Learning Relational Kalman Filtering

Jaesik Choi
School of Electrical and Computer Engineering
Ulsan National Institute of Science and Technology
Ulsan, Korea

Eyal Amir
Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL, USA

Tianfang Xu and **Albert J. Valocchi**
Department of Civil and Environmental Engineering
University of Illinois at Urbana-Champaign
Urbana, IL, USA

Abstract

The Kalman Filter (KF) is pervasively used to control a vast array of consumer, health and defense products. By grouping sets of symmetric state variables, the Relational Kalman Filter (RKF) enables us to scale the exact KF for large-scale dynamic systems. In this paper, we provide a parameter learning algorithm for RKF, and a regrouping algorithm that prevents the degeneration of the relational structure for efficient filtering. The proposed algorithms significantly expand the applicability of the RKFs by solving the following questions: (1) how to learn parameters for RKF from partial observations; and (2) how to regroup the degenerated state variables by noisy real-world observations. To our knowledge, this is the first paper on learning parameters in relational continuous probabilistic models. We show that our new algorithms significantly improve the accuracy and the efficiency of filtering large-scale dynamic systems.

1 Introduction

Many real-world systems can be modeled by continuous variables and relationships (or dependencies) among them. The Kalman Filter (KF) (Kalman 1960) accurately estimates the state of variables in a linear dynamic system with Gaussian noise given a sequence of control-inputs and observations. The KF has been applied in a broad range of domains such as robotics, finance (Bahmani-Oskooee and Brown 2004), and environmental science (P. and Bierkens 2001; Clark et al. 2008). Given a sequence of observations and linear dependencies with Gaussian noise between variables, the KF calculates the conditional probability density of the state variables at each time step.

Unfortunately, the KF computations are cubic in the number of state variables, which limits the use of existing exact methods for domains with a large number of state variables. This has led to the combination of approximation and sampling in the Ensemble Kalman Filter (Evensen 1994), and recently to the Relational Kalman Filters (RKFs) over grouped state variables (Choi, Guzman-Rivera, and Amir 2011; Ahmadi, Kersting, and Sanner 2011). The RKFs leverage the ability of relational languages to specify models with size of representation independent of the size of populations involved (Friedman et al. 1999; Poole 2003; Richardson and Domingos 2006; Kersting 2012).

Lifted inference algorithms for relational continuous models (Wang and Domingos 2008; Choi, Hill, and Amir 2010; Ahmadi, Kersting, and Sanner 2011; Choi and Amir 2012) degenerate (or split) relational structures upon individual observations. Lifted RKF (Choi, Guzman-Rivera, and Amir 2011) maintains relational structure when the same number of observations are made. Otherwise, it also degenerates (possibly rapidly) the relational structure, thus lifted RKF may not be useful with sparse observations.

The main contributions of this paper are (1) to learn parameters for RKFs and (2) to regroup the degenerated state variables from noisy real-world observations with tight error bounds. To our knowledge, this is the first paper on learning parameters in relational continuous probabilistic models.

We propose a new learning algorithm for RKFs. We show that relational learning expedites filtering, and achieves accurate prediction in theory and practice. The key intuition is that the Maximum Likelihood Estimate (MLE) of RKF parameters is the empirical mean and variance over state variables of a group. For partial observations, the parameters can be calculated similarly. We show that variances of degenerated state variables on partial observations converge exponentially under reasonable conditions. Thus, our approximate regrouping algorithm has bounded errors compared to the exact KF. We show that the RKF with regrouping is more robust against degeneracy than the Lifted RKF in practice with partial observations.

2 Relational Linear Dynamic Systems

In this section, we define relational linear dynamic systems. Dependencies among variables are represented using **relational atoms**, or just **atoms**.[1] The relational atoms are useful when the joint probability of variables involves common types of functions. When representing the joint probability distribution, there are products of the parameterized functions (or potentials).

Relational atoms represent the set of state variables corresponding to all ground substitutions of its parameter variables. For example, let $X^{r_1}(\mathit{Latitude}, \mathit{Longitude})$ be an atom

[1] For comprehensive definitions, see (Poole 2003; de Salvo Braz, Amir, and Roth 2005).

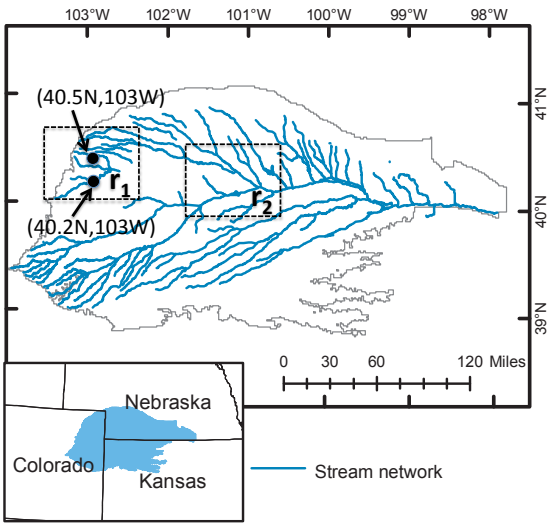

Figure 1: The Republican River Basin covering portions of east Colorado, northwest Kansas, and southwest Nebraska. This figure shows two clustered water wells; $region_1(r_1)$ and $region_2(r_2)$. Water wells in each region have the same (linear Gaussian) relationships with wells in other regions.

for the (water level of) wells in $region_1$, θ=$(40.2N,103W)$. When we substitute *Latitude* and *Longitude* with θ, the atom becomes a state variable $X^{r_1}(40.2N, 103W)$ which represents the level (or prediction) of well head at (*Latitude*=$40.2N$, *Longitude*=$103W$). Formally, applying a substitution θ to an atom $\mathbf{X}(L)$ yields a new atom $\mathbf{X}(L\theta)$ where $L\theta$ is obtained by renaming the parameter variables in L according to θ. If θ is a ground substitution, $\mathbf{X}(L\theta)$ is a ground state variable like $X^{r_1}(40.2N, 103W)$.[2] $|\mathbf{X}(L)|$ or just $|\mathbf{X}|$ denotes the the number of distinct state variables generated from $\mathbf{X}$ by all substitutions.

