

Deep Kalman Filters

発表者：鈴木雅大

論文について

- Rahul G. Krishnan, Uri Shalit, David Sontag (arXiv, 2015/11/16)
 - ニューヨーク大学
- Deep Learning + Kalman Filter
 - VAEを使って**因果推論**しているのが大きな特徴.
- 前回に引き続きDeep...という論文.
 - 最近多い気がする.

関連研究

カルマンフィルタ

- 非観測（潜在）変数： $z_1, \dots, z_T \in \mathbb{R}^s$
- 観測変数： $x_t \in \mathbb{R}^d$
- 行動： $u_t \in \mathbb{R}^c$
- このときカルマンフィルタモデルは次のような線形方程式になる。

$$z_t = G_t z_{t-1} + B_t u_{t-1} + \epsilon_t \text{ (action-transition)}$$
$$x_t = F_t z_t + \eta_t \text{ (observation)}$$

時間遷移雑音

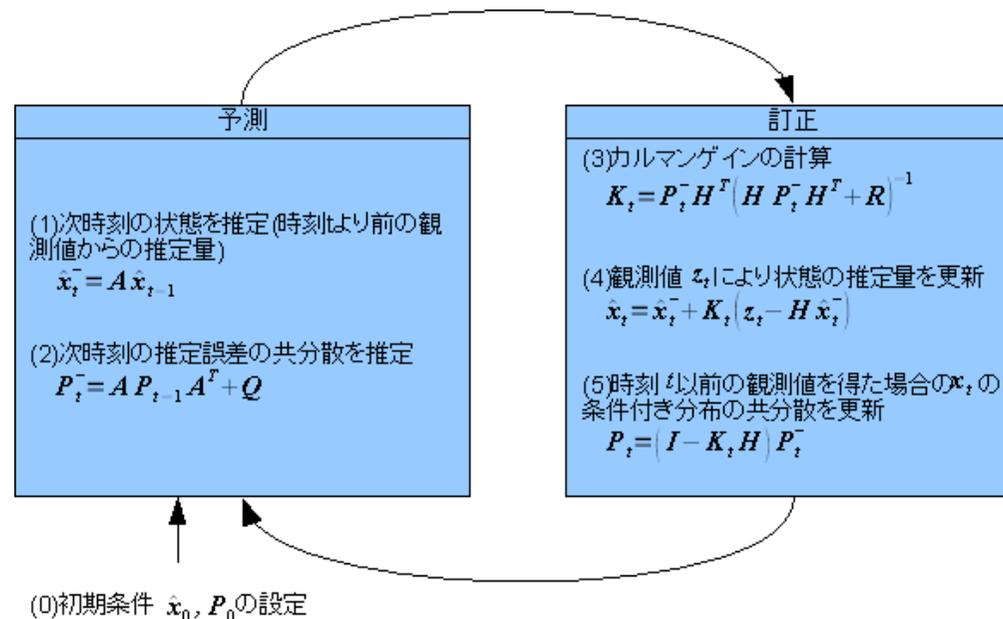
観測雑音

ただし、

- $\epsilon_t \sim \mathcal{N}(0, \Sigma_t), \eta_t \sim \mathcal{N}(0, \Gamma_t)$
- 時変行列：
 - 時間遷移の線形モデル： $G_t \in \mathbb{R}^{s \times s}$
 - 観測モデル： $F_t \in \mathbb{R}^{d \times s}$.

カルマンフィルタ

- カルマンフィルタは予測と修正を繰り返す



- 応用例 :

- ロケットの軌道補正

- 医療診断

- 本論文では時間変化する患者の生成モデルを目標としている。

カルマンフィルタ

- カルマンフィルタは線形システム

$$z_t = G_t z_{t-1} + B_t u_{t-1} + \epsilon_t \text{ (action-transition)}$$

$$x_t = F_t z_t + \eta_t \text{ (observation)}$$

- 非線形なモデルにすると, 事後分布 $p(z_1, \dots, z_T | x_1, \dots, x_T, u_1, \dots, u_T)$ が扱いづらくなる.

