THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Chapter 11: Dimensionality Reduction for Transition Model Estimation

Yangjianchen Xu

The University of North Carolina at Chapel Hill

September 30, 2022

Table of Contents

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Outline

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Sufficient dimensionality reduction

- Sufficient dimensionality reduction: a framework of dimensionality reduction in a supervised learning setting of analyzing an input-output relation.
- Input: state-action pair $(\boldsymbol{s}, a)$; Output: next state $\boldsymbol{s}'$.
- Goal: find a matrix $\boldsymbol{W}$ which induces a linear projection of input $(\boldsymbol{s}, a)$:

$$\boldsymbol{z} = \boldsymbol{W} \begin{pmatrix} \boldsymbol{s} \\ a \end{pmatrix},$$

  such that $\boldsymbol{s}'$ and $(\boldsymbol{s}, a)$ are conditionally independent given $\boldsymbol{z}$ and $\boldsymbol{W}\boldsymbol{W}^\top = \boldsymbol{I}$.
- $\boldsymbol{z}$ contains all information about $\boldsymbol{s}'$:

$$p(\boldsymbol{s}, a, \boldsymbol{s}' \mid \boldsymbol{z}) = p(\boldsymbol{s}, a \mid \boldsymbol{z}) p(\boldsymbol{s}' \mid \boldsymbol{z}) \Leftrightarrow p(\boldsymbol{s}' \mid \boldsymbol{s}, a) = p(\boldsymbol{s}' \mid \boldsymbol{z})$$

Outline

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Outline

Conditional Independence

- Squared-loss conditional entropy (SCE) is defined as

$$\begin{aligned}\mathrm{SCE}\left(\boldsymbol{s}' \mid \boldsymbol{z}\right) &= -\frac{1}{2}\iint p\left(\boldsymbol{s}' \mid \boldsymbol{z}\right) p\left(\boldsymbol{s}', \boldsymbol{z}\right) \mathrm{d}\boldsymbol{z}\mathrm{d}\boldsymbol{s}' \\ &= -\frac{1}{2}\iint \left(p\left(\boldsymbol{s}' \mid \boldsymbol{z}\right) - 1\right)^2 p(\boldsymbol{z})\mathrm{d}\boldsymbol{z}\mathrm{d}\boldsymbol{s}' - 1 + \frac{1}{2}\int \mathrm{d}\boldsymbol{s}'\end{aligned}$$

- It was shown in Tangkaratt et al. (2015) that

$$\mathrm{SCE}\left(\boldsymbol{s}' \mid \boldsymbol{z}\right) \geqslant \mathrm{SCE}\left(\boldsymbol{s}' \mid \boldsymbol{s}, a\right),$$

  and the equality holds if and only if the conditional independence holds.

- Sufficient dimensionality reduction can be performed by minimizing $\mathrm{SCE}\left(\boldsymbol{s}' \mid \boldsymbol{z}\right)$ with respect to $\boldsymbol{W}$:

$$\boldsymbol{W}^* = \underset{\boldsymbol{W} \in \mathbb{G}}{\operatorname{argmin}}\, \mathrm{SCE}\left(\boldsymbol{s}' \mid \boldsymbol{z}\right),$$

  where $\mathbb{G}$ denotes the Grassmann manifold, which is the set of matrices $\boldsymbol{W}$ such that $\boldsymbol{W}\boldsymbol{W}^\top = \boldsymbol{I}$ without redundancy in terms of the span.

Conditional Independence

- Employ the LSCDE method introduced in Chapter 10 to obtain an estimator $\widehat{p}(\boldsymbol{s}' \mid \boldsymbol{z})$ of conditional density $p(\boldsymbol{s}' \mid \boldsymbol{z})$.
- Then, SCE can be approximated as

$$\widehat{\mathrm{SCE}}(\boldsymbol{s}' \mid \boldsymbol{z}) = -\frac{1}{2M} \sum_{m=1}^{M} \widehat{p}(\boldsymbol{s}'_m \mid \boldsymbol{z}_m) = -\frac{1}{2} \widetilde{\boldsymbol{\alpha}}^\top \widehat{\boldsymbol{v}}$$

where

$$\boldsymbol{z}_m = \boldsymbol{W} \begin{pmatrix} \boldsymbol{s}_m \\ a_m \end{pmatrix} \quad \text{and} \quad \widehat{\boldsymbol{v}} = \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{\phi}(\boldsymbol{z}_m, \boldsymbol{s}'_m).$$

