Robust RegBayes: Selectively Incorporating First-Order Logic Domain Knowledge into Bayesian Models

Shike Mei[†]   MEI@CS.WISC.EDU
Jun Zhu[§]   DCSZJ@MAIL.TSINGHUA.EDU.CN
Xiaojin Zhu[†]   JERRYZHU@CS.WISC.EDU
[†]Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI, USA 53706
[§]Dept. of Comp. Sci. & Tech., TNList Lab, State Key Lab of Intell. Tech. & Sys., Tsinghua University, China

Abstract

Much research in Bayesian modeling has been done to elicit a prior distribution that incorporates domain knowledge. We present a novel and more direct approach by imposing First-Order Logic (FOL) rules on the posterior distribution. Our approach unifies FOL and Bayesian modeling under the regularized Bayesian framework. In addition, our approach automatically estimates the uncertainty of FOL rules when they are produced by humans, so that reliable rules are incorporated while unreliable ones are ignored. We apply our approach to latent topic modeling tasks and demonstrate that by combining FOL knowledge and Bayesian modeling, we both improve the task performance and discover more structured latent representations in unsupervised and supervised learning.

1. Introduction

Incorporating domain knowledge into the learning process is an effective way to improve the accuracy of predictive tasks (Richardson & Domingos, 2006) or the interpretability of latent representations (Andrzejewski et al., 2011). Bayesian methods provide a rigorous mathematical framework to incorporate domain knowledge via Bayes' rule. Much research has been done on eliciting an informative prior, either directly (Garthwaite et al., 2005) or indirectly by imposing parameter constraints and confidence values (Mao & Lebanon, 2009). Furthermore, Bayesian methods naturally handle noise in domain knowledge, which is especially important when domain knowledge is collected from the crowd, e.g. (Raykar et al., 2010).

However, since the ultimate goal of Bayesian methods is to infer a posterior distribution, it is arguably more direct to impose domain knowledge directly on the posterior distribution. The regularized Bayesian framework (RegBayes) does this via posterior constraints (or equivalent posterior regularization) using a variational representation of Bayes' rule (Zhu et al., 2013b). RegBayes has had significant success in learning discriminative Bayesian models by conjoining max-margin learning and Bayesian nonparametrics (Zhu et al., 2011; Zhu, 2012). Nonetheless, the domain knowledge considered in RegBayes so far has been max-margin posterior constraints, which could be too narrow and inapplicable to unsupervised learning. Furthermore, no existing RegBayes model has explicitly modeled the noise in domain knowledge.

In this paper we introduce Robust RegBayes, a principled framework to robustly incorporate rich and uncertain domain knowledge in both unsupervised and supervised learning tasks. Our contributions are two-fold: First, we greatly extend the scope of RegBayes domain knowledge by allowing First-Order Logic (FOL) rules. To achieve this, we use groundings of the FOL formulas and define features as expected number of groundings in which the formula is true. In producing FOL domain knowledge, domain experts are often able to focus on high-level modeling goals of the application domain. Second, we explicitly model the uncertainty in domain knowledge using a spike-and-slab prior. This allows us to *automatically* and *selectively* incorporate high-quality domain knowledge while ignoring low-quality ones. Our experiments on Robust RegBayes, especially on various latent Dirichlet allocation (LDA) (Blei et al., 2003) tasks, convincingly demonstrate improved task performance and topic interpretability in both unsupervised and supervised settings. Compared to First-Order Logic LDA (Fold·all, a state-of-the-art framework to incorporate FOL rules into LDA) (Andrzejewski et al., 2011) which requires experts to manually set the weights of FOL rules, Robust RegBayes automatically learns the weights. Com-

pared to max-margin supervised LDA that incorporates word features (Zhu & Xing, 2010), it discovers more interpretable topics without sacrificing prediction accuracy.

2. The Robust RegBayes Framework

2.1. RegBayes with FOL Domain Knowledge

We first review the RegBayes framework (Zhu et al., 2013b). Consider a generic Bayesian latent variable model with observed random variables $\mathbf{X} \in \mathcal{X}$ and hidden variables $\mathbf{H} \in \mathcal{H}$. Standard Bayesian inference calculates the posterior distribution $p(\mathbf{H} \mid \mathbf{X})$ from a prior $p_0(\mathbf{H})$ and a likelihood model. It is often difficult to make sure that the posterior satisfies all domain knowledge constraints. In contrast, the RegBayes framework allows domain knowledge to directly influence the posterior. RegBayes does so by penalizing distributions that differ in the expected value of feature functions. Each feature function, denoted as ϕ_l, and the "belief label" of the feature, denoted as γ_l, are induced from domain knowledge. Formally, the RegBayes inference procedure is defined as a constrained optimization problem:

$$\min_{q(\mathbf{H}) \in \mathcal{P}, \boldsymbol{\xi} \in \mathbb{R}^L_+} \mathrm{KL}\left(q(\mathbf{H}) \parallel p(\mathbf{H} \mid \mathbf{X})\right) + C \sum_l \xi_l \quad (1)$$

$$\text{s.t. } \left| \mathbb{E}_{q(\mathbf{H})}\left[\phi_l(\mathbf{H}, \mathbf{X})\right] - \gamma_l \right| \le \epsilon + \xi_l,$$

where $\mathcal{P}$ denotes the appropriate probability simplex; $p(\mathbf{H} \mid \mathbf{X})$ is the posterior distribution via Bayes' rule; $\boldsymbol{\xi} \in \mathbb{R}^L_+$ is the vector of L slack variables, one for each domain knowledge constraint; ϵ is a small positive precision parameter; and C is a regularization parameter. The key difference between RegBayes and standard Bayesian model is that the "optimal distribution" $q(\mathbf{H})$ obtained by solving Eq (1) can be different from $p(\mathbf{H} \mid \mathbf{X})$. The standard Bayesian posterior is a special case of RegBayes, as can be seen by setting $C = 0$.

Despite its success, the application of RegBayes so far has been limited to max-margin constraints (Zhu et al., 2011). Max-margin constraints cannot represent many kinds of rich domain knowledge such as those for unsupervised models. To substantially broaden the scope of knowledge used in RegBayes, we consider FOL rules in this paper. FOL is a particularly flexible and powerful knowledge representation. It has the additional benefit of insulating the domain experts from the intricacy of Bayesian inference.

