Supplementary Material

Architectural Design and Training of DPN-SA

A deep stacked sparse autoencoder consists of multiple encoders and decoders to learn the identity function of the feature vectors, stacked to each other, with the setup of sparsity constraints in the hidden layers, whose neurons can be activated or not depending on the input.[1]

Let $\boldsymbol{X} = \{\boldsymbol{x}(\mathbf{1}), \boldsymbol{x}(\mathbf{2}), \dots, \boldsymbol{x}(\boldsymbol{N})\}$ represent the population of $\boldsymbol{x}(\boldsymbol{i})$ individuals. Each $\boldsymbol{x}(\boldsymbol{i}) \in \boldsymbol{R}^{\boldsymbol{d}}$ is a $\boldsymbol{d}$-dimensional vector, where $\boldsymbol{d}$ represents the number of pre-treatment covariates, i.e. the number of features. The DPN-SA uses up to $\boldsymbol{l}$ encoder and $\boldsymbol{l}$ decoder layers stacked together one after the other, with an additional linear layer at the end of the last decoder. The activation function of each encoder and decoder is the hyperbolic tangent layer $\boldsymbol{tanh}$. As $\boldsymbol{x}(\boldsymbol{i}) \in \boldsymbol{R}^{\boldsymbol{d}}$, the reconstructed output also needs to be $\boldsymbol{x}'(\boldsymbol{i}) \in \boldsymbol{R}^{\boldsymbol{d}}$. Supposing we set $\boldsymbol{l} = \mathbf{2}$, the sample input covariate vector $\boldsymbol{x}(\boldsymbol{i})$ is reconstructed in the forward propagation as follows:

$$e^{(1)}(i) = f\left(W^{(1)}x(i) + b^{(1)}\right) \quad (1)$$

$$e^{(2)}(i) = f\left(W^{(2)}e^{(1)}(i) + b^{(2)}\right) \quad (2)$$

$$d^{(1)}(i) = f\left(W^{(2)T}e^{(2)}(i) + b^{(3)}\right) \quad (3)$$

$$d^{(2)}(i) = f\left(W^{(1)T}d^{(1)}(i) + b^{(4)}\right) \quad (4)$$

$$x'(i) = W^{(3)}d^{(2)}(i) + b^{(5)} \quad (5)$$

where: $\boldsymbol{e}^{(\boldsymbol{l})}(\boldsymbol{i})$, $\boldsymbol{d}^{(\boldsymbol{l})}(\boldsymbol{i}), \boldsymbol{x}'(\boldsymbol{i})$ are the activations of the encoder, decoder and the reconstructed input of $\boldsymbol{l}^{\boldsymbol{th}}$ layer for the $\boldsymbol{i}^{\boldsymbol{th}}$ sample, respectively; $\boldsymbol{W}^{(\mathbf{1})}$ and $\boldsymbol{W}^{(\mathbf{2})}$ are the weight matrices of encoder 1 and encoder 2, having sizes $\boldsymbol{s1\ x\ d}$ and $\boldsymbol{s2\ x\ s1}$; $\boldsymbol{W}^{(\mathbf{2})\boldsymbol{T}}$ and $\boldsymbol{W}^{(\mathbf{1})\boldsymbol{T}}$ are the weight matrices of decoder 1 and decoder 2, having sizes $\boldsymbol{s1\ x\ s2}$ and $\boldsymbol{d\ x\ s1}$; $\boldsymbol{s}(\boldsymbol{l})$ denotes size or number of neurons of $\boldsymbol{l}^{\boldsymbol{th}}$ layer; $\boldsymbol{b}^{(1)}, \boldsymbol{b}^{(2)}, \boldsymbol{b}^{(3)}, \boldsymbol{b}^{(4)}$ are

the biases of encoder layer 1 and 2 and decoder layer 1 and 2 respectively and $\boldsymbol{f}(.)$ is the activation function ($\boldsymbol{tanh}$).