- $\boldsymbol{\phi}(\boldsymbol{z}, \boldsymbol{s}')$ is the basis function vector used in LSCDE and $\widetilde{\boldsymbol{\alpha}}$ is the LSCDE solution given by

$$\widetilde{\boldsymbol{\alpha}} = (\widehat{\boldsymbol{U}} + \lambda \boldsymbol{I})^{-1} \widehat{\boldsymbol{v}}.$$

Outline

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Dimensionality Reduction with SCE

- With the above SCE estimator, a practical formulation for sufficient dimensionality reduction is given by

$$\widehat{\boldsymbol{W}} = \underset{\boldsymbol{W}\in\mathbb{G}}{\operatorname{argmax}} S(\boldsymbol{W}), \text{ where } S(\boldsymbol{W}) = \widetilde{\boldsymbol{\alpha}}^\top \widehat{\boldsymbol{v}}.$$

- The gradient of $S(\boldsymbol{W})$ with respect to $W_{\ell,\ell'}$ is given by

$$\frac{\partial S}{\partial W_{\ell,\ell'}} = -\widetilde{\boldsymbol{\alpha}}^\top \frac{\partial \widehat{\boldsymbol{U}}}{\partial W_{\ell,\ell'}} \widetilde{\boldsymbol{\alpha}} + 2\frac{\partial \widehat{\boldsymbol{v}}^\top}{\partial W_{\ell,\ell'}} \widetilde{\boldsymbol{\alpha}}$$

- On the Grassmann manifold, the natural gradient (the projection of the ordinary gradient to the tangent space of the Grassmann manifold) gives the steepest direction.
- If the tangent space is equipped with the canonical metric $(\boldsymbol{W}, \boldsymbol{W}') = \frac{1}{2}\operatorname{tr}(\boldsymbol{W}^\top \boldsymbol{W}')$, the natural gradient at $\boldsymbol{W}$ is

$$\frac{\partial S}{\partial \boldsymbol{W}} \boldsymbol{W}_\perp^\top \boldsymbol{W}_\perp,$$

where $\boldsymbol{W}_\perp$ is the matrix such that $[\boldsymbol{W}^\top, \boldsymbol{W}_\perp^\top]$ is an orthogonal matrix.

Dimensionality Reduction with SCE

- The geodesic from $\boldsymbol{W}$ to the direction of the natural gradient over the Grassmann manifold can be expressed using $t \in \mathbb{R}$ as

$$\boldsymbol{W}_t = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{O} \end{bmatrix} \exp\left(-t \begin{bmatrix} \boldsymbol{O} & \frac{\partial S}{\partial \boldsymbol{W}} \boldsymbol{W}_\perp^\top \\ -\boldsymbol{W}_\perp \frac{\partial S}{\partial \boldsymbol{W}}^\top & \boldsymbol{O} \end{bmatrix} \right) \begin{bmatrix} \boldsymbol{W} \\ \boldsymbol{W}_\perp \end{bmatrix}.$$

- Then line search along the geodesic in the natural gradient direction is performed by finding the maximizer from $\{\boldsymbol{W}_t \mid t \geqslant 0\}$ (Edelman et al., 1998).
- Keep updating $\boldsymbol{W}$ until it converges. Final solution is normalized as

$$\widehat{p}\left(\boldsymbol{s}' \mid \boldsymbol{z}\right) = \frac{\widehat{\boldsymbol{\alpha}}^\top \boldsymbol{\phi}\left(\boldsymbol{z}, \boldsymbol{s}'\right)}{\int \widehat{\boldsymbol{\alpha}}^\top \boldsymbol{\phi}\left(\boldsymbol{z}, \boldsymbol{s}''\right) \mathrm{d}\boldsymbol{s}''},$$

where $\widehat{\alpha}_b = \max\left(0, \widetilde{\alpha}_b\right)$.