Formally, let R_l be the lth FOL rule represented in Conjunctive Normal Form with logical predicates over instantiations $(\mathbf{h}, \mathbf{x})$ of the variables $(\mathbf{H}, \mathbf{X})$. To tie the rule to RegBayes, we define a feature function ϕ_l to provide finer resolution over the domain knowledge. Specifically, let G_l be the set of groundings of R_l, we define the feature function $\phi_l = \frac{1}{|G_l|} \sum_{g_l \in G_l} \mathbf{1}(g_l(\mathbf{h}, \mathbf{x}))$. Note that this
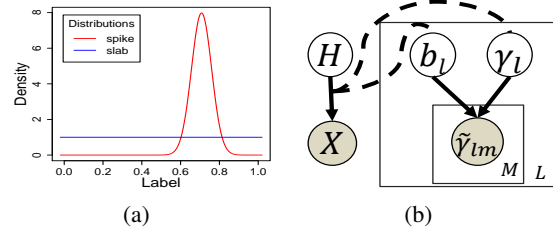

Figure 1. (a) An example of the spike-slab likelihood $p(\tilde{\gamma}_{lm} \mid \gamma_l, b_l)$, where the slab component is a uniform distribution on $[0, 1]$ (blue line) and the spike component is a truncated Gaussian with a small variance (red line). (b) The Robust RegBayes model.

feature function takes value in $[0, 1]$ (rather than $\{0, 1\}$) and captures the fraction of groundings where the rule is true. We let the "golden standard" expectation of rule R_l be $\gamma_l = \mathbb{E}\left[\phi_l(\mathbf{H}, \mathbf{X})\right]$ under the desired distribution. Soliciting γ_l from domain experts is difficult and will be addressed in the next section.

Compared to Markov Logic Network (MLN) which has the goal of modeling FOL rules in probabilistic terms, RegBayes FOL rules are meant to influence a separate Bayesian model. Therefore, RegBayes truly combines FOL and Bayesian modeling. Compared to some other prior work on incorporating FOL into probabilistic models such as Fold·all (Andrzejewski et al., 2011), one major advantage of RegBayes is to automatically learn the FOL rule weights. These weights would be hard (if not impossible) for humans to manually set, especially in a crowd setting. RegBayes learns the rule weights from relatively easier-to-obtain belief labels via solving a dual optimization problem, as we show next.

2.2. Robust RegBayes

The golden standard γ_l for each rule is rarely observed precisely in reality. We solve the problem by treating expert-supplied values of γ_l as noisy observations. Formally, let the FOL knowledge base collected from experts be $\mathrm{KB} = \{R_l, \tilde{\gamma}_l\}_{l=1}^L$. The KB consists of L FOL rules. Each rule R_l is associated with a set of noisy observations $\tilde{\boldsymbol{\gamma}}_l = \{\tilde{\gamma}_{lm} : \tilde{\gamma}_{lm} \in [0, 1]\}_{m=1}^M$ from M different human experts, e.g., workers in a crowdsourcing setting. We interpret $\tilde{\gamma}$ as a degree of belief that the rule holds true over the variables. Our KB is "soft," similar to that in Fold·all (Andrzejewski et al., 2011).

Given the noisy knowledge base KB, we are interested in modeling the reliability of the rules. Previous studies on learning from crowds (Raykar et al., 2010; Welinder et al., 2010) made various assumptions on the experts and tasks. In this paper, for robustness we restrict ourselves to two levels of rule reliability: If $\tilde{\gamma}_{lm}$ is labeled coherently by multiple experts *and* the belief is corroborated by the Bayesian latent variable model, we hypothe-

size that it is reliable and should be incorporated into our Bayesian models; otherwise, we deem the rule unreliable and ignore it. This *knowledge selection process* can be formally characterized by introducing a binary selecting variable $b_l \in \{0, 1\}$ for each rule. We define a "noisy belief likelihood" $p(\tilde{\gamma}_{lm} \mid \gamma_l, b_l)$ as a spike-slab mixture of two components, selected by b_l: If $b_l = 0$, we use a slab distribution to generate diverse beliefs; If $b_l = 1$, we use a spike distribution to generate coherent labels. See Figure 1(a).

Our Robust RegBayes framework is defined as:

$$\min_{q,\boldsymbol{\xi}} \mathrm{KL}\left(q(\mathbf{H}, \boldsymbol{\gamma}, \mathbf{b}) \| p(\mathbf{H}, \boldsymbol{\gamma}, \mathbf{b} \mid \mathbf{X}, \tilde{\boldsymbol{\gamma}})\right) + C \sum_l \xi_l$$

$$\text{s.t. } \mathbb{E}_{q(b_l)}\left[b_l \left| \mathbb{E}_{q(\mathbf{H}|b_l)}\left[\phi_l(\mathbf{H}, \mathbf{X})\right] - \mathbb{E}_{q(\gamma_l|b_l)}\left[\gamma_l\right]\right|\right]$$
$$\leq \epsilon + \xi_l, \ \xi_l \geq 0, \ \forall l = 1 \dots L \quad (2)$$

where $p(\mathbf{H}, \boldsymbol{\gamma}, \mathbf{b} \mid \mathbf{X}, \tilde{\boldsymbol{\gamma}}) \propto p_0(\mathbf{H})p_0(\mathbf{b}, \boldsymbol{\gamma})p(\mathbf{X} \mid \mathbf{H})p(\tilde{\boldsymbol{\gamma}} \mid \boldsymbol{\gamma}, \mathbf{b})$ is the posterior distribution via Bayes' rule. Figure 1(b) shows the graphical model for Robust RegBayes. The prior distribution $p_0(\mathbf{b}, \boldsymbol{\gamma})$ for b_l and γ_l will be discussed in the section of application to LDA. We make two observations. First, if we collapse the model by reducing the uncertainty on $(\boldsymbol{\gamma}, \mathbf{b})$ and holding them constant (i.e., $b_l = 1$ and $\tilde{\gamma}_l = \gamma_l$), Eq (2) reduces to RegBayes Eq (1). In general, Robust RegBayes (2) takes the uncertainty of domain knowledge into consideration and the binary selecting variable b_l specifies the importance of each logic constraint. For unreliable domain knowledge, the corresponding b_l will have a small probability of being 1 and thus the expectation $\mathbb{E}_{q(b_l)}[b_l]$ (i.e., the importance of the logic) will be small. Second, the reliability of rules $(\mathbf{b}, \boldsymbol{\gamma})$ and the underline Bayesian model $(\mathbf{H}, \mathbf{X})$ influence each other in the Robust RegBayes framework. This is represented with the dashed arrow in Figure 1(b). We will see the influence more clearly later in the applications on LDA.

