Model matters with restricted Boltzmann machines

Andee Kaplan, Daniel Nordman, and Stephen Vardeman

Iowa State University
ajkaplan@iastate.edu

December 20, 2016

Slides available at http://bit.ly/kaplan-umass

What is this?

A restricted Boltzman machine (RBM) is an undirected probabilistic graphical model with

1. Discrete or continuous random variables
2. two layers - one hidden and one visible
3. conditional independence within a layer (Smolensky 1986)

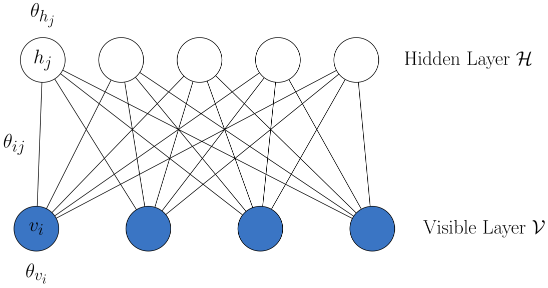

Figure 1: An example RBM, which consists of two layers. Hidden nodes are

How is it used?

- Supervised learning, specifically image classification

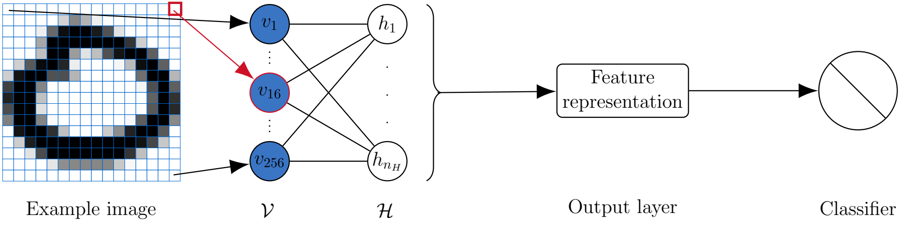

Figure 2: Image classification using a RBM. On the left, each image pixel comprises a node in the visible layer, $\mathcal{V}$. On the right, the output of the RBM is used to create features which are then passed to a supervised learning algorithm.

Joint distribution

- $\boldsymbol{x} = (h_1, \dots, h_{n_H}, v_1, \dots, v_{n_V})$ represents the states of the visible and hidden nodes in a RBM
- Each single "binary" random variable, visible or hidden, will take its values in a common coding set $\mathcal{C}$
  - Two possibilities for the coding set $\mathcal{C} = \{0, 1\}$ or $\mathcal{C} = \{-1, 1\}$.
- A parametric form for probabilities

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{\exp\left(\sum_{i=1}^{n_V}\sum_{j=1}^{n_H}\theta_{ij}v_i h_j + \sum_{i=1}^{n_V}\theta_{v_i}v_i + \sum_{j=1}^{n_H}\theta_{h_j}h_j\right)}{\gamma(\boldsymbol{\theta})} \tag{1}$$

where

$$\gamma(\boldsymbol{\theta}) = \sum_{\boldsymbol{x}\in\mathcal{C}^{n_H+n_V}} \exp\left(\sum_{i=1}^{n_V}\sum_{j=1}^{n_H}\theta_{ij}v_i h_j + \sum_{i=1}^{n_V}\theta_{v_i}v_i + \sum_{j=1}^{n_H}\theta_{h_j}h_j\right)$$

Deep learning

- Stacking layers of RBMs in a deep architecture
- Proponents claim the ability to learn "internal representations that become increasingly complex, which is considered to be a promising way of solving object and speech recognition problems" (Salakhutdinov and Hinton 2009, pp. 450).
- Treating a hidden layer of one RBM as the visible layer in a second RBM, etc.

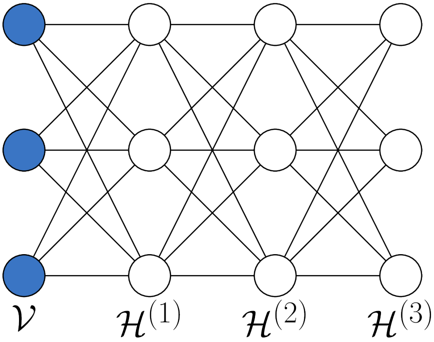

Figure 3: Three layer deep Boltzmann machine, with visible-to-hidden and hidden-to-hidden connections but no within-layer connections.

Why do I care?

1. The model properties are largely unexplored in the literature and
2. The commonly cited fitting methodology remains heuristic-based and abstruse (Hinton, Osindero, and Teh 2006)

We want to

1. Provide steps toward a thorough understanding of the model class and its properties from the perspective of statistical theory, and
2. Explore the possibility of a rigorous fitting methodology

Degeneracy, instability, and uninterpretability. Oh my!

The highly flexible nature of a RBM ($n_H + n_V + n_H * n_V$ parameters) makes at least three kinds of potential model impropriety of concern

1. *degeneracy*
2. *instability*, and
3. *uninterpretability*

> *A model should "provide an explanation of the mechanism underlying the observed phenomena" (Lehmann 1990; G. E. P. Box 1967).*

RBMs often

- fail to generate data with realistic variability and thus an unsatisfactory conceptualization of the data generation process (Li 2014)
- exhibit instability in the parameter space (Szegedy et al. 2013; Nguyen, Yosinski, and Clune 2014)

Near-degeneracy

Definition (Model Degeneracy)

There is a disproportionate amount of probability placed on only a few elements of the sample space, $\mathcal{C}^{n_H+n_V}$, by the model.