Outline

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Relation to Squared-Loss Mutual Information

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

- The above dimensionality reduction method minimizes SCE:

$$\mathrm{SCE}\left(\boldsymbol{s}' \mid \boldsymbol{z}\right) = -\frac{1}{2}\iint \frac{p\left(\boldsymbol{z}, \boldsymbol{s}'\right)^2}{p(\boldsymbol{z})}\mathrm{d}\boldsymbol{z}\mathrm{d}\boldsymbol{s}'.$$

- On the other hand, the dimensionality reduction method proposed in Suzuki and Sugiyama (2013) maximizes squared-loss mutual information (SMI):

$$\mathrm{SMI}\left(\boldsymbol{z}, \boldsymbol{s}'\right) = \frac{1}{2}\iint \frac{p\left(\boldsymbol{z}, \boldsymbol{s}'\right)^2}{p(\boldsymbol{z})p\left(\boldsymbol{s}'\right)}\mathrm{d}\boldsymbol{z}\mathrm{d}\boldsymbol{s}'.$$

- The essential difference between SCE and SMI is whether $p\left(\boldsymbol{s}'\right)$ is included in the denominator of the density ratio.
- SCE-based dimensionality reduction is expected to work better than SMI-based dimensionality reduction.

Outline

Artificial and Benchmark Datasets

The following dimensionality reduction schemes are compared:

- **None**: No dimensionality reduction is performed.
- **SCE (Section 11.2)**: Dimensionality reduction is performed by minimizing the least-squares SCE approximator using natural gradients over the Grassmann manifold (Tangkaratt et al., 2015).
- **SMI (Section 11.2.3)**: Dimensionality reduction is performed by maximizing the least-squares SMI approximator using natural gradients over the Grassmann manifold (Suzuki & Sugiyama, 2013).
- **True**: The "true" subspace is used (only for artificial datasets).

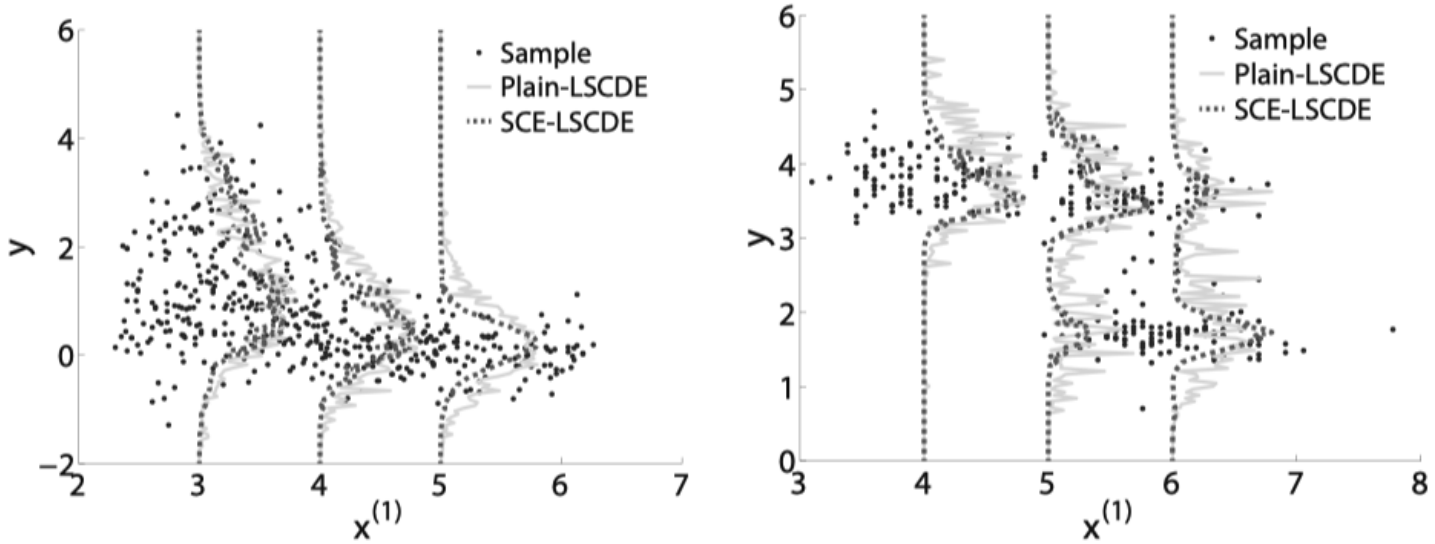
Artificial and Benchmark Datasets

After dimensionality reduction, the following conditional density estimators are run:

- **LSCDE (Section 10.1.3)**: Least-squares conditional density estimation (Sugiyama et al., 2010).
- **ϵKDE (Section 10.1.2)**: ϵ-neighbor kernel density estimation, where ϵ is chosen by least-squares cross-validation.