確率的誤差逆伝播

Stochastic Backpropagation

- SGVBのこと.

$$\begin{aligned}\log p_{\theta}(x) &= \log \int_z \frac{q_{\phi}(z|x)}{q_{\phi}(z|x)} p_{\theta}(x|z) p_0(z) dz \geq \int_z q_{\phi}(z|x) \log \frac{p_{\theta}(x|z) p_0(z)}{q_{\phi}(z|x)} dz \\ &= \mathbb{E}_{q_{\phi}(z|x)} [\log p_{\theta}(x|z)] - \text{KL}(q_{\phi}(z|x) || p_0(z)) = \mathcal{L}(x; (\theta, \phi)),\end{aligned}$$

- いつものVariational autoencoder(VAE)のように期待値部分はサンプリング.

$$q_{\phi}(z|x) \sim \mathcal{N}(\mu_{\phi}(x), \Sigma_{\phi}(x))$$

反事実推定

反事実推定 (Counterfactual Estimation)

- 実際に観測したものとは異なる状況下での確率を推定する.
- 英語で言う**仮定法**
 - もし違う薬を処方したら, 患者の血糖値はどうなるのか?
- しかし確率言語では, 厳密には因果関係 (原因と結果) を記述することは**できない**.
 - 確率言語で記述できるのは「観察的記述」のみ
 - $p(\text{雨}|\text{濡れる})$ は, 濡れたという状態を**観測した**という条件下に過ぎない.
 - もし濡らしたならば, という「介入的記述」はできない・・・
- 補足: 統計学では, 因果関係をものすごく嫌っている.
 - ピアソンとかヒュームの影響?

統計的因果推論

Pearl(2009)

- これまでの確率言語を拡張し、新たに**do演算子**を導入することを提案した。

$$p(y|x) \rightarrow p(y|do(x))$$

- 左は, x であるときの y
 - 右は, x に**変えた**ときの y
- do演算子を導入することで, 反事実推論を定式化できる.
 - グラフィックモデルでわかっている場合, 関心のある変数の値を変更して, 得られたサブグラフで実行することで可能
- 細かい話は「統計的因果推論」でググってください・・・

提案モデル

提案モデル

- 次のようなシーケンスを考える

- 観測 $\vec{x} = (x_1, \dots, x_T)$

- 行動 $\vec{u} = (u_1, \dots, u_{T-1})$

- 潜在状態 $\vec{z} = (z_1, \dots, z_T)$

- このとき、提案モデルの生成過程は次のようになる

$$z_1 \sim \mathcal{N}(\mu_0; \Sigma_0)$$

$$z_t \sim \mathcal{N}(G_\alpha(z_{t-1}, u_{t-1}, \Delta_t), S_\beta(z_{t-1}, u_{t-1}, \Delta_t))$$

$$x_t \sim \Pi(F_\kappa(z_t)).$$

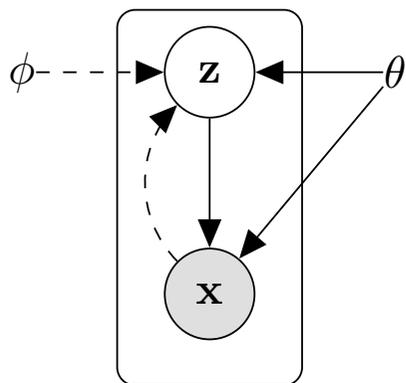
非線形

非線形

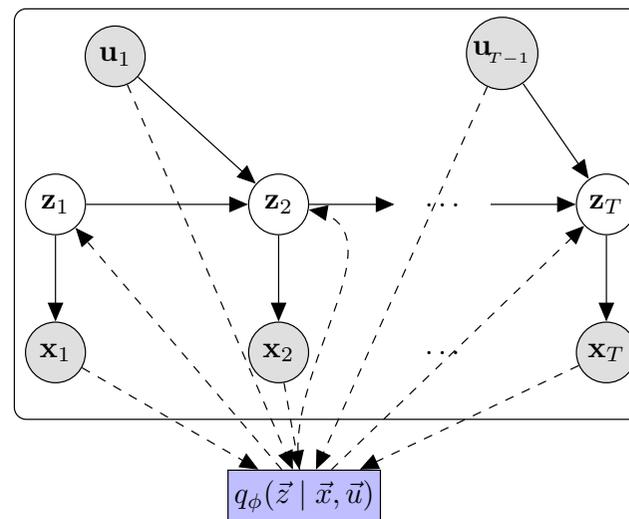
非線形

VAEとの違い

- 生成モデルで比較
 - 点線部分が変分推論
 - DKFでは, 推論部分を $q_\phi(\vec{z}|\vec{x}, \vec{u})$ に集約する.