2.3. A Generic Inference Procedure

Since the KL divergence is convex with respect to q (Wainwright & Jordan, 2008) and the posterior constraints are intrinsically linear, problem (2) is convex. Thus, we can apply convex analysis tools to derive a generic solution. Specifically, by introducing a set of dual variables $\boldsymbol{\mu}$ we obtain the optimal distribution:

$$q(\mathbf{H}, \boldsymbol{\gamma}, \mathbf{b} \mid \boldsymbol{\mu}^*) = \frac{p(\mathbf{H}, \boldsymbol{\gamma}, \mathbf{b} \mid \mathbf{X}, \tilde{\boldsymbol{\gamma}})}{Z(\boldsymbol{\mu}^*)} e^{\sum_l \mu_l^* b_l(\phi_l(\mathbf{H}, \mathbf{X}) - \gamma_l)}$$

where $\boldsymbol{\mu}^*$ is the optimum solution of the dual problem:

$$\max_{\boldsymbol{\mu}} L(\boldsymbol{\mu}) = -\log Z(\boldsymbol{\mu}) - \epsilon \sum_l \mu_l \quad (3)$$

$$\text{s.t. } -C \leq \mu_l \leq C,$$

and $Z(\boldsymbol{\mu})$ is the normalization factor for q. Note that μ_l is the weight of logic rule l and the binary variable b_l determines whether the rule affects the posterior distribution of

$\mathbf{H}$ or not. By solving the dual problem (3), we automatically learn the optimal weights $\boldsymbol{\mu}^*$. Then, by inferring b_l we selectively incorporate reliable FOL rules while ignoring unreliable ones.

Despite its elegance, it is important to realize that the generic inference procedure is in general intractable in latent variable models. One needs to utilize variational approximation or sampling techniques to find approximate solutions. In the next section, we present a specific instantiation of Robust RegBayes to LDA models and detail one way to perform efficient variational inference.

3. Application to LDA Models

3.1. Robust RegBayes Applied to LDA

We now give an instantiation of Robust RegBayes in learning LDA topics by incorporating FOL domain knowledge. LDA posits that each document is drawn from an admixture of K topics. Each topic $\boldsymbol{\varphi}_k$ is defined as a multinomial distribution over a given vocabulary and follows a Dirichlet prior $p(\boldsymbol{\varphi}_k \mid \boldsymbol{\beta}) = Dir(\boldsymbol{\varphi}_k \mid \boldsymbol{\beta})$. For document d, we draw a topic proportion $\boldsymbol{\theta}_d$ from a Dirichlet distribution $p(\boldsymbol{\theta}_d \mid \boldsymbol{\alpha}) = Dir(\boldsymbol{\theta}_d \mid \boldsymbol{\alpha})$. For the ith word in document d, we draw a topic assignment z_{di} from the multinomial parametrized by $\boldsymbol{\theta}_d$, $p(z_{di} = k \mid \boldsymbol{\theta}_d) = \theta_{dk}$, and then draw the word w_{di} from the selected topic $\boldsymbol{\varphi}_{z_{di}}$, that is $p(w_{di} \mid z_{di}, \boldsymbol{\varphi}) = \varphi_{z_{di}, w_{di}}$. The joint distribution of LDA is $p(\mathbf{W}, \mathbf{Z}, \boldsymbol{\varphi}, \boldsymbol{\theta} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = \left(\prod_k p(\boldsymbol{\varphi}_k \mid \boldsymbol{\beta})\right)\left(\prod_d p(\boldsymbol{\theta}_d \mid \boldsymbol{\alpha})\prod_i p(z_{di} \mid \boldsymbol{\theta}_d)p(w_{di} \mid z_{di}, \boldsymbol{\varphi})\right)$ where $\mathbf{W} = \{w_{di}\}$ are the observed words, $\mathbf{Z} = \{z_{di}\}, \boldsymbol{\theta} = \{\theta_{dk}\}, \boldsymbol{\varphi} = \{\varphi_{dk}\}$ are the hidden variables. In Bayesian methods, we aim to infer the posterior over hidden variables $p(\mathbf{Z}, \boldsymbol{\varphi}, \boldsymbol{\theta} \mid \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\beta})$.

For domain knowledge, we assume that all the FOL rules are defined over the instantiation of words $\mathbf{W}$ and hidden topic assignments $\mathbf{Z}$. To account for uncertainty in knowledge, we model the belief labels $\tilde{\gamma}_{lm}$ by a spike-slab likelihood (cf. Figure 1(a)), where we define the slab component as $\mathrm{uniform}[0, 1]$ and the spike component as a truncated Gaussian distribution in $[0, 1]$ with the golden standard γ_l as the mean and variance σ_l^2. The variance σ_l^2 is determined by empirical Bayes. The likelihood is then defined as $p(\tilde{\boldsymbol{\gamma}}_l \mid \gamma_l, b_l) = \prod_m \mathcal{N}(\tilde{\gamma}_{lm}; \gamma_l, \sigma_l^2)^{b_l}$. We set non-informative uniform priors for both b_l and γ_l.