Conditional Density Estimation

- Input $\boldsymbol{x} = \left(x^{(1)}, \ldots, x^{(5)}\right)^\top$; Output: y.
- $x^{(1)}$ is relevant to predicting the output y; $x^{(2)}, \ldots, x^{(5)}$ are standard Gaussian noise.



(a) Bone mineral density (b) Old Faithful geyser

FIGURE 11.1: Examples of conditional density estimation by plain LSCDE and SCE-LSCDE.

Artificial Datasets

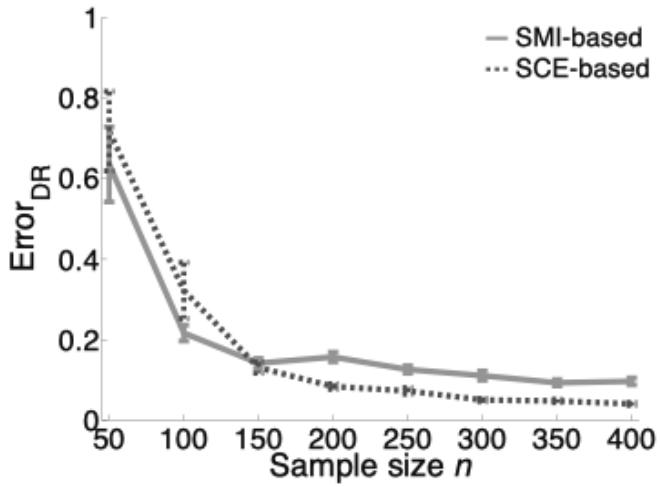
- Input $\boldsymbol{x} = \left(x^{(1)}, \ldots, x^{(5)}\right)^\top$; Output: y.
- Each element of $\boldsymbol{x}$ follows the standard Gaussian distribution and y is given by
  (a) $y = x^{(1)} + (x^{(1)})^2 + (x^{(1)})^3 + \varepsilon$,
  (b) $y = (x^{(1)})^2 + (x^{(2)})^2 + \varepsilon$,
  where $\varepsilon \sim N(0, (1/4)^2)$.
- Dimensionality reduction error:

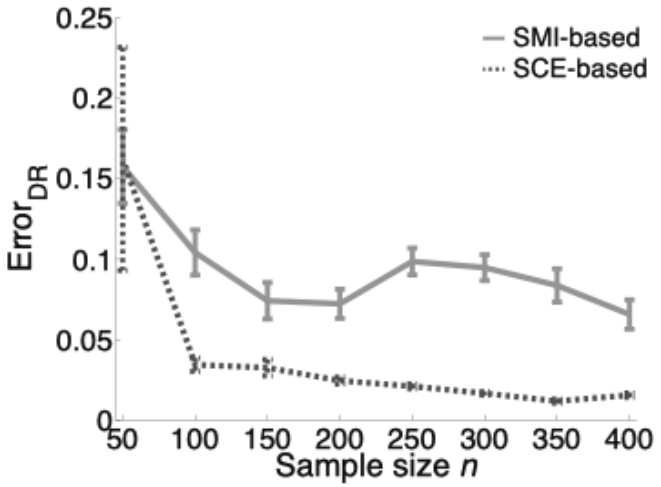
$$\mathrm{Error}_{\mathrm{DR}} = \left\| \widehat{\boldsymbol{W}}^\top \widehat{\boldsymbol{W}} - \boldsymbol{W}^{*\top} \boldsymbol{W}^* \right\|_{\mathrm{Frobenius}}$$