(a) Variational Autoencoder



(b) Deep Kalman Filter

尤度最大化

- 目的は次の尤度の最大化

$$\log p_{\theta}(x_1 \dots, x_T | u_1 \dots u_{T-1})$$

- しかし, 厳密には解けないので変分推論
- 変分推論部分は次のようになる.

$$q_{\phi}(\vec{z} | \vec{x}, \vec{u}) = \prod_{t=1}^T q(z_t | z_{t-1}, x_t, \dots, x_T, \vec{u})$$

- 下界は次のように求められる.

$$\begin{aligned} \log p_{\theta}(\vec{x} | \vec{u}) &\geq \int_{\vec{z}} q_{\phi}(\vec{z} | \vec{x}, \vec{u}) \log \frac{p_0(\vec{z} | \vec{u}) p_{\theta}(\vec{x} | \vec{z}, \vec{u})}{q_{\phi}(\vec{z} | \vec{x}, \vec{u})} d\vec{z} \\ &= \mathbb{E}_{q_{\phi}(\vec{z} | \vec{x}, \vec{u})} [\log p_{\theta}(\vec{x} | \vec{z}, \vec{u})] - \text{KL}(q_{\phi}(\vec{z} | \vec{x}, \vec{u}) || p_0(\vec{z} | \vec{u})) \end{aligned}$$

$$\text{(using } x_t \perp\!\!\!\perp x_{-t} | \vec{z}) \sum_{t=1}^T \mathbb{E}_{z_t \sim q_{\phi}(z_t | \vec{x}, \vec{u})} [\log p_{\theta}(x_t | z_t, u_{t-1})] - \text{KL}(q_{\phi}(\vec{z} | \vec{x}, \vec{u}) || p_0(\vec{z} | \vec{u})) = \mathcal{L}(x; (\theta, \phi))$$

下界の計算

- KLダイバージェンス部分は次のようにもとまる.

$$\text{KL}(q_\phi(\vec{z}|\vec{x}, \vec{u})||p_0(\vec{z}))$$

$$= \int_{z_1} \dots \int_{z_T} q_\phi(z_1|\vec{x}, \vec{u}) \dots q_\phi(z_T|z_{T-1}, \vec{x}, \vec{u}) \log \frac{p_0(z_1, \dots, z_T)}{q_\phi(z_1|\vec{x}, \vec{u}) \dots q_\phi(z_T|z_{T-1}, \vec{x}, \vec{u})} d\vec{z}$$

$$\underbrace{\text{(Factorization of } p(\vec{z}))}_{=}$$
$$\text{KL}(q_\phi(z_1|\vec{x}, \vec{u})||p_0(z_1))$$

$$+ \sum_{t=2}^T \mathbb{E}_{z_{t-1} \sim q_\phi(z_{t-1}|\vec{x}, \vec{u})} [\text{KL}(q_\phi(z_t|z_{t-1}, \vec{x}, \vec{u})||p_0(z_t|z_{t-1}, u_{t-1}))].$$

- よって下界は

$$\log p_\theta(\vec{x}|\vec{u}) \geq \mathcal{L}(x; (\theta, \phi)) =$$

$$\sum_{t=1}^T \mathbb{E}_{q_\phi(z_t|\vec{x}, \vec{u})} [\log p_\theta(x_t|z_t)] - \text{KL}(q_\phi(z_1|\vec{x}, \vec{u})||p_0(z_1))$$

$$- \sum_{t=2}^T \mathbb{E}_{q_\phi(z_{t-1}|\vec{x}, \vec{u})} [\text{KL}(q_\phi(z_t|z_{t-1}, \vec{x}, \vec{u})||p_0(z_t|z_{t-1}, u_{t-1}))]$$

識別モデルの選択

- 識別モデル（変分推論部分）は次のようなモデル化方法がある.