With the above definitions, we have $\mathbf{H} = \{\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\varphi}\}$ and $\mathbf{X} = \mathbf{W}$. Plugging these variables to problem (2), we get the optimization problem of learning robust logic LDA:

$$\min_{q,\boldsymbol{\xi}} \mathrm{KL}\left(q(\mathbf{H}, \boldsymbol{\gamma}, \mathbf{b}) \| p(\mathbf{H}, \boldsymbol{\gamma}, \mathbf{b} \mid \mathbf{W}, \tilde{\boldsymbol{\gamma}}, \boldsymbol{\alpha}, \boldsymbol{\beta})\right) + C \sum_l \xi_l$$

$$\text{s.t. } \mathbb{E}_{q(b_l)}\left[b_l \left| \mathbb{E}_{q(\mathbf{Z}|b_l)}\left[\phi_l(\mathbf{Z}, \mathbf{W})\right] - \mathbb{E}_{q(\gamma_l|b_l)}\left[\gamma_l\right]\right|\right]$$
$$\leq \epsilon + \xi_l, \ \xi_l \geq 0, \ \forall l = 1 \dots L.$$

3.2. Variational Approximation

To collapse the parameter space and improve inference accuracy, we first marginalize out the variables φ and θ by exploring the conjugacy between multinomial and Dirichlet in a way similar to (Teh et al., 2007). This marginalization does not affect our logic constraints since they are not directly defined on φ or θ. In theory, we can apply convex analysis tools to derive a closed-form expression of the posterior distribution q as in Section 2.3 and solve the dual problem of the generic form (3) for the dual parameters. Unfortunately, in practice it is intractable from the posterior of logic LDA. Thus we resort to variational approximate methods, as detailed below.

Approximate Inference: Given the dual variables $\boldsymbol{\mu}$, we need to compute the collapsed posterior $q(\mathbf{Z}, \boldsymbol{\gamma}, \mathbf{b} \mid \boldsymbol{\mu})$. This can be done with variational methods. Specifically, we make the mean field assumption that $\tilde{q}(\mathbf{Z}, \boldsymbol{\gamma}, \mathbf{b} \mid \boldsymbol{\mu}) = \prod_d \prod_i \tilde{q}(z_{di} \mid \boldsymbol{\psi}_{di}) \prod_l \tilde{q}(\gamma_l \mid \rho_l) \tilde{q}(b_l \mid \lambda_l)$, where $\tilde{q}(z_{di} \mid \boldsymbol{\psi}_{di})$ is a discrete distribution with parameters $\boldsymbol{\psi}_{di}$; $\tilde{q}(\gamma_l \mid \rho_l)$ is a point-mass function centered on ρ_l; and $\tilde{q}(b_l \mid \lambda_l)$ is a Bernoulli distribution. Then, the best approximation can be found by minimizing the KL-divergence between $\tilde{q}(\mathbf{Z}, \mathbf{b}, \boldsymbol{\gamma} \mid \boldsymbol{\mu})$ and the posterior distribution $q(\mathbf{Z}, \boldsymbol{\gamma}, \mathbf{b} \mid \boldsymbol{\mu})$ with respect to variational parameters. It can be shown that we have the following mean field update equations. For the topic assignment variational parameter ψ, we have:

$$\psi_{di}^k \propto \exp \Big(\mathbb{E}_{\tilde{q}(\mathbf{Z}_{-di})\tilde{q}(\mathbf{b})} \Big[\log(\alpha_k + n_{dk}^{-di})$$
$$+ \log(\beta_{w_{di}} + n_{\cdot k w_{di}}^{-di}) - \log(\sum_v \beta_v + n_{\cdot k \cdot}^{-di})$$
$$+ \sum_l \mu_l b_l \phi_l(\mathbf{Z}, \mathbf{W}) \Big] \Big),$$

where $n_{dkv} \triangleq \sum_i \mathbf{1}(z_{di} = k, w_{di} = v)$ is the number of times that word v is assigned to topic k in document d; the dot denotes summation over that index (e.g., $n_{dk\cdot} = \sum_v n_{dkv}$); and $-di$ denotes that word w_{di} is excluded in calculating the counts. Note that the last term incorporates the FOL logic constraints. Exact computation of the expectations is still very expensive, however. We thus make further approximations. The first three terms in the exponential are the same as in the collapsed variational inference (CVI) algorithm for LDA and we can approximate it effectively by zero-order information (Asuncion et al., 2009). For the last term, we approximate it by using the mode $\hat{\mathbf{Z}}$ of the current distribution $\tilde{q}(\mathbf{Z})$. We get:

$$\psi_{di}^k \propto \frac{\alpha_k + N_{\cdot dk}^{-di}}{\sum_v \beta_v + N_{\cdot k \cdot}^{-di}} (\beta_{w_{di}} + N_{\cdot k w_{di}}^{-di}) e^{\sum_l \mu_l \lambda_l \phi_l(\hat{\mathbf{z}}, \mathbf{W})},$$

$$(4)$$

where $N_{dkv}^{-di} \triangleq \sum_{j \neq i} \mathbf{1}(w_{dj} = v) \psi_{dj}^k$. For the variational parameters $\boldsymbol{\rho}$ and $\boldsymbol{\lambda}$, letting $S(x) \triangleq 1/(1+e^{-x})$ denote the sigmoid function, we have the mean-field update equations (The dual variables $\boldsymbol{\mu}$ are given):

$$\lambda_l = S \Big(M \log(\sqrt{2\pi}\sigma_l) + \frac{\sum_m ((\tilde{\gamma}_{lm})^2 + \sigma_l^2)}{2\sigma_l^2}$$
$$+ \mu_l (\mathbb{E}_{\tilde{q}(\mathbf{Z})}[\phi_l(\mathbf{Z}, \mathbf{W})] - \rho_l) \Big),$$
$$\rho_l = \frac{-\mu_l \lambda_l \sigma_l^2 + \lambda_l (\sum_m \tilde{\gamma}_{lm})}{M \lambda_l}.$$

Due to space limit, we briefly explain the intuition behind λ_l update. It is influenced by both the coherence of belief labels (the first and second terms) and the difference between the current expected feature value and the golden standard for the rule (the third term). Therefore, Robust RegBayes infers the reliability of each rule by considering both noisy labels and the underline Bayesian latent variable model.