After reconstructing the a sample input feature $\boldsymbol{x(i)}$, the objective function of the sparse autoencoder $\boldsymbol{J_{sparse}(W,b)}$ has to be minimized:

$$J(W,b) = \left[\frac{1}{2N}\sum_{i=1}^{N}|x'_{W,b}(i) - x(i)|^2\right] + \frac{\lambda}{2N}\sum_{l=1}^{L}|W^{(l)}|_F^2 \tag{6}$$

$$J_{sparse}(W,b) = J(W,b) + \beta\sum_{\mathrm{j}=1}^{\mathrm{s2}}\mathrm{KL}(\rho||\rho_j) \tag{7}$$

where $\boldsymbol{\lambda}$ is the regularization parameter, $\boldsymbol{W^{(l)}}$ is the weight matrix corresponding to $\boldsymbol{l^{th}}$ layer of the network, $\boldsymbol{s2}$ is the number of neurons in the $\boldsymbol{2^{nd}}$ hidden layer, $\boldsymbol{\beta}$ is the weight of the sparsity penalty, $\boldsymbol{KL}$ is the Kullback-Leibler divergence, ρ is the sparsity constraint or sparsity parameter and ρ_j is the average activation of layer 2 (with s2 neurons), and the subscript $\boldsymbol{F}$ is the Frobenius norm (equivalent to the squared norm of the weight matrix). The $\boldsymbol{J_{sparse}(W,b)}$ is minimized using backpropagation, for $\boldsymbol{K_{sa}}$ number of epochs.

The DPN-SA is trained in two phases. In the first phase the sparse autoencoder is trained and optimized ($\boldsymbol{K_{sa}}$ epochs). The parameters ($\boldsymbol{W,b}$) are updated in each iteration by the Adam optimizer. The Adam is an optimization algorithm for stochastic gradient descent designed for training deep learning models.[2] After the sparse autoencoder has learnt the latent representation of the covariates, the decoder part is removed and a softmax classifier is attached to the end of the last encoder layer. The softmax classifier is trained for $\boldsymbol{K_c}$ number of epochs. The final network gives the estimation of the propensity score $\boldsymbol{\pi(x)}$. Algorithm 1 describes the two-phase procedures to obtain the final DPN-SA.

Algorithm 1. Training of the Deep Propensity Network using a Sparse Autoencoder (DPN-SA)

Input: Dataset batches B_b of random samples with assigned treatment T training set X_{train}, number of epochs K_s, K_c, learning rates lr_{SA}, lr_c for the sparse autoencoder and the *softmax* classifier, respectively.

Output: Propensity Scores PS_{out} for each sample $i = 1 \ldots N$.

Procedure:

Initialize the weights $W^{(l)}, b^{(l)}$ of the sparse autoencoder.

for iteration epochs = 1, 2, . . . K_s **do**

for dataset batches $= 1,2,..B$ in X_{batch}

Compute x'_{batch} using forward propagation algorithm.

Compute J_{sparse}.

Compute Gradient: $\nabla(J_{sparse})$

Optimize weights using Adam optimizer: $(W^{(l)}, b^{(l)}) \leftarrow Adam(X_{batch}, W^{(l)}, b^{(l)})$

end for

end for

Remove the decoder from the sparse autoencoder

Attach a *softmax* classifier to the last encoder

Initialize $W^{(c)}, b^{(c)}$ of the classifier

$PS_{out} \leftarrow$ Empty

for iteration epochs = 1, 2, . . . K_c **do**

for dataset batches $= 1,2,..B$ in X_{batch}

$(W^{(l)}, b^{(l)}, W^{(c)}, b^{(c)}) \leftarrow Adam(X_{batch}, W^{(l)}, b^{(l)}, W^{(c)}, b^{(c)})$

Get the propensity score $\pi(T \mid X_{batch})$ and add to PS_{out}

end for

end for

Detailed characteristics of the sample population

Table S1. The original IHDP dataset (https://github.com/vdorie/npci/tree/master/examples/ihdp_sim).