- **q-INDEP**: $q(z_t|x_t, u_t)$ parameterized by an MLP
- **q-LR**: $q(z_t|x_{t-1}, x_t, x_{t+1}, u_{t-1}, u_t, u_{t+1})$ parameterized by an MLP
- **q-RNN**: $q(z_t|x_1, \dots, x_t, u_1, \dots, u_t)$ parameterized by a RNN
- **q-BRNN**: $q(z_t|x_1, \dots, x_T, u_1, \dots, u_T)$ parameterized by a bi-directional RNN

→実験で検証

- 事後確率の因子分解は次のようになる

$$p(\vec{z}|\vec{x}, \vec{u}) = p(z_1|\vec{x}, \vec{u}) \prod_{t=2}^T p(z_t|z_{t-1}, x_t, \dots, x_T, u_{t-1}, \dots, u_{T-1})$$

- 実は \vec{x} の全系列を使わなくても、事後分布の推定は可能
- p に q をなるべく近似する方法として、Universality of normalizing flows[Rezende+ 2015]がある

実験

Healing MNIST

- 本当は医療データを使いたいが、不確かでノイズが入っている上に、いろいろ面倒.
- 本研究では、MNISTデータを加工したものをを用いる.
 - 回転を行動 u 、回転した画像を観測 x とする.
 - シーケンスのランダムな3連続画像の左上に四角形を重ねている.
 - 行動 u とは関係ない影響を意図（季節的なインフルや病気）
 - 観測する画像には20%のノイズが入っている.
- このデータを使って次の実験をする.
 - Small Healing MNIST
 - 40000の長さ5のシーケンスで訓練
 - Large Healing MNIST
 - 140000の長さ5のシーケンスで訓練

作成したデータのイメージ



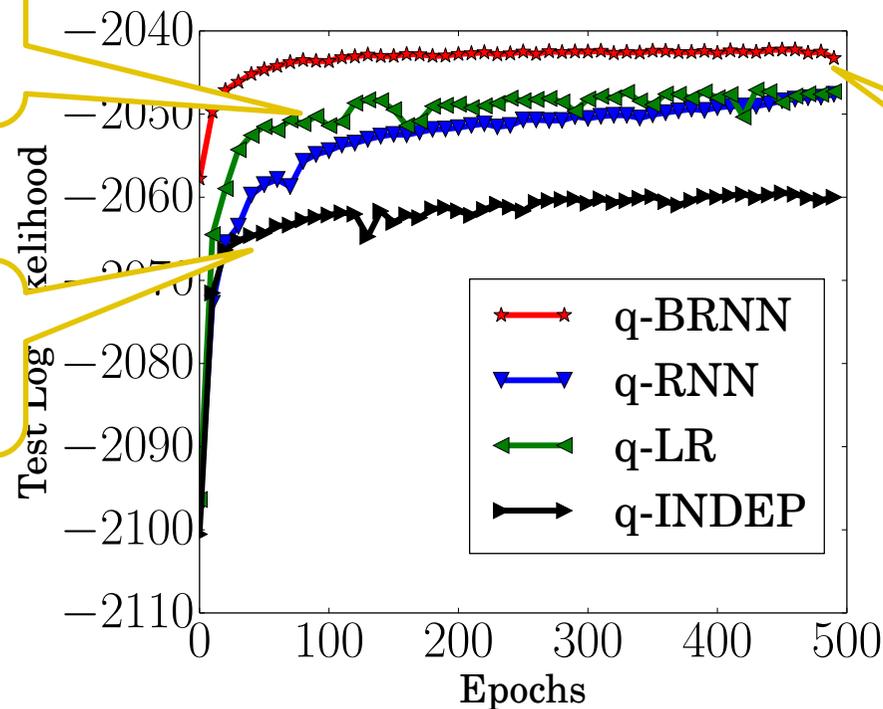
Small Healing MNIST

□ テスト対数尤度で4つの識別モデルを評価

- **q-INDEP**: $q(z_t|x_t, u_t)$ parameterized by an MLP
- **q-LR**: $q(z_t|x_{t-1}, x_t, x_{t+1}, u_{t-1}, u_t, u_{t+1})$ parameterized by an MLP
- **q-RNN**: $q(z_t|x_1, \dots, x_t, u_1, \dots, u_t)$ parameterized by a RNN
- **q-BRNN**: $q(z_t|x_1, \dots, x_T, u_1, \dots, u_T)$ parameterized by a bi-directional RNN