Weight Learning: To learn the dual parameters $\boldsymbol{\mu}$ (i.e., the weights of FOL rules), we perform stochastic gradient descent (SGD) to the dual problem (3). Since the exact calculation of the gradient is intractable, we approximate it as follows:

$$\partial_{\mu_l} \log Z(\boldsymbol{\mu}) = \sum_{\mathbf{Z}, \boldsymbol{\gamma}, \mathbf{b}} q(\mathbf{Z}, \boldsymbol{\gamma}, \mathbf{b} \mid \boldsymbol{\mu}) b_l (\phi_l(\mathbf{Z}, \mathbf{W}) - \gamma_l)$$
$$\approx \sum_{\mathbf{Z}, \boldsymbol{\gamma}, \mathbf{b}} \tilde{q}(\mathbf{Z}, \boldsymbol{\gamma}, \mathbf{b} \mid \boldsymbol{\mu}) b_l (\phi_l(\mathbf{Z}, \mathbf{W}) - \gamma_l)$$
$$\approx \mathbb{E}_{\tilde{q}(b_l)}[b_l] (\hat{\phi}_l(\hat{\mathbf{Z}}, \mathbf{W}) - \mathbb{E}_{\tilde{q}(\gamma_l)}[\gamma_l]), \quad (5)$$

where the first equality holds due to duality; the first approximation is due to variational approximation; the second approximation is due to approximating the expectation of the logic rule. Here, we use the mode $\hat{\mathbf{Z}}$ of the variational distribution $\tilde{q}(\mathbf{Z}, \boldsymbol{\gamma}, \mathbf{b} \mid \boldsymbol{\mu})$ which is efficient since $\mathbf{Z}$ is independent under the mean field assumption. Another approximation is made to calculate $\phi_l(\mathbf{Z}, \mathbf{W})$ when the number of groundings is too large — we approximate it by uniformly sampling the groundings for such rules, denoted as $\hat{\phi}_l(\mathbf{Z}, \mathbf{W})$. These approximations work well in practice, as we show below.

With the approximate gradients, we update the weights by the SGD rule:

$$\mu_l^{t+1} = Proj_{[-C,C]}(\mu_l^t + \tau_t(-\partial_{\mu_l} \log Z(\boldsymbol{\mu}) + \epsilon)), \quad (6)$$

where $Proj_{[s,t]}(x)$ denotes the Euclidean projection of x to the interval $[s, t]$; and τ_t is the step length which satisfies mild conditions to ensure convergence (Bottou & Bousquet, 2011). In our implementation we set $\tau_t = (t + \tau_0)^{-\kappa}$ and tune parameters τ_0 and κ.

4. Experiments

We now present empirical results on learning both unsupervised and supervised topic models to demonstrate the efficacy of Robust RegBayes on incorporating noisy FOL do-

Table 1. Datasets. Intuitively, seed rule anchors a specific word to a specific topic, which will attract similar words to the topic; cannot-link rule forces two specific words into different topics; doc seed rule is similar to seed rule, but only applies to specific documents. See (Andrzejewski et al., 2011) for the formal definitions of seed, cannot-link, docseed, inclusion, and exclusion rules.

Dataset	#Documents	#Topics	Description	#FOL Rules
COMP	5,000	20	comp.* in 20 newsgroup data	8 seeds
COM	2,740	25	U.S. House of Representatives	3 seeds, 2 docseeds
POL	2,000	20	movie reviews	1 cannot-link
HDG	24,073	50	PubMed abstracts	8 seeds, 6 inclusion, 6 exclusion

main knowledge. In short, Robust RegBayes shows superior ability to discover latent semantic structures and make accurate predictions in the supervised settings.

4.1. Experiments with Unsupervised Topic Models

We denote our Robust RegBayes applied to LDA as "RLogicLDA." A special case of RLogicLDA is to set $b_l = 1$ for all rules in Eq. (2), i.e., do not allow the model to ignore any rules via the slab component. Equivalently, the special case treats all rules as valid. We denote the special case as "LogicLDA." For baselines, we use (*i*) the vanilla LDA using Gibbs sampling, and (*ii*) Fold·all, a MAP estimator to incorporate FOL into LDA. Fold·all requires experts to manually set the weights of rules. We adopt the well-performing "Mir" method for Fold·all and download the authors' implementation (Andrzejewski et al., 2011).

4.1.1. LOGICLDA VS. LDA AND FOLD·ALL

We first show that LogicLDA achieves similar performance as Fold·all (which has the benefit of expert-set rule weights) by automatically learning rule weights. We use all four real datasets in (Andrzejewski et al., 2011) and the same logic rules. The logic rules contain "seed rules" which assign specific words to specific topics, "cannot-link rules" which force two words into separate topics, and so on. Details are in Table 1. Since no belief labels $\tilde{\gamma}_l$ were provided with their data, we define them by examining the meaning of the logic rules: All the rules in COMP, CON and POL aim to make the learned topics more understandable for humans, we set all the belief labels of these rules at $\tilde{\gamma}_l = 1$. For HDG, the rules are given by biological experts and should be satisfied according to the description. Thus, we also set their belief labels at 1. As in (Andrzejewski et al., 2011), we randomly split documents into training/testing sets by a ratio of 8/2. For LogicLDA, we utilize the FOL rules during training and estimate the topics φ from the posterior distribution $\tilde{q}(\mathbf{Z})$ as in (Asuncion et al., 2009). As in (Andrzejewski et al., 2011), the knowledge is assumed to be encoded into the estimated topics. Therefore, for testing, we do not utilize the logic rules and only optimize the variational bound given φ as in vanilla LDA (Blei et al., 2003).

We measure *test set perplexity* to evaluate LDA perfor-

mance (Blei et al., 2003). To show different methods' ability in incorporating logic rules, we also measure the *proportion of satisfied logic rules*. For fairness, all parameters are the same as in (Andrzejewski et al., 2011) across all the methods in comparison, e.g. we use the same symmetric Dirichlet parameters $\boldsymbol{\alpha} = \mathbf{1}$, $\boldsymbol{\beta} = \mathbf{1}$ below. For the extra parameters in our methods, we simply set $\epsilon = 0.001$ and the regularization parameter C at a large number (e.g., 1000000) so that the dual parameters $\boldsymbol{\mu}$ never reach the bounds in Eq (3). The SGD step length decays as $\tau_t = (t + 10)^{-0.5}$ by cross validation on the training data.

We run each method five times under random initialization and report the average results in Table 2. LogicLDA is superior to LDA and Fold·all by both measures: First, LogicLDA achieves the lowest test set perplexity in three out of four data sets. These differences are statistically significant under 2-tailed paired t-test with significance level $p < 0.02$. In addition, on the CON data set LogicLDA is not significantly different than the best (LDA).