A pairwise Gaussian parfactor $((\mathbf{X}, \mathbf{X}'), \phi)$ is composed of a pair of two atoms ($\mathbf{X}$, $\mathbf{X}'$) and a linear Gaussian potential ϕ between two atoms in the following form,

$$\phi(\mathbf{X}, \mathbf{X}') \propto \exp\left[-\frac{(\mathbf{X} - \mathbf{X}' - \mu)^2}{\sigma^2}\right].$$

For example, a pairwise Gaussian parfactor $\phi_{r_1,r_2}(X_{r_1}, X_{r_2})$ represents the linear Gaussian relationship between two ground variables chosen from $region_1$ and $region_2$ respectively.

A pairwise Gaussian factor, or just a **factor**, $f = ((\mathbf{x}, \mathbf{x}'), \phi)$ is a pair where ϕ is a potential function on $(\mathbf{x}, \mathbf{x}')$ from $\mathbb{R}^2$ to $\mathbb{R}^+$ where $(\mathbf{x}, \mathbf{x}')$ is a pair of ground random variables derived by ground substitutions from $(\mathbf{X}(L\theta), \mathbf{X}'(L'\theta'))$. A factor f defines a weighting function on a **valuation** $(x, x') = (v, v')$: $w_f(x, x') = \phi(v, v')$. The weighting function for a *parfactor* g is the product of the weighting functions over all of its ground substitutions (factors), $w_g(v) = \prod_{f \in g} w_f(v)$. Hence, a set of parfactors G defines a probability density,

$$w_G(v) = \frac{1}{Z} \prod_{g \in G} \prod_{f \in g} w_f(v),$$

where Z is the normalizing constant and $f \in g$ means f is a ground instance of g.[3] In this way, we can represent the joint probability of all random variables (e.g., all wells in $region_1$ and $region_2$).

Relational Transition Models (RTMs) characterize the dependencies of relational atoms between consecutive time steps. $X_t^i(a)$ and $X_{t+1}^j(a')$ are relational atoms at time step t and $t+1$ respectively when a and a' are ground substitutions, e.g., a=$(40.2N,98W)$,a'=$(40.5N,98W)$. $U_t^i(a)$ is the control-input information. A RTM takes the following form,

$$X_{t+1}^j(a') = B_X^{ij} X_t^i(a) + B_U^{ij} U_t^i(a) + G_{RTM}^{ij}, \quad (1)$$

where $G_{RTM}^{ij} \sim \mathcal{N}(0, {\sigma_{RTM}^{ij}}^2)$ and $\mathcal{N}(m, \sigma^2)$ is the normal distribution with mean m and variance σ^2. B_X^{ij} and B_U^{ij} are the linear transition coefficients.

In the linear Gaussian representation, the transition models take the following form,

$$\phi_{RTM}(X_{t+1}^j(a') | X_t^i(a), U_t^i(a)) \propto \exp\left[-\frac{(X_{t+1}^j(a') - B_X^{ij} X_t^i(a) - B_U^{ij} U_t^i(a))^2}{2{\sigma_{RTM}^{ij}}^2}\right]. \quad (2)$$

The most common transition is the one from the state $X_t^i(a)$ to the state itself $X_{t+1}^i(a)$ at the next time step,

$$X_{t+1}^i(a) = B_X^i X_t^i(a) + B_U^i U_t^i(a) + G_{RTM}^i. \quad (3)$$

Relational Observation Models (ROMs) represent the relationships between the hidden (state) variables, $X_t^i(a)$. The observations can be made on the directly related variable, $O_t^i(a)$ (**direct observations or individual observations**),

$$O_t^i(a) = C_X^i X_t^i(a) + G_{ROM}^i, G_{ROM}^i \sim \mathcal{N}(0, {\sigma_{ROM}^i}^2) \quad (4)$$

C_X^i is the linear coefficient.

ROMs also represent the relationships between the hidden variables $X_t^i(a)$ and the observations made indirectly on another variable in the atom $O_t^i(a')$ where $a \neq a'$ (**relational observations**),

$$O_t^i(a') = C_t'^i X_t^i(a) + G_{ROM}'^i, G_{ROM}'^i \sim \mathcal{N}(0, {\sigma_{ROM}'^i}^2). \quad (5)$$

In most cases, it is reasonable to assign the variance of the direction observation $var(G_{ROM}^i)$ to be smaller value than the variance of relational one $var(G_{ROM}'^i)$ i.e. $\sigma_{ROM}^i \ll \sigma_{ROM}'^i$.

For the well example, $X_t^{r_1}(\mathbf{40.2N}, 103W)$ will have a smaller variance (more certain), when an observation is

[2] Here, we assume that the ground state variables are univariate, e.g., domain of x is $\mathbb{R}$. Models with multivariate ground variables can be handled similarly.

[3] The condition of being a probability density is that at least a random variable has a prior distribution, see (Choi, Hill, and Amir 2010).

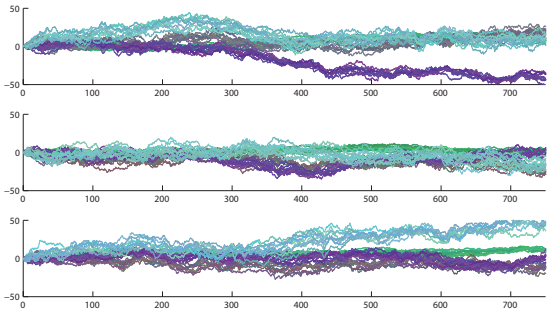

Figure 2: Samples generated from Relational Linear Models each with randomly generated parameters. Here, each plot includes 40 variables represented by four colored atoms with 10 variables each. The x axis is the time step.

made at $O_t^{r_1}(\mathbf{40.2N}, 103W)$ than made at a nearby location $O_t^{r_1}(\mathbf{40.5N}, 103W)$. $O_t^{r_1}(\mathbf{40.2N}, 103W)$ is a direct observation for $X_t^{r_1}(\mathbf{40.2N}, 103W)$ and a relational observation for $X_t^{r_1}(\mathbf{40.5N}, 103W)$. Thus, after an update step, the variance of $X_t^{r_1}(\mathbf{40.2N}, 103W)$ will be significantly reduced compared to the variance of $X_t^{r_1}(\mathbf{40.5N}, 103W)$.