q-LRの方が良くなっているが、実際のサンプルでは四角形は捉えられていない

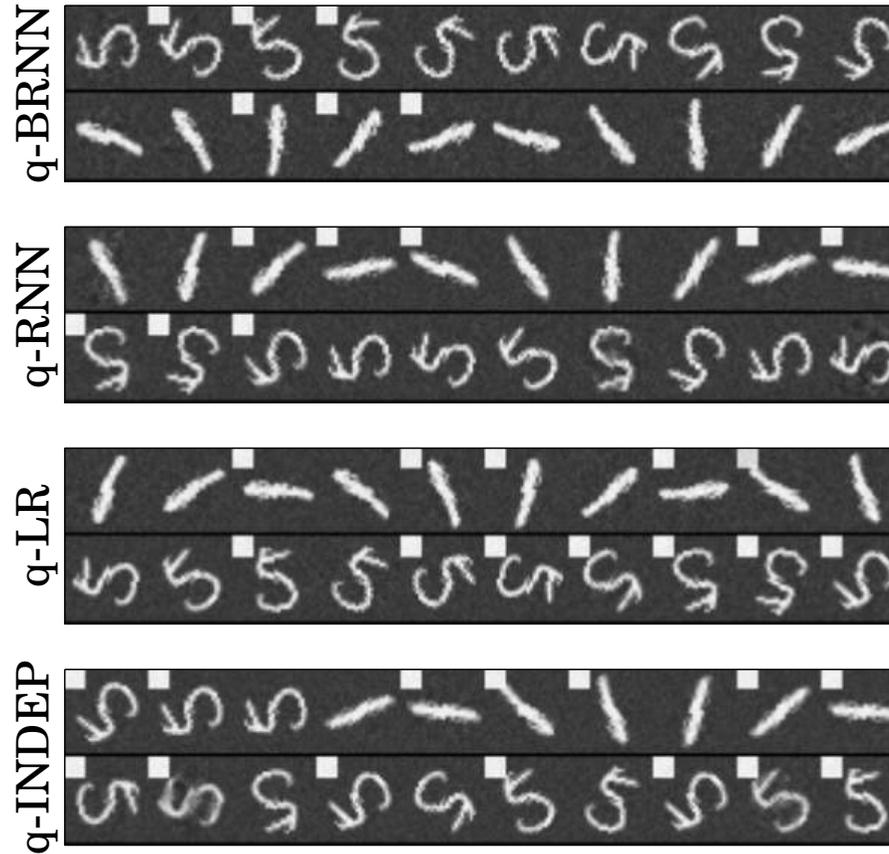
q-INDEPが悪くなるのも当然（四角形を捉えられない）



q-BRNNが一番良くなるのは当然（forward backward法に似てる）

Small Healing MNIST