Second, LogicLDA and Fold·all both achieve much higher proportion of FOL rule satisfaction than LDA (except for the POL data set, where all models achieve near 100% satisfaction). Importantly, LogicLDA does so by automatically learning the rule weights, while Fold·all has to rely on human experts to specify the weights.

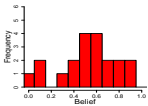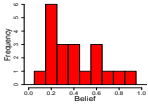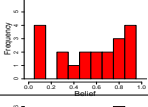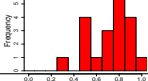#### 4.1.2. RLOGICLDA VS. LOGICLDA: ROBUSTNESS

We examine the robustness of RLogicLDA by comparing it with LogicLDA under the same settings as above, but with potentially unreliable domain knowledge. To this end, we intentionally design one potentially unreliable FOL rule for each of the four datasets, see Table 3. We show each designed rule to $M = 20$ volunteers and collected their subjective belief label $\tilde{\gamma}_{lm}$ on that rule. Specifically, each volunteer can select their $\tilde{\gamma}_{lm}$ between 0 and 1 with step size 0.1 via a user interface. Table 3 shows the histogram of $\tilde{\gamma}_{lm}$: a flat histogram indicates disagreements among the volunteers and thus unreliable rule.

RLogicLDA performs better than LogicLDA in test set perplexity, as shown in Table 3. On COMP and HDG data sets, the difference is statistically significant under 2-tailed paired t-test ($p < 0.02$) while on CON and POL the dif-

Table 2. Test set perplexity and proportion of satisfied logic rules on four datasets.

	Test Set Perplexity			Proportion of Satisfied Logic Rules		
	LDA	Fold·all	LogicLDA	LDA	Fold·all	LogicLDA
COMP	1531 ± 12	1537 ± 11	$\mathbf{1463} \pm 5$	0.00 ± 0.00	1.00 ± 0.00	0.97 ± 0.01
CON	$\mathbf{1206} \pm 6$	1535 ± 10	$\mathbf{1216} \pm 11$	0.07 ± 0.04	0.67 ± 0.03	0.70 ± 0.00
POL	3218 ± 13	3220 ± 13	$\mathbf{3176} \pm 12$	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
HDG	940 ± 6	973 ± 7	$\mathbf{885} \pm 2$	0.60 ± 0.01	0.95 ± 0.00	0.96 ± 0.01

Table 3. RLogicLDA is robust to unreliable domain knowledge

Data	Designed Rule	Histogram	Test Set Perplexity		Satisfaction Proportion	
			LogicLDA	RLogicLDA	LogicLDA	RLogicLDA
COMP	seed: {*problem, windows, window, available, files, mac, apple, system, im*} → topic 1		1467 ± 6	$\mathbf{1446} \pm 6$	0.39 ± 0.16	0.07 ± 0.06
CON	seed: {*bill, people, law, health, tax, trade, economy, budget, pension*} → topic 7		$\mathbf{1228} \pm 9$	$\mathbf{1228} \pm 16$	0.49 ± 0.03	0.08 ± 0.03
POL	must-link{*acting, make, performance, character*}: same topic		$\mathbf{3173} \pm 13$	$\mathbf{3168} \pm 11$	0.57 ± 0.19	0.09 ± 0.04
HDG	cannot-link{*human, gene*}: different topics		895 ± 2	$\mathbf{891} \pm 2$	0.75 ± 0.03	0.95 ± 0.01

ference is not significant. It achieves this by only listening to reliable rules. The empirical means of the belief labels for the four rules are 0.50, 0.40, 0.52 and 0.72 respectively. The satisfaction proportions of LogicLDA are close to these empirical means – it indiscriminately obeys all domain knowledge. In contrast, RLogicLDA is able to ignore the first three rules it deems unreliable, while obeying the fourth rule. This is reflected in RLogicLDA's proportions.

4.2. Experiments with Supervised Topic Models

We now show that robustly incorporating knowledge can help achieve both higher prediction performance and better interpretability of learned topics compared to other supervised LDA methods.
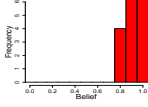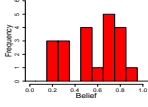
4.2.1. SETTINGS AND DOMAIN KNOWLEDGE

We use the HotelReview dataset (Zhu & Xing, 2010) and predict the global rating (an integer from 1 to 5) of each hotel review based on its content. As in (Zhu & Xing, 2010), we treat it as a regression problem and normalize the ratings. The dataset contains 5,000 reviews and is equally split into training and testing sets. Besides the global rating, each review also has the ratings of five aspects: *value*, *location*, *service*, *room*, and *cleanliness*. Discovering the

latent correspondence between review contents and aspects is an interesting research topic (Wang et al., 2010). Here, we use seed rules to assign several representative words of each aspect to a specific set of topics. Specifically, we assign words {*value*, *price*, *quality*, *worth*, *resort*} to topics 1 and 2 to seed the *value* aspect; {*location*, *traffic*, *restaurant*, *beach*} to topic 3 to seed the *location* aspect; {*service*, *food*, *breakfast*, *dinner*} to topics 4–6 to seed the *service* aspect; and {*door*, *floor*, *bed*, *stay*, *bathroom*, *room*} to topics 7–10 to seed the *room* aspect. We ignore the *cleanliness* aspect because we find reviews on cleanliness usually are contained within the reviews on the *room* aspect and thus redundant. Furthermore, to distinguish positive and negative aspects, we use a "sentiment seed rule" to assign 19 seed positive words to topics 1,3,5,7,9.[1]

Note that these rules represent our intention to relate topics and aspects. Therefore, we set all belief labels for the five rules to 1.0. Finally, as in Section 4.1.2, we also collect empirical belief labels from $M = 20$ volunteers for one reliable rule (the "Not rule") and one unreliable rule (the "But rule"), see Table 4.

---

[1] The 19 seed words are *amazing*, *beach*, *beautiful*, *comfortable*, *enjoyed*, *excellent*, *fantastic*, *fresh*, *friendly*, *good*, *great*, *large*, *lovely*, *nice*, *perfect*, *wonderful*, *best*, *recommend* and *enjoy*.