Variable	All patients (985)	T=1 (377; 38.2%)	T=0 (608; 61.8%)
bw	Avg = 1795.86701 Std = 457.219401	Avg = 1819.32361 Std = 438.809313	Avg = 1781.32361 Std = 468.76
b.head	Avg = 29.4277563 Std = 2.46631306	Avg = 29.4677563 Std = 2.32608452	Avg = 29.04 Std = 2.55
preterm	Avg = 6.98172589 Std = 2.67927742	Avg = 6.95755968 Std = 2.52188345	Avg = 6.996 Std = 2.77
Birth.o	Avg = 1.8964467 Std = 0.998697	Avg = 1.90185676 Std = 1.03831991	Avg = 1.896 Std = 0.97
nn.health	Avg = 99.9949239 Std = 15.8785717	Avg = 100.657825 Std = 15.9741196	Avg = 99.58 Std = 15.88

momage	Avg = 24.7715736 Std = 6.0335	Avg = 24.5862069 Std = 5.93211784	Avg = 24.88 Std = 6.09
Sex - male	499 (50.6%)	187(49.6%)	312(51.3%)
twin	97(9.84%)	39(10.34%)	58(9.5%)
b.marr	455(46.19%)	160(42.44%)	295(48.5%)
Mom.lths	394(40%)	162(42.99%)	232(38.1%)
Mom.hs	270(27.4%)	104(27.6%)	166(27.3%)
Mom.scoll	197(20%)	63(16.7%)	134(22.03%)
cig	346(35.1%)	131(34.7%)	215(35.5%)
first	542(55.02%)	177(46.9%)	365(60.1%)
booze	128(12.9%)	43(11.4%)	85(13.07%)
drugs	947(96.14%)	360(95.4%)	587(96.2%)
work.dur	566(57.5%)	218(57.8%)	348(57.8%)
prenatal	940(95.43%)	358(94.9%)	582(95.8%)
ark	128(12.99%)	48(12.7%)	80(13.19%)
ein	138(14.01%)	46(12.2%)	92(14.8%)
har	138(14.01%)	45(11.93%)	93(15.23%)
mia	100(10.15%)	44(11.67%)	56(9.29%)
pen	101(10.25%)	48(12.73%)	53(8.71%)
tex	137(13.9%)	49(12.99%)	88(14.38%)
was	131(13.29%)	51(13.5%)	80(13.1%)
momwhite	363(36.85%)	139(36.8%)	224(36.8%)
momblack	517(52.48%)	201(53.31%)	316(51.9%)
momhisp	105(10.65%)	31(8.2%)	74(12.1%)

Table S2. The processed IHDP dataset (https://www.fredjo.com/).

Variable	**All patients (747)**	**T=1 (139; 18.6%)**	**T=0 (608; 81.3%)**
X1	Avg = -1.20482E-11 Std = 1	Avg = -1.20482E-11 Std = 1	Avg = -1.20482E-11 Std = 468.76
X2	Avg = -2.12851E-10 Std = 1	Avg = -2.12851E-10 Std = 1	Avg = -2.12851E-10 Std = 1
X3	Avg = -2.74431E-10 Std = 1	Avg = -2.74431E-10 Std = 1	Avg = -2.74431E-10 Std = 1
X4	Avg = -8.03258E-12 Std = 1	Avg = -8.03258E-12 Std = 1	Avg = -8.03258E-12 Std = 1
X5	Avg = 2.67738E-11 Std = 1	Avg = 2.67738E-11 Std = 1	Avg = 2.67738E-11 Std = 1
X6	Avg = 1.20482E-11 Std = 1	Avg = 1.20482E-11 Std = 1	Avg = 1.20482E-11 Std = 1
X7	384 (51.4%)	72(51.7%)	212(34.8%)
X8	70(9.37%)	12(8.6%)	58(9.5%)
X9	389(52.07%)	94(57.6%)	295(48.51)
X10	272(36.4%)	40(28.7%)	232(38.15%)
X11	201(26.9%)	35(25.1%)	166(27.3%)
X12	164(21.9%)	30(21.5%)	134(22.03)
X13	268 (35.8%)	53(38.1%)	215(35.36%)
X14	346(46.8%)	80(57.5%)	266(43.75%)
X15	105(14.05%)	20(14.3%)	85(13.98%)
X16	717(95.9%)	130(9.3%)	587(96.5%)
X17	444(59.4%)	96(69.06%)	348(57.2%)

