in `R` (**?**). An initial run of 10,000 burn in iterations, followed by 1,000,000 iterations of the sampler. Every 100th iteration is saved. Convergence in the log probability of the model is assessed using the Geweke diagnostic in the `coda` package (**??**). If the convergence criterion is satisfied the simulation continues to the next set of arguments, otherwise the number of iterations is doubled. If the convergence criteria is still not satisfied, then the aforementioned step in the simulation is flagged for review. At each point in the simulation's parameter space, we execute 1,000 Monte Carlo iterations. [3]

We compute the MLE estimated by iteratively reweighted least squares via the `glm` function in `R` (**?**). For the LSM we use the posterior mode as our point estimate. In the cases where edges are binomial and there is omitted network structure, we scale the estimates using the reciprocal of the bias of β when $\mathbf{x}$ and $\mathbf{d}$ are independent: $\sqrt{\frac{3.28+\beta^2\text{Var}(\mathbf{d})}{3.29}}$, where 3.29 is the variance of a standard logistic distribution. We do this to adjust for bias in the coefficient estimate that arises due to the lack of a scale coefficient in logistic regression (See the derivation of the bias under an independent but omitted covariate in **?**).

For each combination of simulation conditions we evaluate the mean square prediction error of the model on new edges drawn condtional on $\mathbf{x}$ and $\mathbf{d}$, the bias of the estimated coefficient for the measured covariate, and the Type-1 and 2 error rates. Inference regarding Credible intervals for β are defined by using the region of highest posterior density which covers 95% of the marginal posterior distribution **?**. We consider β to be statistically significant at the 0.05 level if the credible interval does not contain 0.

3.3 Results

To evaluate estimation error we compute the bias. Figures **??** and **??** shows the results. In Figure **??** we can see that with a uniform distribution over the correlation between the distances and the observed covariate, all methods (except the true model wherein the distances are treated as observed) show substantial bias. In most cases the LSM performs better than the GLM, sometimes by as much as 50%. However, in absolute terms the amount of bias is large. Our results demonstrate that using the LSM will not permit the discovery of and adjustment for the true latent positions. For both the GLM and the LSM, bias decays rapidly as the degree of correlation between the distances and the measured covariate decreases. In **??** we can see that when there is no ommitted network structure, the LSM exhibits bias greater than that of the GLM, though it does appear that scaling the prior of the latent positions to match the scale of the linear predictor decreases this tendency.

Figures **??** and **??** show the inferential error rates of the LSM with different priors for β and the latent positions, as well as a GLM. Type-1 error rates shown in Figure **??** for the LSM are in general

[3] One of the primary difficulties in executing the above simulation design is the computational cost of estimating the LSM. We utilize the `BatchExperiments` R package to construct and execute our computational experiments on a Torque cluster **?**.