Table 4. Intentionally Designed Reliable and Unreliable Rules

Rule	Description	Histogram	mean($\tilde{\gamma}_{lm}$)	$p(b_l = 1 \mid \lambda_l)$	Satisfaction Proportion	
					sLogicLDA	sRLogicLDA
Not rule	seed: {adjectives with negation within distance 4 before it} → the last topic		0.91	0.99	0.98 ± 0.03	1.00 ± 0.00
But rule	seed: {all words before adversative transition (e.g. "but") in sentences} → the last topic		0.56	0.00	0.70 ± 0.13	0.05 ± 0.4

4.2.2. PREDICTION PERFORMANCE

We build our supervised RLogicLDA ("sRLogicLDA") by adding the same max-margin posterior constraints as in (Zhu et al., 2013a) to RLogicLDA. The parameter settings of ϵ, C and α are the same as in the unsupervised experiments. The other parameters (α, β) and the regularization parameter introduced by max-margin constraints (Zhu et al., 2013a) are set by cross-validation on the training set. For baselines, we compare with (i) maximum entropy discrimination LDA regression (MedLDAr) (Zhu et al., 2013a), a RegBayes model that incorporates max-margin posterior regularization into LDA; (ii) supervised conditional topical random fields (sCTRF) (Zhu & Xing, 2010), a feature based model that incorporates both single and pairwise word features into MedLDAr.

We run each algorithm five times with random initialization and random split, and report the average test set results in Figure 2(a). Note with our setting sRLogicLDA requires at least 10 topics to accommodate for the FOL rules, thus its curve starts at #topics=10; while the baselines have no logic rules and they start at #topic=3. We use predictive R^2 (Zhu & Xing, 2010) as the performance measure of regression. sRLogicLDA achieves comparable performance as feature based sCTRF, and outperforms MedLDAr. Note that sC-TRF uses 15 features on words,[2] while sRLogicLDA only needs 7 simple logic rules. Therefore, incorporating domain knowledge as constraints is useful for prediction compared with feature engineering approaches.

4.2.3. TOPIC INTERPRETABILITY

Tables 5,6 show the top 10 words of each topic learned by sRLogicLDA and sCTRF with $K = 15$ topics.[3] We manu-

---

[2]The sCTRF features are: 9 Part-of-Speech features that categorize the words, 5 WordNet sentiment features, and 1 feature on whether two words belong to the same phrase.

[3]We observed similar phenomenon with other K. We did not include the topics learned by MedLDAr for two reasons: first, as Figure 2(a) shows MedLDAr has inferior predictive performance compared to sCTRF and sRLogicLDA. Second, it was shown in (Zhu & Xing, 2010) that MedLDAr produces less interpretable

ally judged which words represent the *value*, *location*, *service* and *room* aspects, respectively, and colored them orange, blue, cyan and red, respectively. When applicable, we mark FOL seed words with an ∗.

sRLogicLDA has a clear correspondence between topics and aspects due to the FOL rules. Topics T1–T10 obey the grouping into the four aspects (denoted by vertical lines in Table 5). The only exception is T7, which we suspect is because the other three topics T8–T10 are sufficient in describing the *room* aspect. We also note that sRLogicLDA is successful in attracting non FOL seeded, but aspect-related, words into the topics (i.e., those colored words not marked by an ∗ in Table 5). In contrast, such a clear correspondence is largely absent in sCTRF (Table 6). Its topics contain a mix of *room*, *location*, *service* aspects, and the *value* aspect is missing among the top topic words.

Finally, we study sRLogicLDA's ability to utilize the sentiment seed rule to attract additional positive words into specific topics. The set of positive topics, denoted as $T_p = \{T1, T3, T5, T7, T9\}$, is defined as the topics specified by the sentiment seed rule. The set of other topics is denoted as T_o. We hope to see that T_p attracts many more positive sentiment words (excluding the 19 seed words which the rule forces into T_p anyway), and that T_o attracts fewer positive sentiment words (including the 19 seed words since any positive words in T_o is undesirable). To this end, we first obtain a commonly-used positive word list W containing 2006 positive words.[4] Note W includes the 19 seed words. We measure the amount of positive words in T_o by the average weights of words in W over these topics: $A_o = \frac{\sum_{k \in T_o} \sum_{w \in W} \varphi_{kw}}{|T_o|}$. Let $W_{\setminus 19}$ be the set W excluding the 19 seed words. We measure the amount of positive words in T_p by $A_p = \frac{\sum_{k \in T_p} \sum_{w \in W_{\setminus 19}} \varphi_{kw}}{|T_p|}$. Our hypothesis is that, with the sentiment seed rule, $A_p \geq A_o$. Note that because of the exclusion of seed words from the computation of A_p, this hypothesis is a very strict comparison.

---

topics than sCTRF on the same data.

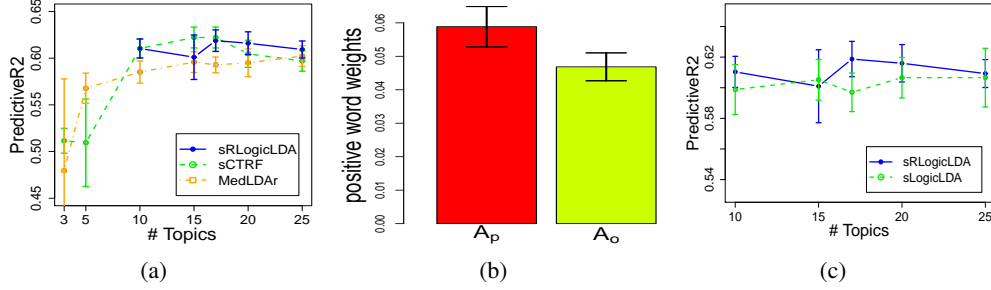[4]http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar.

Figure 2. (a) Predictive R^2 of sRLogicLDA, sCTRF, and MedLDAr. (b) average weights of positive words in the positive topic set (A_p) and the other topic set (A_o); and (c) predictive R^2 of sRLogicLDA and sLogicLDA.