In the linear Gaussian representation, ROMs take the following form,

$$\phi_{ROM}(O_t^i(a)|X_t^i(a)) \propto \exp\left[-\frac{(O_t^i(a) - C_X^i X_t^i(a))^2}{2{\sigma_{ROM}^i}^2}\right].$$

Relational Pairwise Models (RPMs) represent linear dependencies between pairs of relational atoms,

$$X_t^i(a) = R_t^{ij} X_t^j(a') + G_{RPM}^{ij}, G_{RPM}^{ij} \sim \mathcal{N}(0, \sigma_{RPM}^{ij}), \quad (6)$$

where R_t^{ij} is the coefficient.

Note that RTMs and ROMs represent the nature of dynamic systems (e.g. the state at the next time step depends on the current time step). A set of RPMs over multiple atoms is an efficient way to represent the relational structure over a large number of state variables as shown in Figure 2.

Relational Kalman Filter (RKF) is a filtering procedure with a relational linear dynamic system which is composed of RTMs, ROMs and RPMs. That is, the joint probability of state variables is represented by the product of pairwise Gaussian parfactors. **Lifted RKF** computes the posterior of the state variables given a prior (current) belief and full or partial observations. The input to the problem is: (1) relational parfactors (RTMs, ROMs and RPMs); (2) a current belief over atoms (X_0^i); (3) a sequence of control-inputs ($U_1^i, \ldots, U_T^i$); and (4) a sequence of observations ($O_1^i, \ldots, O_T^i$). The output is the multivariate Gaussian distribution over the atoms (X_T^i) at each time step T. The filtering problem is solved by algorithms presented in (Choi, Guzman-Rivera, and Amir 2011; Ahmadi, Kersting, and Sanner 2011). In this paper, we focus on the parameter learning problem.

3 Learning Relational Kalman Filter

The two important parameters of the RKF are the transition models and observation models. In this section, we present a learning algorithm that derives the MLEs of RTMs and ROMs. For simplicity, we will present a solution with fully observed model. A solution for partial observations can be derived with a slight modification.

3.1 Algorithm *LearningRKF*

Algorithm *LearningRKF* estimates the parameter of RKF given a sequence of observations such as measurements of water wells for several years. The overall procedure is similar to parameter learning for the ground KF. Here, the main difference is that the coefficients and covariances of RTMs and ROMs are the block matrices. A subroutine, *BlockAverage*, computes the averages of the diagonal and non-diagonal entries of an input matrix, and then ouputs a block matrix where each block includes the empirical means, variances and covariances in each block. This is essentially parameter tying (Raedt 2008) step for RKF. In the following sections, we will show that the block matrix computed by *BlockAverage* is the MLE.

Algorithm 1 LearningRKF

input: a sequence of obs $(O_1, \cdots, O_T)$
$(\mathbf{B}, \Sigma_T, \mathbf{C}, \Sigma_O) \leftarrow (I, I, I, I)$
currentLL $\leftarrow \infty$
repeat
  prevLL $\leftarrow$ currentLL
  $\left(\widehat{\mathbf{B}}, \widehat{\Sigma}_T, \widehat{\mathbf{C}}, \widehat{\Sigma}_O\right) \leftarrow$ LearnGroundTM(O_t,$\mathbf{B}$,Σ_T,$\mathbf{C}$,Σ_O)
  $(\mathbf{B}, \Sigma_T, \mathbf{C}, \Sigma_O) \leftarrow$ *BlockAverage*($\widehat{\mathbf{B}}, \widehat{\Sigma}_T, \widehat{\mathbf{C}}, \widehat{\Sigma}_O$)
  currentLL $\leftarrow \sum_t \log P(O_t|X_t, \mathbf{B}, \Sigma_T, \mathbf{C}, \Sigma_O)$
until | prevLL - currentLL | $< \epsilon$
output: estimated parameters $(\mathbf{B}, \Sigma_T, \mathbf{C}, \Sigma_O)$

3.2 Learning Transition Models

Here, we derive the parameter of the RTMs: linear coefficient B and Gaussian noise G_{RTM}. It has been shown that a relational linear dynamic model with RTMs, ROMs and RPMs can be converted into a linear multivariate models with block coefficient and covariance matrices (Choi, Hill, and Amir 2010). Thus, given data, we find the block coefficient and covariance matrices of RTMs.

Learning Transition Noise means to compute the mean and the covariance matrix in the following block forms,

$$\mu_T = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}, \Sigma_T = \begin{bmatrix} \Sigma^{1,1} & \Sigma^{1,2} & \cdots & \Sigma^{1,n} \\ \Sigma^{2,1} & \Sigma^{2,2} & \cdots & \Sigma^{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma^{n,1} & \Sigma^{n,2} & \cdots & \Sigma^{n,n} \end{bmatrix}, \quad (7)$$

where μ_i is a vector of size $|X^i|$ (=n_i); $\Sigma^{i,j}$ is a matrix of size n_i by n_j.

Given a prior, a linear coefficient $\mathbf{B}$ and a sequence of full observations, we derive the estimate X_t at time step t

assuming Gaussian noise in the transition model. The MLE estimation of μ and Σ for the RTM can be derived:

$$(\mu_{Tmax}, \Sigma_{Tmax}) = \arg\max_{\mu_T, \Sigma_T} \sum_{t=2,\cdots,T} \log f_{\mathcal{N}}(\vec{X}_t; \mu_T, \Sigma_T)$$

where $\vec{X}_t = X_t - \mathbf{B}X_{t-1}$ and $f_{\mathcal{N}}$ is the Gaussian pdf.