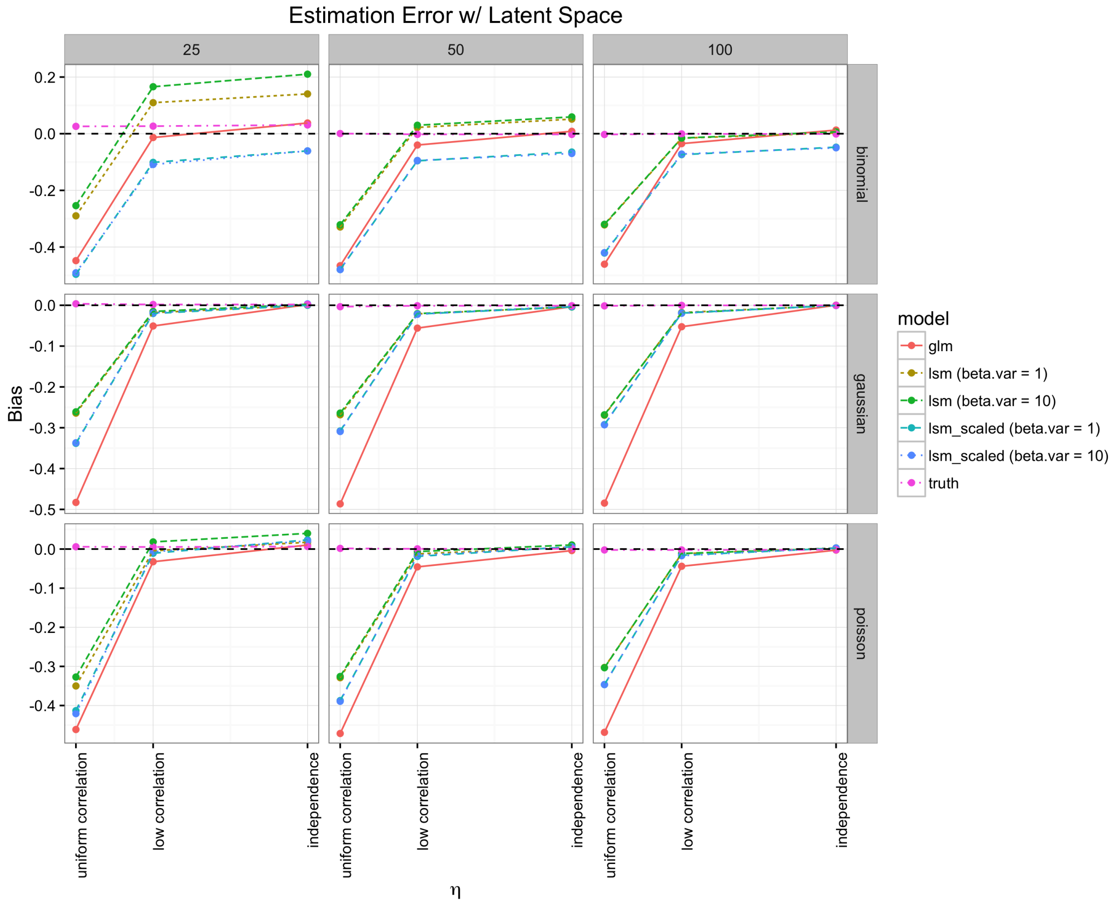

Figure 2: The bias of estimates of the effect of the observed covariate $\mathbf{x}$ when there is an ommitted variable. The x-axis gives the value of the parameter η which controls the degree of dependence between $\mathbf{x}$ and ommitted covariate. Lower values of η indicate higher levels of dependence between the observed and ommitted covariate. The y-axis gives a Monte Carlo estimate of the bias. The number of nodes are indicated in the top panels, while the distributional family of the edges is shown on the right panel. Each panel represents 4 values of η with 1,000 Monte Carlo iterations executed at each point.