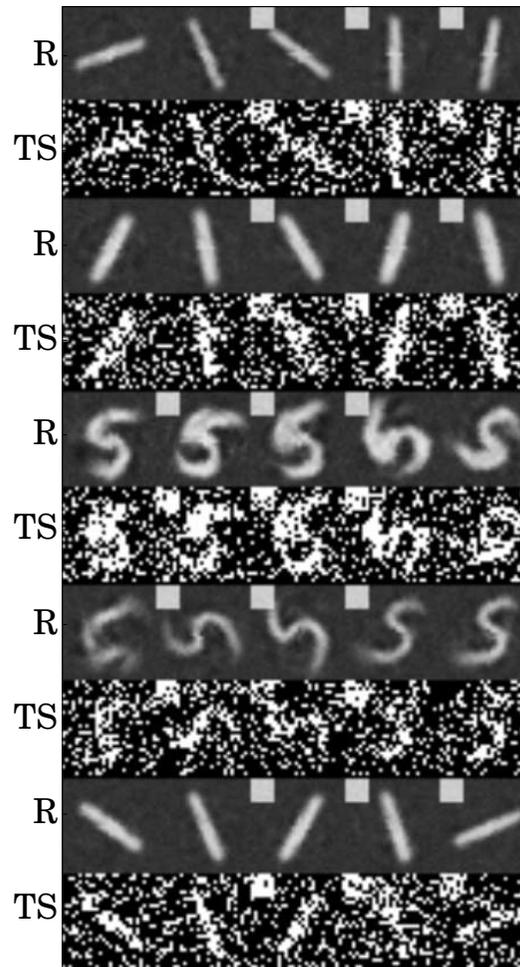
□ 各モデルでのサンプル



□ q-BRNNとq-RNNが良さそう

Large Healing MNIST

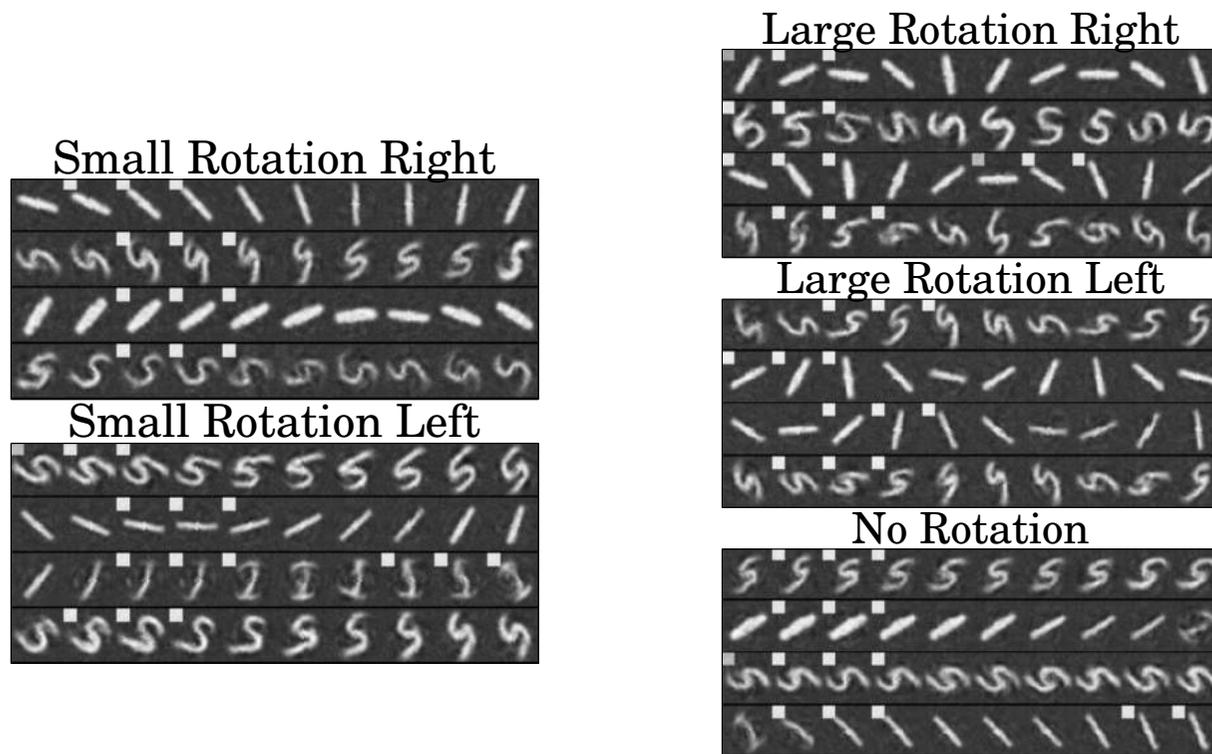
- 再構成した画像
- TS:観測した画像
- R:再構成画像



- 異なるスタイルの数字を学習している (= 個々の患者)