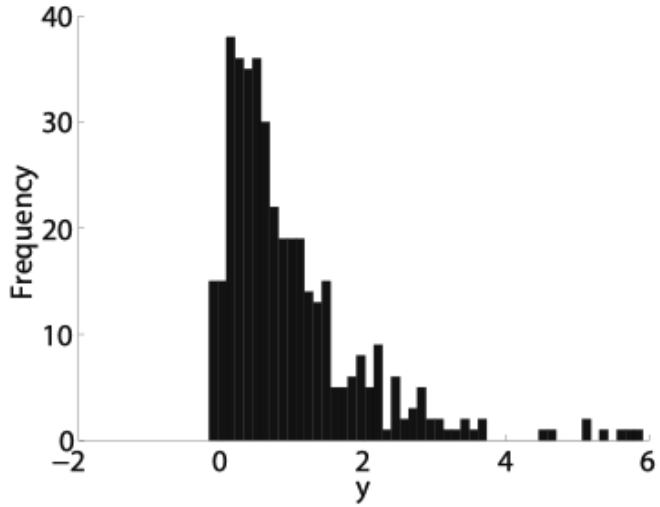
- Conditional density estimation error between true $p(y \mid \boldsymbol{x})$ and its estimate $\widehat{p}(y \mid \boldsymbol{x})$, evaluated by the squared loss:

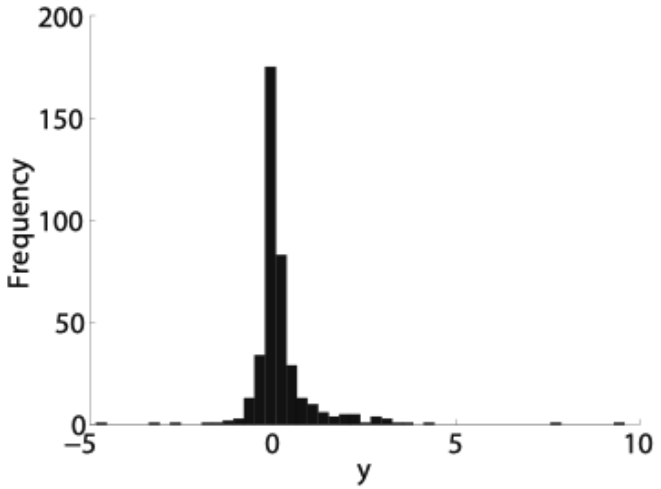
$$\mathrm{Error}_{\mathrm{CDE}} = \frac{1}{2n'} \sum_{i=1}^{n'} \int \widehat{p}\left(y \mid \widetilde{\boldsymbol{x}}_i\right)^2 \, \mathrm{d}y - \frac{1}{n'} \sum_{i=1}^{n'} \widehat{p}\left(\widetilde{y}_i \mid \widetilde{\boldsymbol{x}}_i\right)$$