Proposition 1. *Given a RKF with a single atom, the MLEs of the Gaussian transition noise are the empirical mean, variance and covariance as follows:*

$$\mu_{MLE} = \begin{bmatrix} m \\ m \\ \vdots \\ m \end{bmatrix}, \Sigma_{MLE} = \begin{bmatrix} \sigma^2 & \sigma' & \cdots & \sigma' \\ \sigma' & \sigma^2 & \cdots & \sigma' \\ \vdots & \vdots & \ddots & \vdots \\ \sigma' & \sigma' & \cdots & \sigma^2 \end{bmatrix} \quad (8)$$

such that

$$m = \frac{1}{n\bar{T}} \sum_{t=2}^{T} \sum_{a \in A} \vec{X}_t(a), \sigma^2 = \frac{1}{n\bar{T}} \sum_{t=2}^{T} \sum_{a \in A} \left(\vec{X}_t(a) - m\right)^2,$$

$$\sigma' = \frac{1}{n(n-1)\bar{T}} \sum_{t=2}^{T} \sum_{\substack{a,a' \in A \\ a \neq a'}} \left(\vec{X}_t(a) - m\right)\left(\vec{X}_t(a') - m\right),$$

where $n = |\vec{X}_t(A)|$ *and* $\bar{T} = T - 1$.

Proof. The MLEs of the parameters (μ_T, Σ_T) are derived by the partial derivatives of the log likelihood:

$$\frac{\partial}{\partial \mu_T} \sum_{t=2}^{T} \log f_{\mathcal{N}}(\vec{X}_t; \mu_T, \Sigma_T) = 0,$$

$$\frac{\partial}{\partial \Sigma_T} \sum_{t=2}^{T} \log f_{\mathcal{N}}(\vec{X}_t; \mu_T, \Sigma_T) = 0.$$

All ground variables generated from the atom $\vec{X}_t$ have the same mean, variance and covariances as shown in Equation (8). Now, we can specify the following linear constraints:

$$m = \frac{1}{\bar{T}} \sum_{t} \vec{X}_t(a_1) = \cdots = \frac{1}{\bar{T}} \sum_{t} \vec{X}_t(a_n).$$

That is, $m = \frac{1}{n\bar{T}} \sum_t \sum_a \vec{X}_t(a)$.

The covariance matrix of the RTM is also calculated from the empirical covariance matrix. The diagonal entries σ^2 are derived as follows:

$$\sigma^2 = \frac{1}{\bar{T}} \sum_{t} (\vec{X}_t(a_1) - m)^2 = \cdots = \frac{1}{\bar{T}} \sum_{t} (\vec{X}_t(a_{n_i}) - m)^2.$$

Thus, $\sigma^2 = \frac{1}{n\bar{T}} \sum_t \sum_a \left(\vec{X}_t(a) - m\right)^2$. Non-diagonal entries (covariances) are derived similarly with $n(n-1)$ empirical covariances.[4] □

[4] One gets the same result when differentiating m, σ^2 and σ' directly from the log-likelihood.

This result is consistent with the result in the non-relational KF because the MLE of the ground KF (μ_T and Σ_T) are known to be the empirical mean and the empirical covariance matrix (Roweis and Ghahramani 1999).

In the general case, when we have multiple atoms, the mean vector and the covariance matrix are block forms as shown in Equation (7). That is, the mean and covariance values are same in each subblock. In case of two atoms X^i and X^j, the means and covariances are represented as follows:

$$\mu_T = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \Sigma_T = \begin{bmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{2,1} & \Sigma_{2,2} \end{bmatrix}$$

The MLE parameters of the RTM are derived similarly with empirical means and covariances of subblocks.

Proposition 2. *Given a RKF with multiple atoms, the MLEs of the Gaussian transition noise are the empirical means, variances and covariances,*

$$\mu_i = [m_i, \cdots, m_i]^T \text{ s.t. } m_i = \frac{1}{n_i\bar{T}} \sum_{t=2}^{T} \sum_{a \in A} \vec{X}_t^{\,i}(a),$$

$$\Sigma^{i,i} = \begin{bmatrix} \sigma^2 & \sigma' & \cdots & \sigma' \\ \sigma' & \sigma^2 & \cdots & \sigma' \\ \vdots & \vdots & \ddots & \vdots \\ \sigma' & \sigma' & \cdots & \sigma^2 \end{bmatrix}, \Sigma^{i,j} = \begin{bmatrix} \sigma'' & \cdots & \sigma'' \\ \vdots & \ddots & \vdots \\ \sigma'' & \cdots & \sigma'' \end{bmatrix},$$

$$\sigma^2 = \frac{1}{n_i\bar{T}} \sum_{t=2}^{T} \sum_{a in A} \left(\vec{X}_t^i(a) - m_i\right)\left(\vec{X}_t^i(a) - m_i\right),$$

$$\sigma' = \frac{1}{n_i(n_i-1)\bar{T}} \sum_{t=2}^{T} \sum_{\substack{a,a' \in A \\ (a \neq a')}} \left(\vec{X}_t^i(a) - m_i\right)\left(\vec{X}_t^i(a') - m_i\right),$$

$$\sigma'' = \frac{1}{n_i n_j\bar{T}} \sum_{t=2}^{T} \sum_{a \in A, b \in B} \left(\vec{X}_t^i(a) - m_i\right)\left(\vec{X}_t^j(b) - m_j\right),$$

where $n_i = |\vec{X}_t^i|$ *and* $\bar{T} = T - 1$.

Proof. The principles used in the proof of Proposition 1 are applied because the $\Sigma^{i,i}$ and $\Sigma^{i,j}$ are block matrices. □

Thus, the block covariance matrix Σ_T can be derived by (1) learning the ground (non-block) covariance matrix $\widehat{\Sigma}_T$ and (2) computing the averages of each subblock. ***BlockAverage*** in Algorithm *LearningRKF* conducts the averaging-out procedure.