Large Healing MNIST

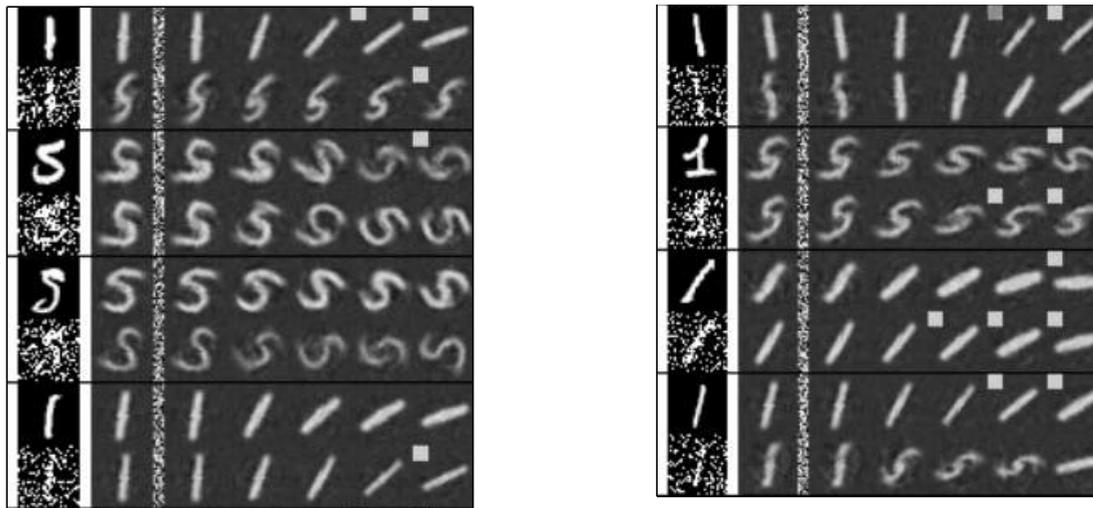
- 回転角度を調整（治療の強さに該当）



- 見たことのない行動でも生成できる = 反事実推論の簡単な例

Large Healing MNIST

- 見たことのない例で系列を生成



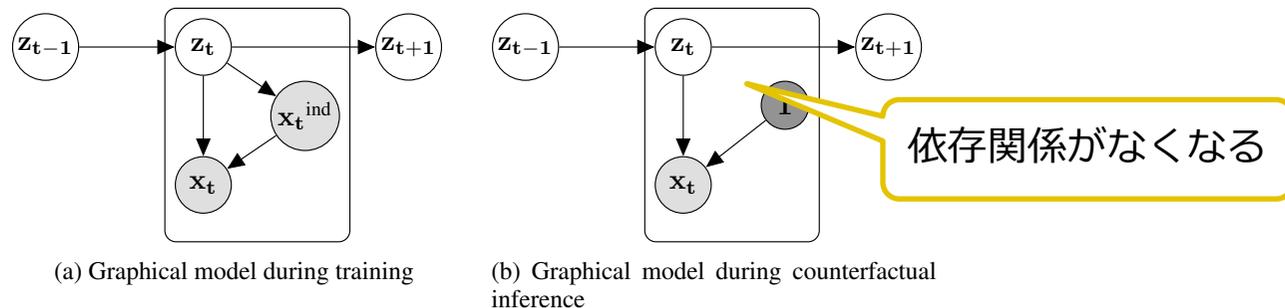
- 見たことのない画像でも系列を生成できた
 - 新しい患者に対しても，薬を処方した時のシミュレーションができる！
- ノイズを加えてもうまく生成できることがわかった。

医療データへの応用

- 大手健康保険業者のヘルスケア請求データを利用
 - データの詳細は全く不明
 - 8000人の糖尿病患者に抗糖尿病薬を飲ませた時の効果を調査
 - 観測変数：
 - A1cレベル, グルコース, その他年齢, 性別など
 - 行動変数：
 - 9種類の糖尿病の薬（インシュリンやメトホルミンなど）
 - 3か月間隔の4年半のデータを1シーケンスとする。
- このデータを学習して, **反事実推論**を行いたい。
 - 所定の薬で治療していなかった場合, 患者のA1cとグルコースレベルはどうなっていたのか？

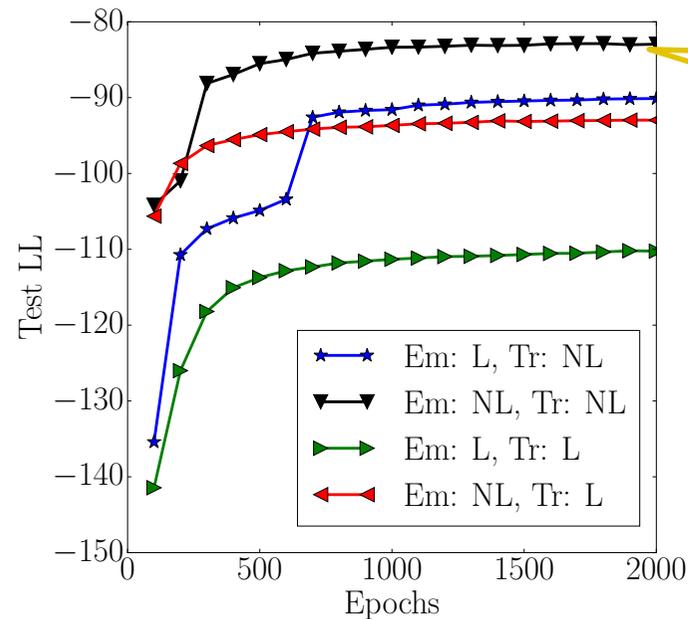
医療データへの応用

- しかし、いくつか問題点がある。
 - A1cとグルコースレベルは常に測っているわけではない
 - しかも糖尿病の疑いが高い人の方がより測る頻度が多い = 混乱の要因
- そこで、新たにlab indicator変数 x^{ind} を用意する
 - 結果にかかわらず、labのテストを受けたかどうかを表す。
 - これによって、観測変数が**潜在変数とlab indicator変数に条件づけられる。**
- 反事実推論をするときには、**do演算子**を用いることで推論が可能
 - Lab indicatorを1にして、潜在変数との依存関係をなくす