Artificial Datasets









Artificial Datasets

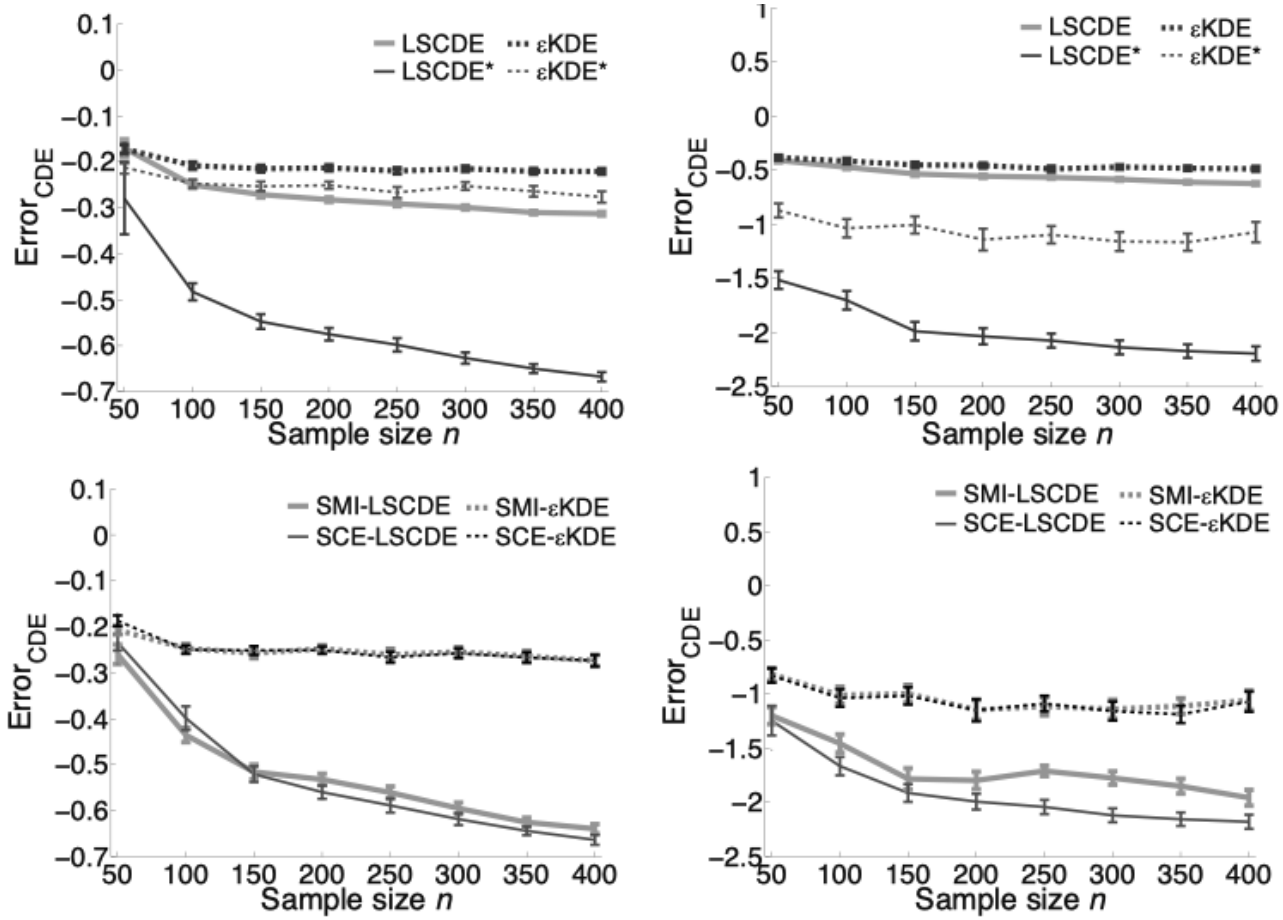


FIGURE 11.2: Top row: The mean and standard error of the dimensionality reduction error over 20 runs on the artificial datasets. 2nd row: Histograms of output $\{y_i\}_{i=1}^{400}$. 3rd and 4th rows: The mean and standard error of the conditional density estimation error over 20 runs.

UCI benchmark datasets

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

TABLE 11.1: Mean and standard error of the conditional density estimation error over 10 runs for various datasets (smaller is better). The best method in terms of the mean error and comparable methods according to the two-sample paired t-test at the significance level 5% are specified by bold face.