Learning Linear Coefficient means to estimate the linear coefficient $\mathbf{B}$ between X_{t-1} and X_t. In this case, given other parameters, Gaussian noise of RTMs and ROMs, the MLE of $\mathbf{B}$ is derived as follow (Roweis and Ghahramani 1999),

$$\widehat{\mathbf{B}} = \left(\sum_{t=2,\cdots,T} X_t X_{t-1}^T\right)\left(\sum_{t=1,\cdots,T-1} X_t X_t^T\right)^{-1}.$$

Here, $\widehat{\mathbf{B}}$ denotes the linear coefficient of the ground TM, and $\widehat{\mathbf{B}}$ will be converted to a block matrix by averaging the

coefficient in each subblock. $\widehat{\mathbf{B}}^{i,j}$ denotes the subblock for the linear transition from X^j_{t-1} to X^i_t. The MLE of the block coefficient is represented as follows:

$$\mathbf{B}^{i,i} = \begin{bmatrix} b & b' & \cdots & b' \\ b' & b & \cdots & b' \\ \vdots & \vdots & \ddots & \vdots \\ b' & b' & \cdots & b \end{bmatrix}, \mathbf{B}^{i,j} = \begin{bmatrix} b'' & \cdots & b'' \\ \vdots & \ddots & \vdots \\ b'' & \cdots & b'' \end{bmatrix},$$

such that

$$b = \frac{1}{n_i} \sum_{k=1}^{n_i} \widehat{\mathbf{B}}^{i,i}_{k,k}, b' = \frac{1}{n_i(n_i-1)} \sum_{\substack{(k,l)=(1,1) \\ k \neq l}}^{(n_i,n_i)} \widehat{\mathbf{B}}^{i,i}_{k,l},$$

$$b'' = \frac{1}{n_i n_j} \sum_{(k,l)=(1,1)}^{(n_i,n_j)} \widehat{\mathbf{B}}^{i,j}_{k,l}.$$

The block coefficient matrix $\mathbf{B}$ is also the MLE of RTM.

3.3 Learning Observation Models

Given RTMs and a sequence of full observations, we derive the estimate X^t at time step t assuming that there is no observation noise.

Learning Observation Noise means to estimate the mean vector and covariance matrix for the ROM. The MLEs problem is formulated as follows:

$$(\mu_{\text{MLE}}, \Sigma_{\text{MLE}}) = \arg\max_{\mu_O, \Sigma_O} \sum_{t=1}^{T} \log f_{\mathcal{N}}(\vec{O}_t; \mu_O, \Sigma_O),$$

where $\vec{O}^i_t = O_t - \mathbf{C} \cdot X^i_t$. The derivation is similar to RTMs. One can substitute $\vec{O}^i_t$ for $\vec{X}^i_t$ in Proposition 2.

Learning Linear Coefficient $\mathbf{C}$ means to compute the linear coefficient between X_t and O_t.

$$\widehat{\mathbf{C}} = \left(\sum_{t=1,\cdots,T} O_t X_t^T \right) \left(\sum_{t=1,\cdots,T} X_t X_t^T \right)^{-1}.$$

The coefficient $\mathbf{C}$ is also computed from $\widehat{\mathbf{C}}$ by averaging out each subblock as in learning $\mathbf{B}$.

4 LRKF with Regroupings

With the estimated parameters, the RKF predicts the state variables in the relational linear dynamic model. This section presents a new lifted Kalman filtering algorithm, which approximately regroups degenerated relational structures. Existing lifted Kalman filtering algorithms (Choi, Guzman-Rivera, and Amir 2011; Ahmadi, Kersting, and Sanner 2011) suffer degenerations of relational structures when sparse observations are made.[5] Algorithm *LRKF-Regroup* also degenerates the domains of relational atoms by calling *DegenAtom* when state variables are observed in different time steps. Thus, the state variables present different Obs (e.g., $Obs_{(i,a)} \neq Obs_{(i,a')}$) where $Obs_{(i,a)}$ stores the most recently observed time for a ground substition a in the i-th atom.

To overcome such degeneracy, *LRKF-Regroup* introduces a new subroutine, called *MergeAtom*, which merges covariance structures when random variables are not directly observed for a certain time step, say k.

[5]Note that, the lifted algorithm in (Choi, Guzman-Rivera, and Amir 2011) is only valid when the same number of observations are made at the same time steps.

Algorithm 2 LRKF-Regroup (Prediction w/ testing data)

Input: params $(\mathbf{B}, \Sigma_T, \mathbf{C}, \Sigma_O)$, obs $(O_1, \cdots, O_T)$
repeat
  $\mu_0 \leftarrow 0, \Sigma_0 \leftarrow 0$
  $(Obs_{(1,1)}, \cdots, Obs_{(n,n_i)}) \leftarrow (0, \cdots, 0)$
  for $t \leftarrow 1$ **to** T **do**
    $(\mu'_t, \Sigma'_t) \leftarrow$ Predict-RTM(μ_{t-1}, Σ_{t-1}, $\mathbf{B}$, Σ_T)
    for all (i, a) s.t. $O^i_t(a)$ is observed **do**
      $Obs_{(i,a)} \leftarrow t$
    end for
    $(\mathbf{B}, \Sigma_T, \mathbf{C}, \Sigma_O) \leftarrow$ *DegenAtom*(Obs, $\mathbf{B}$, Σ_T, $\mathbf{C}$, Σ_O)
    $(\mu_t, \Sigma_t) \leftarrow$ Update-ROM(μ'_t, Σ'_t, $\mathbf{C}$, Σ_O)
    $(\mathbf{B}, \Sigma_T, \mathbf{C}, \Sigma_O) \leftarrow$ *MergeAtom*(Obs, t, $\mathbf{B}$, Σ_T, $\mathbf{C}$, Σ_O)
  end for
until t is T
Output: state estimations $((\mu_0, \Sigma_0), \cdots, (\mu_T, \Sigma_T))$

The *MergeAtom* operation iterates over atoms and finds all state variables which are not observed for a certain time step k. The selected variables are stored in $mlist$. Then, the *BlockAverage* respectively averages diagonal entries and non-diagonal entries in $mlist$, and sets the averaged values to state variables in $mlist$. In this way, it rebuilds the compact relational structure again.

Algorithm 3 MergeAtom

input: recent obs time Obs, time t, params $(\mathbf{B}, \Sigma_T, \mathbf{C}, \Sigma_O)$
$mlist \leftarrow \emptyset$
for $i = 1$ **to** n **do**
  for each a s.t. $Obs_{(i,a)} + k \leq t$ **do**
    $mlist^i \leftarrow mlist^i \bigcup \{a\}$
  end for
end for
$(\mathbf{B}', \Sigma'_T, \mathbf{C}', \Sigma'_O) =$ *BlockAverage*($mlist$, $\mathbf{B}$, Σ_T, $\mathbf{C}$, Σ_O)
output: merged relational structures $(\mathbf{B}', \Sigma'_T, \mathbf{C}', \Sigma'_O)$

The following lemma and theorem are the main theoretical contributions of this paper.[6]

Lemma 3. *When at least one relational observation is made on an atom O_i at every time step, the variance of any state variable $X^i(a)$ in* LRKF-Regroup *is bounded by ${\sigma'^i_{ROM}}^2$ and converges to* $\sqrt{{\sigma'^i_{ROM}}^2 {\sigma^i_{RTM}}^2 + \frac{{\sigma'^i_{ROM}}^4}{4}} - \frac{{\sigma^i_{RTM}}^2}{2}$.