実験結果

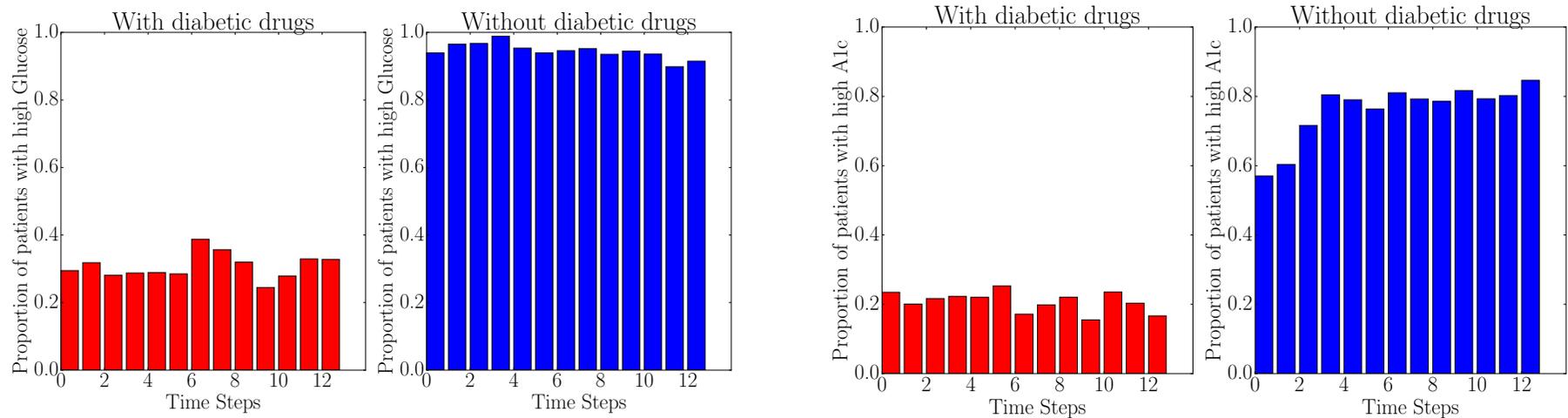
- 識別モデルにq-BRNNを使い, 8000人のデータで学習
- カルマンフィルタを線形モデルと非線形モデルで比較実験
 - L : linear relationship
 - NL : 2層のNN (パラメータは1つ)
 - Em(ission) : F_k
 - Tr(ansition) : G_a



非線形モデルが
一番尤度が高い

実験結果

- 非線形モデルを利用して反事実推論をする。
 - メトホルミンを最初に処方された800人のデータを利用
 - その後処方を続けた場合 (with) と続けなかった (without) 場合で比較
 - グルコースレベルとA1cレベルの高い患者の割合で比較



- 処方を受けなかった患者が明らかに悪くなっている。

まとめ

- 非線形なカルマンフィルタが反事実推論に有効であることがわかった。
 - 事後分布は見たことのない状態も推論できた。
 - 異なる行動でも、系列をサンプルできた。
- 反事実推論だけでなく、患者を潜在変数に落とし込んで、似た患者を探すこともできる。
 - 潜在変数が患者個人をちゃんとエンコードしているか、潜在変数の変化が知られている病気の軌跡と対応しているか、などが研究課題
- 今後は他の領域に適用
 - ビデオ映像の教師なし学習

感想

- VAEを画像生成ではなく確率モデルとして活用した初めての例？
 - Deep Learningで因果推論が可能になった.
 - 個人的にはVAEの正しい使い方だと思う.
- 実験自体はまだ大したことはない.
 - 今後に期待？
 - 着想はかなり面白い

参考文献

□ 林岳彦氏の資料

- 相関と因果について考える：統計的因果推論、その(不)可能性の中心
<http://www.slideshare.net/takehikoihayashi/ss-13441401>
- 確率と因果を革命的に架橋する：Judea Pearlのdo演算子
<http://takehiko-i-hayashi.hatenablog.com/entry/20111222/1324487579>