Table 5. Top 10 words in Sampled Topics learned by sRLogicLDA

T1+	T2	T3+	T4	T5+	T6	T7+	T8	T9+	T10	T11	T12	T13	T14	T15
resort	n't	*beach	restaurant	pool	*breakfast	but	*room	*room	hotel	staff	pool	but	but	n't
free	pay	*location	fruit	good	*food	n't	told	*bed	*room	but	view	reception	hotel	night
*price	but	*dinner	*location	holiday	*service	kids	asked	*bathroom	rooms	good	area	small	staff	looked
great	money	nice	wine	bar	but	people	desk	shower	*stay	guests	balcony	area	people	smell
*worth	check	street	served	entertainment	day	time	front	*door	hotels	time	small	however	day	work
island	time	parking	morning	day	water	nice	manager	*floor	night	however	chairs	road	time	air
trip	back	area	menu	*food	bar	night	*stay	colorred*stay	booked	bit	spa	car	place	left
beautiful	car	good	evening	euros	buffet	great	called	bedroom	*floor	reviews	lounge	park	n't	dirty
*quality	expensive	*restaurant	meal	lovely	drinks	day	call	coffee	city	bar	pools	tv	back	carpet
place	lobby	internet	eggs	evening	lunch	family	back	towels	view	found	ocean	side	night	back
		great												

Table 6. Topics learned by sCTRF in a randomly selected run

T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15
n't	room	room	room	room	room	hotel	hotel	hotel	hotel	hotel	pool	beach	beach	great
poor	n't	n't	n't	n't	hotel	room	room	room	pool	pool	hotel	pool	resort	lovely
dirty	told	told	hotel	hotel	n't	n't	n't	day	day	day	food	food	great	good
bad	asked	hotel	stay	stay	stay	night	breakfast	staff	area	food	good	good	pool	beautiful
room	hotel	back	front	night	night	stay	staff	area	staff	good	beach	resort	good	excellent
hotel	back	front	desk	rooms	rooms	rooms	day	breakfast	rooms	bar	bar	bar	food	beach
worst	manager	desk	back	back	time	breakfast	night	pool	food	area	day	great	island	wonderful
back	stay	stay	night	bed	staff	staff	time	time	time	staff	nice	day	day	nice
small	called	asked	rooms	front	bed	time	time	n't	breakfast	beach	restaurant	hotel	nice	fantastic
awful	night	manager	door	time	breakfast	day	area	night	good	restaurant	staff	nice	ocean	amazing

Fig 2(b) presents the average A_p and A_o from five randomized run. A_p is indeed statistically significantly larger than A_o (2-tailed unpaired t-test with $p < 0.02$). Therefore, the sentiment seed rule attracts more positive words to T_p.

Taken together, these results demonstrate that by incorporating FOL rules on aspect-topic relation, sRLogicLDA learns topics with improved interpretability.

4.2.4. ROBUSTNESS

We examine sRLogicLDA's ability to automatically infer robustness of FOL rules by comparing it with one variant: sLogicLDA—a special case of sRLogicLDA where all b_l are set to 1 (i.e., forced to use all FOL rules with no attempt to infer their robustness).

First, Table 4 shows that sLogicLDA simply matches satisfaction proportions to the empirical mean of belief labels, while sRLogicLDA is more sophisticated and achieves a quite different proportion on the unreliable "But rule." This demonstrates that sRLogicLDA can select the reliable "Not rule" and ignore the unreliable "But rule." Second, Fig-

ure 2(c) shows that sRLogicLDA outperforms sLogicLDA in predictive R^2, suggesting that automatically inferring the robustness of knowledge achieves better performance.

5. Conclusions

We proposed Robust RegBayes, a framework to selectively incorporate noisy FOL domain knowledge into Bayesian models via posterior regularization. We applied our framework to unsupervised and supervised topic models, and demonstrated that through incorporating domain knowledge robustly, we can improve both the predictive performance and topic interpretability. In the future, we plan to extend Robust RegBayes to incorporate FOL domain knowledge into Bayesian nonparametric models.

Acknowledgments

References

Andrzejewski, D., Zhu, X., Craven, M., and Recht, B. A framework for incorporating general domain knowledge into latent Dirichlet allocation using first-order logic. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2011.

Asuncion, A., Welling, M., Smyth, P., and Teh, Y.W. On smoothing and inference for topic models. In *Uncertainty in Artificial Intelligence (UAI)*, 2009.

Blei, D.M., Ng, A.Y., and Jordan, M.I. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3: 993–1022, 2003.

Bottou, L. and Bousquet, O. The tradeoffs of large-scale learning. *Optimization for Machine Learning*, pp. 351, 2011.

Garthwaite, P., Kadane, J., and O'Hagan, A. Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100(470):680–700, 2005.

Mao, Y. and Lebanon, G. Domain knowledge uncertainty and probabilistic parameter constraints. In *Uncertainty in Artificial Intelligence (UAI)*, 2009.

Raykar, V., Yu, S., Zhao, L., Valadez, G.H., Florin, C., Bogoni, L., and Moy, L. Learning from crowds. *Journal of Machine Learning Research*, 11:1297–1322, 2010.

Richardson, M. and Domingos, P. Markov logic networks. *Machine Learning*, 62(1-2):107–136, 2006.

Teh, Y.W., Newman, D., and Welling, M. A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. *Advances in Neural Information Processing Systems (NIPS)*, 19:1353, 2007.

Wainwright, M.J. and Jordan, M.I. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

Wang, H., Lu, Y., and Zhai, C. Latent aspect rating analysis on review text data: a rating regression approach. In *International Conference on Knowledge Discovery and Data mining (SIGKDD)*, 2010.

Welinder, P., Branson, S., Belongie, S., and Perona, P. The multidimensional wisdom of crowds. In *Advances in Neural Information Processing Systems (NIPS)*, 2010.

Zhu, J. Max-margin nonparametric latent feature models for link prediction. In *International Conference on Machine Learning (ICML)*, 2012.

Zhu, J. and Xing, E.P. Conditional topic random fields. In *International Conference on Machine Learning (ICML)*, 2010.

Zhu, J., Chen, N., and Xing, E.P. Infinite latent SVM for classification and multi-task learning. *Advances in Neural Information Processing Systems (NIPS)*, 25, 2011.

Zhu, J., Chen, N., Perkins, H., and Zhang, B. Gibbs max-margin topic models with fast sampling algorithms. In *International Conference on Machine Learning (ICML)*, 2013a.

Zhu, J., Chen, N., and Xing, E.P. Bayesian inference with posterior regularization and applications to infinite latent SVMs. *arXiv:1210.1766v2*, 2013b.