Proof. Let the variance of the i-th atom be σ^i_{RTM} and the variances of direct and relational observations respectively be ${\sigma^i_{ROM}}^2$ and ${\sigma'^i_{ROM}}^2$ where $\sigma^i_{ROM} < \sigma'^i_{ROM}$ as Equations (4) and (5).

Let $\sigma^i_t(a)^2$ be the variance of a state variable $X^i(a)$ at time t. This variable will be updated by at least one relational observation by the *LRKF-Regroup*. Then, the new variance is $\frac{1}{\sigma^i_{t+1}(a)^2} = \frac{1}{\sigma^i_t(a)^2} + \frac{1}{{\sigma'^i_{ROM}}^2}$. That is, $\sigma^i_{t+1}(a)^2 \leq$

[6]The definitions of the direct observation and the relational observation are in Equations (4) and (5).

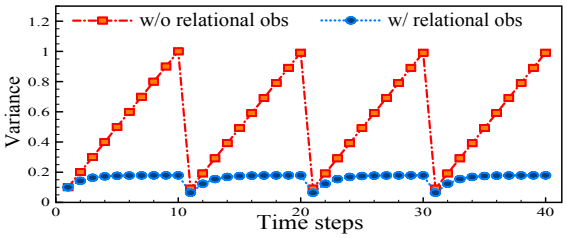

Figure 3: Variances of an atom with and without relational observation (obs). Both models receive one direct obs every 10 time steps. The circle-shaped marks represent the variances of an atom with relational obs at every time step. The square-shaped marks represent the variances of an atom without any relational obs. The setting is $({\sigma^i_{ROM}}^2, {\sigma'^i_{ROM}}^2)=(0.1, 0.5)$.

$\min(\sigma^i_t(a)^2, {\sigma'^i_{ROM}}^2)$. Use the following equation for transition and update in each filtering step:

$$\frac{1}{\sigma^i_{t+1}(a)^2} \leq \frac{1}{\sigma^i_t(a)^2 + {\sigma^i_{RTM}}^2} + \frac{1}{{\sigma'^i_{ROM}}^2}.$$

For the convergence, let $\sigma^i_{t+1}(a)=\sigma^i_t(a)=\sigma^i_*(a)$. The variance is then $\sigma^i_*(a)^2 = \sqrt{{\sigma'^i_{ROM}}^2 {\sigma^i_{RTM}}^2 + \frac{{\sigma'^i_{ROM}}^4}{4}} - \frac{{\sigma^i_{RTM}}^2}{2}$ □

Lemma 3 shows that the variance of an atom is bounded when consecutive relational observations are made. Figure 3 illustrates the intuitions of the lemma.

Theorem 4. *When (1) no direct observation is made on two state variables $X^i(a)$ and $X^i(a')$ in an atom O_i at least for k time steps; and (2) at least one relational observation is made on the other variables in the atom, the difference of the variances of two state variables $\sigma^i_t(a)$ and $\sigma^i_t(a')$ is bounded by $O(c^k)$ where c is $\sigma^i_*(a)^2/(\sigma^i_*(a)^2+{\sigma'^i_{ROM}}^2)$ and $c \leq 1$.*

Proof. We follow the result of the Lemma 3 and use $\sigma^i_*(a)$. The variance of each time step follows the recursive form:

$$\frac{1}{\sigma^i_{t+1}(a)^2} = \frac{1}{\sigma^i_t(a)^2 + {\sigma^i_{RTM}}^2} + \frac{1}{{\sigma'^i_{ROM}}^2}. \quad (9)$$

An exact (non-recursive) formula for $\sigma^i_{t+1}(a)$ is non-trivial. Thus, we introduce another simpler, converging sequence,

$$\bar{\sigma}^i_{t+1}(a)^2 = c(\bar{\sigma}^i_t(a)^2 + {\sigma^i_{RTM}}^2).$$

Since $\sigma^i_t(a)^2 - \bar{\sigma}^i_t(a)^2$ is positive and convex when $c < 1$,

$0 \leq \bar{\sigma}^i_t(a) \leq \sigma^i_t(a) \leq \sigma^i_*(a)$, $\sigma^i_*(a)^2 - \bar{\sigma}^i_t(a)^2 \geq \sigma^i_*(a)^2 - \sigma^i_t(a)^2$.

The convergence of the simpler form is slower than the original one in Equation (9). However, it provides an exact formulation and converges exponentially:

$$\bar{\sigma}^i_t(a)^2 = c^k \bar{\sigma}^i_{t-k}(a)^2 + {\sigma^i_{RTM}}^2 \frac{1-c^k}{1-c}.$$

WLOG, we set $X^i(a')$ has no direct observation longer than $X^i(a)$. The variance of $X^i(a')$ has the same formulation

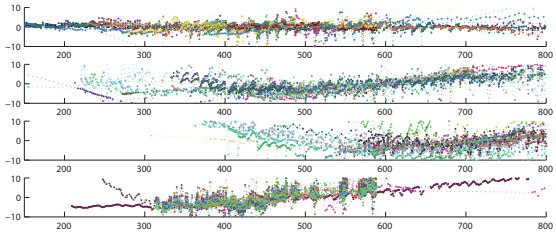

Figure 4: The water levels of wells in four sets of clusters. Each plot shows distinct group behaviors. The x axis is the time step in month and the y axis is the head (of well).

with a substitution of $k+\alpha$ for k. Thus, the variance of $X^i(a')$ is represented as follows:

$$\bar{\sigma}^i_t(a')^2 = c^{k+\alpha} \cdot \bar{\sigma}^i_{t-k-\alpha}(a')^2 + {\sigma^i_{RTM}}^2 \frac{1-c^{k+\alpha}}{1-c}.$$