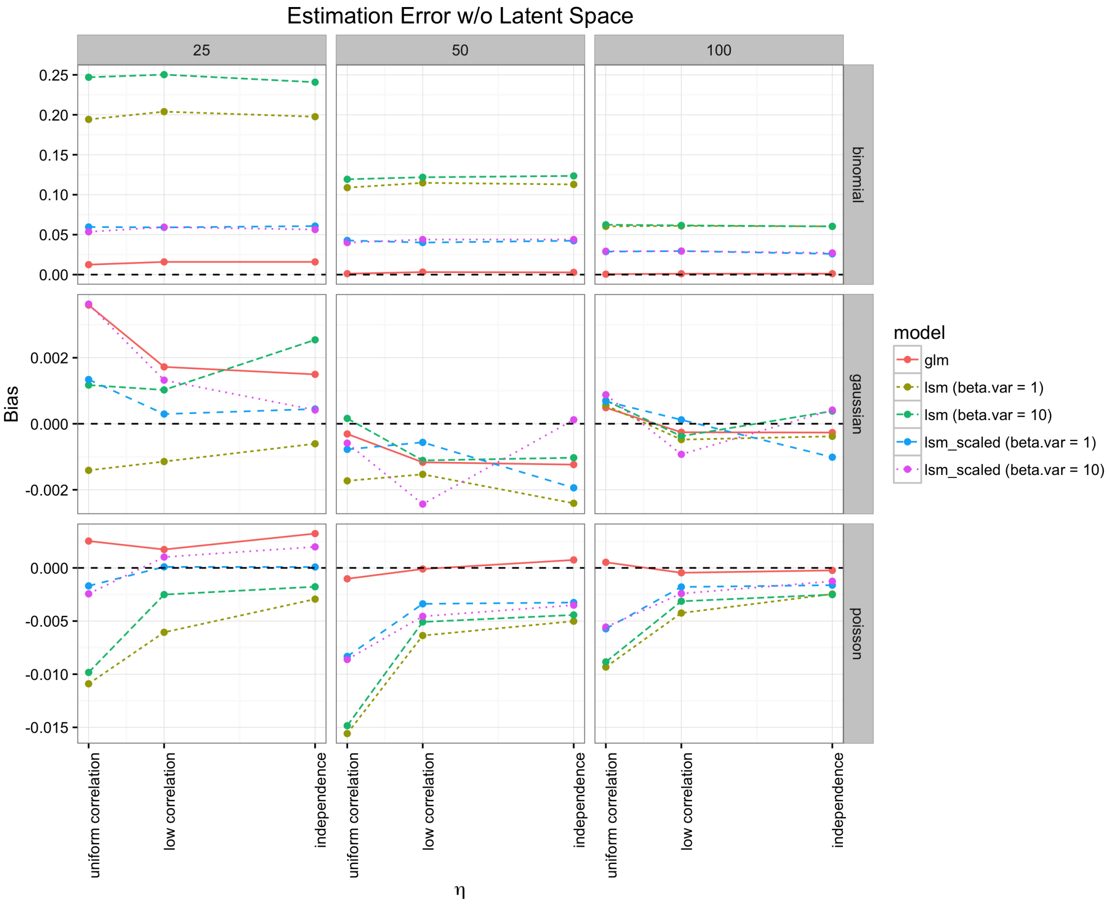

Figure 3: The bias of estimates of the effect of the observed covariate $\mathbf{x}$ when there is no ommitted variable.

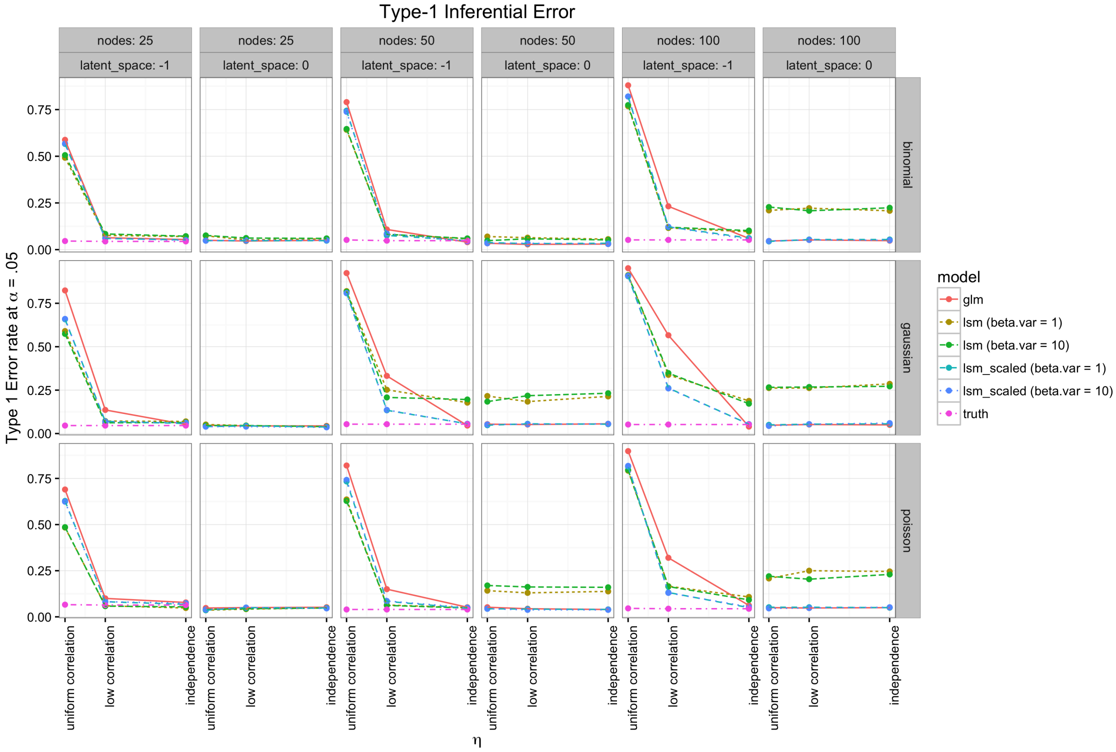

Figure 4: Monte Carlo estimates of the Type-1 error regarding the effect of the observed covariate $\mathbf{x}$ are shown on the y-axis. Here, $\beta = 0$ and the error rate shown in each panel in that row gives 1 minus the probability of a 95% confidence region (for the LSM) or interval (for the GLM) including 0, giving the probability that a true null hypothesis of $\beta = 0$ is falsely rejected.