X18	720(96.3%)	138(99.2%)	582(95.72%)
X19	101(13.5%)	21(15.1%)	80(13.15%)
X20	101(13.5%)	9(6.47%)	92(15.13%)
X21	117(15.55%)	24(17.2%)	93(15.29%)
X22	61(8.1%)	5(3.5%)	56(9.21%)
X23	55(7.36%)	2(1.4%)	53(8.71%)
X24	96(12.8%)	8(5.75%)	88(14.47%)
X25	118(15.7%)	38(27.33%)	80(13.15%)

Table S3. The original Jobs dataset (https://users.nber.org/~rdehejia/data/.nswdata2.html; https://www.sciencedirect.com/science/article/abs/pii/S030440760400082X).

Variable	**LaLonde**	**PSID Sample**
Age	24.52(6.63)	34.85(10.44)
Education	10.27(1.7)	12.12(3.08)
Black	0.8(0.40)	0.25(0.43)
Hispanic	0.11(0.31)	0.03(0.18)
Married	0.16(0.37)	0.47(0.34)
No H.S Degree	0.48(0.41)	0.31(0.46)
Real Earnings in 1974	3631(6221)	19429(13407)
Real Earnings in 1975	3043(5066)	19063(13597)
Real Earnings in 1978	5455(6253)	21554(15555)
Real Earnings in 1979	…	…
Zero Earnings in 1974	0.45(0.5)	0.09(0.28)
Zero Earnings in 1975	0.4 (0.49)	0.1(0.3)
Experimental Impact (1978 earnings)	886(488)	…
Sample size	297 Treatment 425 Controls	2490 Controls

Table S4. The processed Jobs dataset (https://www.fredjo.com/).

Variable	**All patients (3212)**	**T=1 (297; 9.24%)**	**T=0 (2915; 90.75%)**
X1	Avg = 0.00012765 Std = 0.999874618	Avg = 0.000127646 Std = 0.999874618	Avg = 0.000127646 Std = 0.999874618
X2	Avg = -0.00099 Std = 0.999166151	Avg = -0.00099004 Std = 0.999166151	Avg = -0.00099004 Std = 0.999166151
X3	1202(37.4%)	238(80.1%)	964(33.03%)
X4	157(4.8%)	28(9.42%)	129(4.4%)
X5	2274(70.7%)	50(16.8%)	2224(76.2%)
X6	1323(41.1%)	217(73.06%)	1106(37.9%)
X7	Avg = -0.0004172 Std = 1.000487399	Avg = -0.00041719 Std = 1.000487399	Avg =-0.00041719 Std-dev =1.000487399
X8	Avg = 3.11333E-06 Std = 0.999783331	Avg =3.11333E-06 Std = 0.999783331	Avg =3.11333E-06 Std-dev =0.999783331
X9	Avg = -0.00047945 Std = 1.000054882	Avg =-0.00047945 Std = 1.000054882	Avg =-0.00047945 Std = 1.000054882
X10	Avg = -0.00033313	Avg =-0.00033313	Avg =-0.00033313

	Std = 1.00047291	Std = 1.00047291	Std = 1.00047291
X11	Avg = 0.001099004 Std = 0.999417803	Avg =0.001099004 Std = 0.999417803	Avg =0.001099004 Std = 0.999417803
X12	Avg = -0.00014633 Std = 1.000062289	Avg =-0.00014633 Std = 1.000062289	Avg =-0.00014633 Std = 1.000062289
X13	Avg = 0.000298879 Std = 1.000276622	Avg =0.000298879 Std = 1.000276622	Avg =0.000298879 Std = 1.000276622
X14	542(16.8%)	131(44.1%)	411(14.09%)
X15	538(16.74)	111(37.37%)	427(14.6%)
X16	Avg =0.00011208 Std = 0.99999744	Avg =0.00011208 Std = 0.99999744	Avg =0.00011208 Std = 0.99999744
X17	35(1.08%)	6(2.02%)	29(0.99%)

Supplementary References

1. Ng A. CS294A Lecture notes Sparse autoencoder.

   https://web.stanford.edu/class/cs294a/sparseAutoencoder_2011new.pdf

2. Kingma DP, Ba JL. Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. 2015.