Note that, $\bar{\sigma}^i_{t+\alpha}(a) \leq \bar{\sigma}^i_{t+\alpha}(a') \leq \bar{\sigma}^i_{t+\infty}(a')$.

$$\begin{aligned}
&|\sigma^i_{t+\alpha}(a')^2 - \sigma^i_t(a)^2| \leq |\sigma^i_{t+\infty}(a')^2 - \sigma^i_t(a)^2| \\
&= c^k({\sigma^i_{RTM}}^2/(1-c) - \sigma^i_{t+k}(a)^2) \\
&= c^k({\sigma^i_{RTM}}^2(1+\sigma^i_*(a)^2/{\sigma^i_{ROM}}^2) - \sigma^i_{t+k}(a)^2) \\
&\leq c^k {\sigma^i_{RTM}}^2(1+\sigma^i_*(a)^2/{\sigma^i_{ROM}}^2) = O(c^k).
\end{aligned}$$

□

For the well example, suppose that $X^{r1}(40.5N,103W)$ (or $X(A)$) and $X^{r_1}(40.2N,103W)$ (or $X(B)$) are in an atom X^{r_1}. As time goes, each well may be observed in different time steps. At time step 10, $X(A)$ is observed through the observation variable $O(40.5N,103W)$ (or $O(A)$) while $X(B)$ are not observed, yet. Now, two wells have different variances and covariances, thus different pairwise Gaussian factors. The pairwise Gaussian parfactor for $X(A)$ and $X(B)$ cannot be shared anymore. Thus, the atom X^{r_1} should be **degenerated** (or divided) into two parts. The *degeneration* corresponds to the terms, **split** (Poole 2003) and **shatter** (de Salvo Braz, Amir, and Roth 2005).

Since the degeneration, $X(A)$ and $X(B)$ are not observed for another 15 time steps. As time goes, the variances of $X(A)$ and $X(B)$ increase again. However, with at least one relational observation at every time step, the variances of $X(A)$ and $X(B)$ converge, and the difference reduces exponentially as we show in Theorem 4.

Thus, given the consecutive relational observations, we can compute the variances of the degenerated variables without searching individuals. That is, k in Algorithm *MergeAtom* is determined to guarantee that the error of variances is less than a bound of our interest.

5 Experimental Results

In the experiments, we use multiple synthetic data sets and a real-world groundwater flow MODFLOW model, the Republican River Compact Association (RRCA) model (McKusick 2003) as shown in Figure 1. The dataset extracted from the RRCA model includes a set of monthly

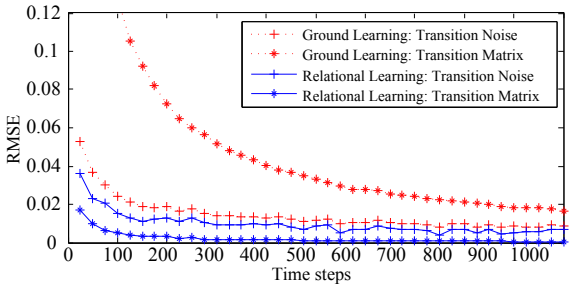

Figure 5: Error of estimated parameters (transition matrix and transition noise) in ground KF and LRKF.

measured heads (water level) at over 3,000 wells for 850 months. Not all wells are measured every month. There are different groups of wells which show various dynamic behaviors as shown in Figure 4. We choose 1,182 wells which have at least 45 measurements (about 5% of 850).

5.1 The Accuracy of Parameter Learning

For the experiments with synthetic data, we generate 750 time steps of samples for Relational Linear (Transition) Models with 6 atoms each with 10 variables. That is, our synthetic training data is composed of 60 by 750 numerical values which is similar to Figure 2. Then, we compare our relational learning algorithm presented in Section 3 with the state-of-the-art (ground) learning algorithm for the vanilla KF (Digalakis, Rohlicek, and Ostendorf 1993; Ghahramani and Hinton 1996). In each time step, we compute the MLE of the transition matrix and the transition noise (B_X and G_{RTM} in Equation (1)). Then, we measure the average root mean-square error (RMSE) between the true parameters and the estimated parameters after 10 repeated experiments. Figure 5 shows that our relational learning can find the true parameters much faster than ground learning. This result implies that our relational learning performs better than ground learning when samples are generated from relational models.

To compare the two algorithms in a real-world data, we conduct experiments on the RCAA data. Handling the RCAA data is challenging because it has irregular, noisy measurements with various dynamic changes. In addition, relational information (i.e., which wells are included in which atom) is not given. To build the relational information, we cluster the set of water wells by spectral clustering (Ng, Jordan, and Weiss 2001).[7] Figure 4 shows four representative groups of water wells.

Then, we compute the MLEs parameters: the transition matrix and noise (B_X and G_{RTM} in Equation (1)); and the observation matrix and noise (H_X and G_{ROM} in Equation (4)). To compute the model accuracy, we prepare test data as the last 20% of measurements for randomly sampled (50%) water wells. That is, 10% of all measurements (about 10,038) are reserved for testing. All other measurements are used for training. We compare the two methods in terms of the RMSE and the negative log probability ($-log(p(\text{data}|\text{prediction}))$). Note that KFs provide value predictions with uncertainty (the variances). For the 10,038 test measurements, the average RMSEs of the LRKF and the vanilla KF respectively were **4.36** and 5.10. The average negative log probabilities of the LRKF and the vanilla KF respectively were **3.88** and 4.91. As shown in Figure 7, the RRCA model data are very dynamic and noisy. The result shows that LRKF models such complex real-world data better than the vanilla KF.

[7]The distance matrix between water wells is computed by measurements of co-occurrences and their average differences.