lower than those of a GLM when an unmeasured covariate exists. Under uniform correlation between the omitted network structure and the observed covariate the Type-1 error rate is often more than 10 times greater than the nominal rate of 0.05. Again, scaling the prior of the latent space to match that of the observed covariate appears to make the LSM's error rate comparable to that of the GLM. Type-2 error rates, shown in Figure **??** are also substantially above their nominal rate when there is a uniform distribution on the strength of confounding, though the inflation is not as bad as that of the Type-1 rates. These results indicate that, in the presence of unobserved confounders that can be represented through the LSM, the LSM does not adequately correct the inferential errors that would arise due to omitted variables in the conventional regression framework.

We estimate generalization error as the expected prediction error on new edges generated using a fixed set of latent positions for the nodes and a fixed observed covariate. Generalization error results are shown in Figure **??**. In nearly all cases the LSM substantially outperforms the GLM, especially when the omitted covariate is not highly collinear with the observed covariate. In nearly

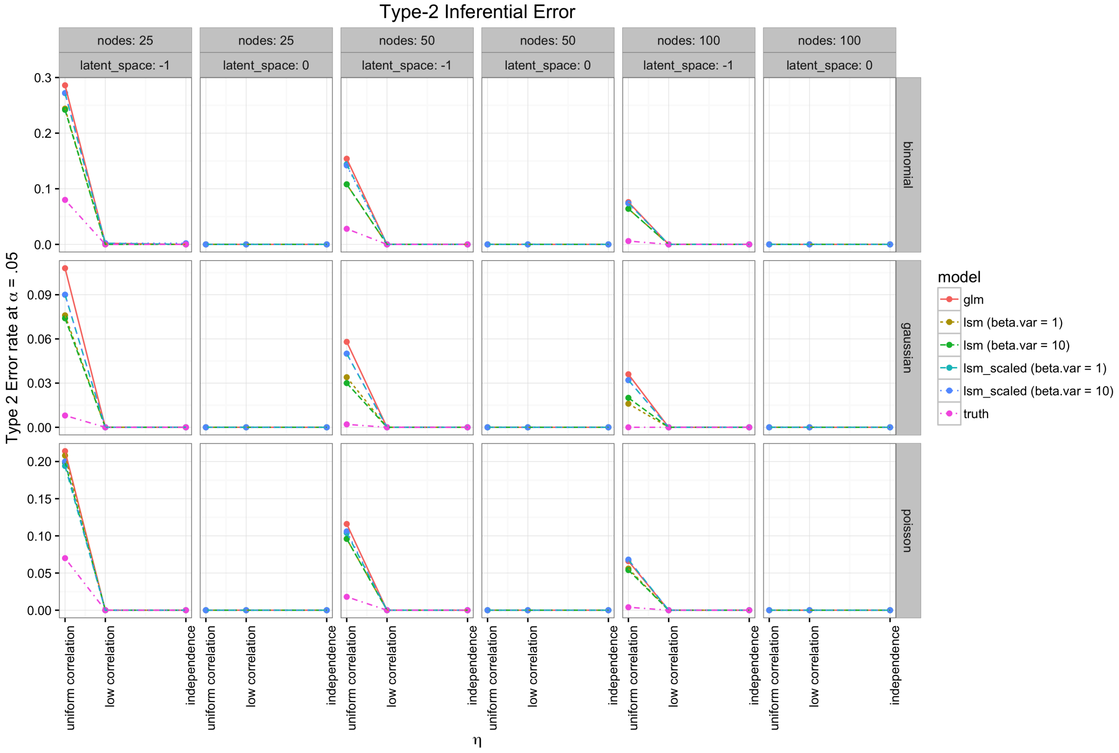

Figure 5: Monte Carlo estimates of the Type-2 error regarding the effect of the observed covariate $\mathbf{x}$ are shown on the $y - axis$. Here, $\beta = 1$, and the error rate shown in each panel in each row gives the probability of the probability/confidence intervals covering 0, giving the probability of accepting a false null hypothesis.