Dataset	$(d_{\mathbf{x}}, d_{\mathbf{y}})$	n	SCE-based		SMI-based		No reduction		Scale
			LSCDE	ϵKDE	LSCDE	ϵKDE	LSCDE	ϵKDE	
Housing	(13, 1)	100	**−1.73(.09)**	−1.57(.11)	**−1.91(.05)**	−1.62(.08)	−1.41(.05)	−1.13(.01)	×1
Auto MPG	(7, 1)	100	**−1.80(.04)**	**−1.74(.06)**	**−1.85(.04)**	**−1.77(.05)**	**−1.75(.04)**	−1.46(.04)	×1
Servo	(4, 1)	50	**−2.92(.18)**	**−3.03(.14)**	**−2.69(.18)**	**−2.95(.11)**	−2.62(.09)	−2.72(.06)	×1
Yacht	(6, 1)	80	**−6.46(.02)**	**−6.23(.14)**	−5.63(.26)	−5.47(.29)	−1.72(.04)	−2.95(.02)	×1
Physicochem	(9, 1)	500	**−1.19(.01)**	−0.99(.02)	**−1.20(.01)**	−0.97(.02)	**−1.19(.01)**	−0.91(.01)	×1
White Wine	(11, 1)	400	−2.31(.01)	**−2.47(.15)**	**−2.35(.02)**	**−2.60(.12)**	−2.06(.01)	−1.89(.01)	×1
Red Wine	(11, 1)	300	**−2.85(.02)**	−1.95(.17)	**−2.82(.03)**	−1.93(.17)	−2.03(.02)	−1.13(.04)	×1
Forest Fires	(12, 1)	100	**−7.18(.02)**	−6.93(.03)	−6.93(.04)	−6.93(.02)	−3.40(.07)	−6.96(.02)	×1
Concrete	(8, 1)	300	**−1.36(.03)**	−1.20(.06)	**−1.30(.03)**	−1.18(.04)	−1.11(.02)	−0.80(.03)	×1
Energy	(8, 2)	200	**−7.13(.04)**	−4.18(.22)	−6.04(.47)	−3.41(.49)	−2.12(.06)	−1.95(.14)	×10
Stock	(7, 2)	100	−8.37(.53)	**−9.75(.37)**	**−9.42(.50)**	**−10.27(.33)**	−7.35(.13)	−9.25(.14)	×1
2 Joints	(6, 4)	100	**−10.49(.86)**	−7.50(.54)	**−8.00(.84)**	−7.44(.60)	−3.95(.13)	−3.65(.14)	×1
4 Joints	(12, 8)	200	**−2.81(.21)**	−1.73(.14)	−2.06(.25)	−1.38(.16)	−0.83(.03)	−0.75(.01)	×10
9 Joints	(27, 18)	500	**−8.37(.83)**	−2.44(.17)	**−9.74(.63)**	−2.37(.51)	−1.60(.36)	−0.89(.02)	×100

UCI benchmark datasets

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

TABLE 11.2: Mean and standard error of the chosen subspace dimensionality over 10 runs for benchmark and robot transition datasets.

Dataset	$(d_{\mathbf{x}}, d_{\mathbf{y}})$	SCE-based		SMI-based	
		LSCDE	ϵKDE	LSCDE	ϵKDE
Housing	(13, 1)	3.9(0.74)	2.0(0.79)	2.0(0.39)	1.3(0.15)
Auto MPG	(7, 1)	3.2(0.66)	1.3(0.15)	2.1(0.67)	1.1(0.10)
Servo	(4, 1)	1.9(0.35)	2.4(0.40)	2.2(0.33)	1.6(0.31)
Yacht	(6, 1)	1.0(0.00)	1.0(0.00)	1.0(0.00)	1.0(0.00)
Physicochem	(9, 1)	6.5(0.58)	1.9(0.28)	6.6(0.58)	2.6(0.86)
White Wine	(11, 1)	1.2(0.13)	1.0(0.00)	1.4(0.31)	1.0(0.00)
Red Wine	(11, 1)	1.0(0.00)	1.3(0.15)	1.2(0.20)	1.0(0.00)
Forest Fires	(12, 1)	1.2(0.20)	4.9(0.99)	1.4(0.22)	6.8(1.23)
Concrete	(8, 1)	1.0(0.00)	1.0(0.00)	1.2(0.13)	1.0(0.00)
Energy	(8, 2)	5.9(0.10)	3.9(0.80)	2.1(0.10)	2.0(0.30)
Stock	(7, 2)	3.2(0.83)	2.1(0.59)	2.1(0.60)	2.7(0.67)
2 Joints	(6, 4)	2.9(0.31)	2.7(0.21)	2.5(0.31)	2.0(0.00)
4 Joints	(12, 8)	5.2(0.68)	6.2(0.63)	5.4(0.67)	4.6(0.43)
9 Joints	(27, 18)	13.8(1.28)	15.3(0.94)	11.4(0.75)	13.2(1.02)

Outline

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Remarks

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

- Coping with high dimensionality of the state and action spaces is one of the most important challenges in model-based reinforcement learning.
- The squared-loss conditional entropy (SCE) for dimensionality reduction can be estimated by LSCDE. This allowed us to perform dimensionality reduction and conditional density estimation simultaneously in an integrated manner.
- In contrast, SMI-based method yields a two-step procedure of first reducing the dimensionality and then the conditional density is estimated.
- SCE-based dimensionality reduction was shown to outperform the SMI-based method, particularly when output follows a skewed distribution.