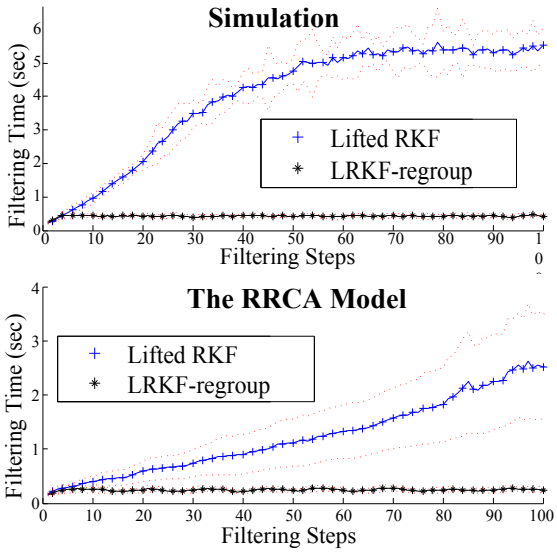

Figure 6: Filtering time of Lifted RKF and LRKF-Regroup in a simulation and the RRCA groundwater model.

5.2 The Efficiency of Filtering

For the filtering efficiency, we compare the *Lifted RKF* (Choi, Guzman-Rivera, and Amir 2011) and our *LRKF-Regroup*.[8] The algorithms are compared on two datasets with sparse observations: one synthetic dataset and one real-world groundwater data. Note that the lifted RKF will not degenerate the model on full, dense observations. In both experiments, we set k in Algorithm *MergeAtom* to be 4. That is, two state variables will be merged if they have the same observation numbers and types when at least one relational observation is made.

In synthetic data, we assume an atom with 300 ground substitutions, i.e., $|X^i|$=300. Then we make a sparse observations with a rate of 90%. That is, 90% of state variables will be observed in each time step. Then, we report the average filtering time of *Lifted RKF* and *LRKF-Regroup* in the simulation and the RRCA model. The experimental results are presented in Figure 6.

[8]Source code is available at `http://pail.unist.ac.kr/LRKF`.

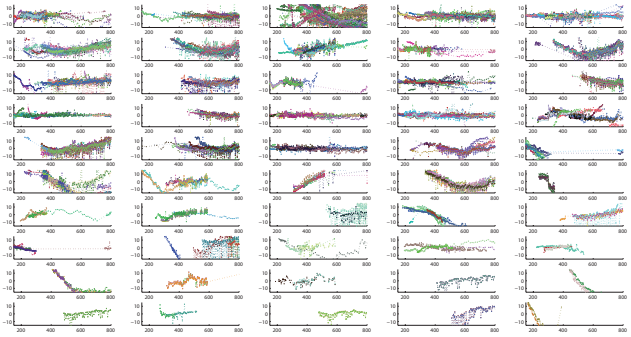

Figure 7: The water levels of all wells in the RRCA data. Each of 50 plots represents wells in a group. The x axis is the time (in months). The y axis is the water level (head).

6 Conclusion

This paper provides new answers and insights on (1) how to learn parameters for the RKF; and (2) how to regroup the state variables from noisy real-world data. We propose a new learning algorithm that regroups the state variables when individual observations are made to the RKF in different time steps. In a simulated dataset and a real-world dataset, we demonstrate that the new algorithm significantly improves the accuracy and the efficiency of filtering for large-scale dynamic systems.

7 Acknowledgements

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science, ICT & Future Planning (no. NRF-2014R1A1A1002662), the ICT R&D program of MSIP/IITP [10035348, *Development of a Cognitive Planning and Learning Model for Mobile Platforms*], and NSF award ECS-09-43627 - *Improving Prediction of Subsurface Flow and Transport through Exploratory Data Analysis and Complementary Modeling*.

References

Ahmadi, B.; Kersting, K.; and Sanner, S. 2011. Multi-evidence lifted message passing, with application to pagerank and the kalman filter. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 1152–1158.

Bahmani-Oskooee, M., and Brown, F. 2004. Kalman filter approach to estimate the demand for international reserves. *Applied Economics* 36(15):1655–1668.

Choi, J., and Amir, E. 2012. Lifted relational variational inference. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 196–206.

Choi, J.; Guzman-Rivera, A.; and Amir, E. 2011. Lifted relational kalman filtering. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2092–2099.

Choi, J.; Hill, D. J.; and Amir, E. 2010. Lifted inference for relational continuous models. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, 126–134.

Clark, M. P.; Rupp, D. E.; Woods, R. A.; Zheng, X.; Ibbitt, R. P.; Slater, A. G.; Schmidt, J.; and Uddstrom, M. J. 2008. Hydrological data assimilation with the ensemble kalman filter: Use of streamflow observations to update states in a distributed hydrological model. *Advances in Water Resources* 31(10):1309 – 1324.

de Salvo Braz, R.; Amir, E.; and Roth, D. 2005. Lifted first-order probabilistic inference. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 1319–1325.

Digalakis, V.; Rohlicek, J.; and Ostendorf, M. 1993. Ml estimation of a stochastic linear system with the em algorithm and its application to speech recognition. *IEEE Transactions on Speech and Audio Processing* 1(4):431–442.

Evensen, G. 1994. Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics. *Journal of Geophysical Research* 99:10143–10162.

Friedman, N.; Getoor, L.; Koller, D.; and Pfeffer, A. 1999. Learning probabilistic relational models. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 1300–1309.

Ghahramani, Z., and Hinton, G. E. 1996. Parameter estimation for linear dynamical systems. Technical report, CRG-TR-96-2, University of Toronto, Dept. of Computer Science.

Kalman, R. E. 1960. A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering* 82(Series D):35–45.

Kersting, K. 2012. Lifted probabilistic inference. In *Proceedings of European Conference on Artificial Intelligence.*, 33–38.

McKusick, V. 2003. Final report for the special master with certificate of adoption of rrca groundwater model. *State of Kansas v. State of Nebraska and State of Colorado, in the Supreme Court of the United States* 3605(1145):4440.

Ng, A. Y.; Jordan, M. I.; and Weiss, Y. 2001. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, 849–856. MIT Press.

P., M. F., and Bierkens. 2001. Spatio-temporal modelling of the soil water balance using a stochastic model and soil profile descriptions. *Geoderma* 103(1–2):27–50.

Poole, D. 2003. First-order probabilistic inference. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 985–991.

Raedt, L. D. 2008. *Logical and Relational Learning: From ILP to MRDM (Cognitive Technologies)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

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

Roweis, S., and Ghahramani, Z. 1999. A unifying review of linear gaussian models. *Neural Comput.* 11(2):305–345.

Wang, J., and Domingos, P. 2008. Hybrid markov logic networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1106–1111.