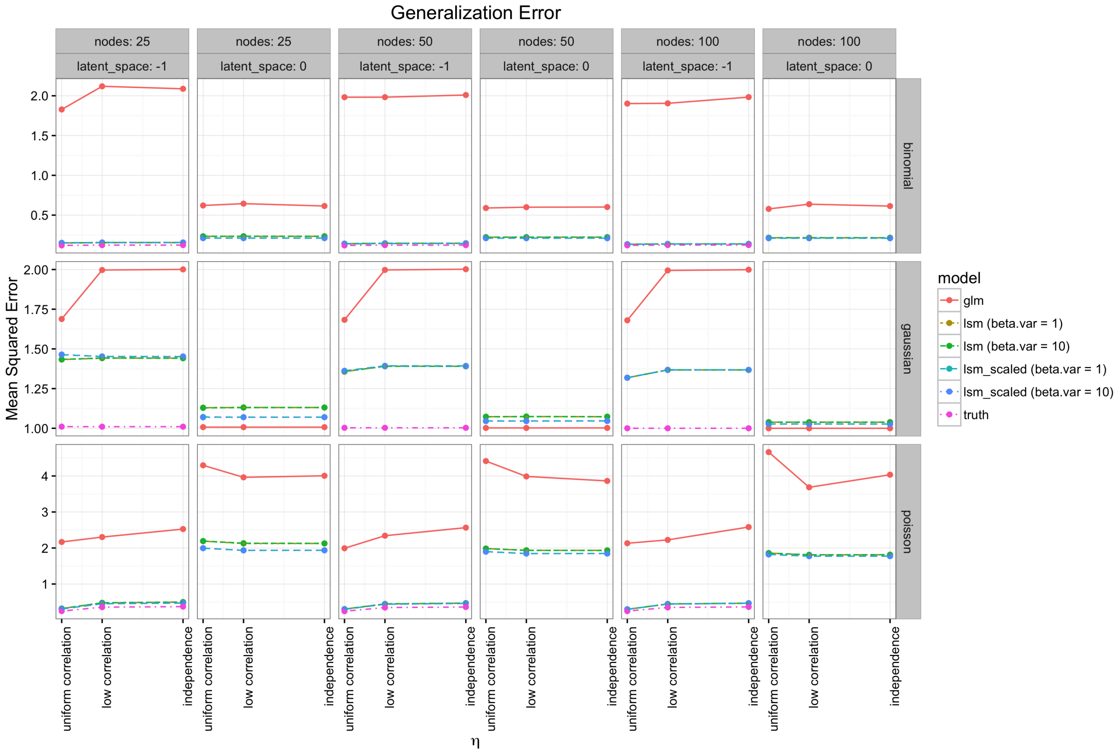

Figure 6: The x-axis shows a Monte Carlo estimate of the mean square error for edge values drawn from the appropriate distribution with the observed covariate $\mathbf{x}$ and the latent positions $\mathbf{z}$ fixed. In the binomial case the Brier score is computed and in the poisson and normal cases the mean-square-error.

all cases it performs as well as the true model.

4 Conclusion

Based on our simulation study, the primary advantage of the LSM is that the latent space provides an efficient complement to observed covariates when it comes to fitting and predicting the network. Considering both (1) that the LSM does not substantially reduce bias or inferential error relative to the GLM, and (2) that the LSM exhibits strong predictive performance, we infer that the latent positions are fit to explain the systematic variation that cannot be explained through the observed covariates. This is of considerable use in research focused on developing a predictive model, or fitting and exploring latent positions. However, since the latent positions are inferred to complement the observed covariates—explaining residual variation—the LSM is ill-suited to adjust for confounding network structure, as adjusting for confounding would require that the latent positions explain variation that could otherwise be attributed to the observed covariates.

We conclude that the primary reason for using the LSM or one of its many variants, rather than a GLM, should be interest in using the latent space model without covariates to explore latent structure in the network through a principled approach to embedding, or the prediction of edges in networks where there is likely to be omitted structure that is not adequately modeled using observed covariates. Although inferential and estimation errors are somewhat smaller than that of a GLM when there is omitted network structure, even moderate collinearity between the omitted structure and the measured covariate leads to substantial bias and inferential error. The LSM cannot control for unmeasured confounders. Furthermore, if such network structure does not exist, the LSM may induce additional bias and inferential error, though this effect can be moderated by scaling the prior on the latent positions to match that of the measured covariates. If the researcher is interested in identifying causal effects of observed covariates, but is concerned that unmeasured variables or network structure could confound the relationship between observed variables and the network, the LSM does not represent an advisable alternative to measuring the confounding variables and/or explicitly modeling the endogenous network structure.