RBM models exhibit *near-degeneracy* when random variables in

$$Q_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{i=1}^{n_V}\sum_{j=1}^{n_H}\theta_{ij}v_i h_j + \sum_{i=1}^{n_V}\theta_{v_i}v_i + \sum_{j=1}^{n_H}\theta_{h_j}h_j,$$

have a mean vector $\boldsymbol{\mu}(\boldsymbol{\theta})$ close to the boundary of the convex hull of $\mathcal{T} = \{\boldsymbol{t}(x) : x \in \mathcal{C}^{n_H+n_V}\}$ (Handcock 2003), where

$$\boldsymbol{t}(\boldsymbol{x}) = \{v_1, \dots, v_{n_V}, h_1, \dots, h_{n_H}, v_1 h_1, \dots, v_{n_V}h_{n_H}\}$$

and

$$\boldsymbol{\mu}(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}}\boldsymbol{t}(\boldsymbol{X})$$

.

Instability

Definition (Instability)

Characterized by excessive sensitivity in the model, where small changes in the components of data outcomes, $\boldsymbol{x}$, lead to substantial changes in probability.

- Concept of model deficiency related to *instability* for a class of exponential families of distributions (Schweinberger 2011)
- For the RBM, consider how a data model might be expanded to incorporate more visibles
  - Necessary to grow the number of model parameters in a sequence $\boldsymbol{\theta}_{n_V} \in \mathbb{R}^{n_V + n_H + n_V * n_H}$, $n_V \geq 1$
  - May also arbitrarily expand the number of hidden variables used.

Unstable RBMs

Definition (S-unstable RBM)

A RBM model formulation is *S-unstable* if

$$\lim_{n_V \to \infty} \frac{1}{n_V} \text{ELPR}(\boldsymbol{\theta}_{n_V}) = \infty.$$

where

$$\text{ELPR}(\boldsymbol{\theta}_{n_V}) = \log \left[\frac{\max\limits_{(v_1, \dots, v_{n_V}) \in \mathcal{C}^{n_V}} P_{\boldsymbol{\theta}_{n_V}}(v_1, \dots, v_{n_V})}{\min\limits_{(v_1, \dots, v_{n_V}) \in \mathcal{C}^{n_V}} P_{\boldsymbol{\theta}_{n_V}}(v_1, \dots, v_{n_V})} \right] \tag{2}$$

S-unstable RBM model sequences are undesirable for several reasons - small changes in data can lead to overly-sensitive changes in probability.

One-pixel change

Consider the biggest log-probability ratio for a one-pixel (one component) change in data outcomes (visibles)

$$\Delta(\boldsymbol{\theta}_{n_V}) \equiv \max \left\{ \log \frac{P_{\boldsymbol{\theta}_{n_V}}(v_1, \dots, v_{n_V})}{P_{\boldsymbol{\theta}_{n_V}}(v_1^*, \dots, v_{n_V}^*)} \right\},$$

where $(v_1, \dots, v_{n_V})\&(v_1^*, \dots, v_{n_V}^*) \in \mathcal{C}^{n_V}$ differ by exactly one component

Result

Let $c > 0$ and let $ELPR(\boldsymbol{\theta}_{n_V})$ be as in (2) for an integer $n_V \geq 1$. If $\frac{1}{n_V} ELPR(\boldsymbol{\theta}_{n_V}) > c$, then $\Delta(\boldsymbol{\theta}_{n_V}) > c$.

If the quantity (2) is too large, then a RBM model sequence will exhibit large probability shifts for very small changes in the data configuration.

Tie to degeneracy

Define an arbitrary modal set of possible outcomes (i.e. set of highest probability outcomes) for a given $0 < \epsilon < 1$ as

$$M_{\epsilon,\boldsymbol{\theta}_{n_V}} \equiv \left\{ \boldsymbol{v} \in \mathcal{C}^{n_V} : \log P_{\boldsymbol{\theta}_{n_V}}(\boldsymbol{v}) > (1-\epsilon) \max_{\boldsymbol{v}^*} P_{\boldsymbol{\theta}_{n_V}}(\boldsymbol{v}^*) + \epsilon \min_{\boldsymbol{v}^*} P_{\boldsymbol{\theta}_{n_V}}(\boldsymbol{v}^*) \right\}$$

Result

For an S-unstable RBM model, and for any given $0 < \epsilon < 1$, $P_{\boldsymbol{\theta}_{n_V}}\left((v_1, \dots, v_{n_V}) \in M_{\epsilon,\boldsymbol{\theta}_{n_V}}\right) \to 1$ *holds as* $n_V \to \infty$.

- All probability will stack up on mode sets or potentially those few outcomes with the highest probability
- Proofs of results 1-2 can be found in (Kaplan, Nordman, and Vardeman 2016)

Uninterpretability

Definition (Uninterpretability)

Characterized by marginal mean-structure (controlled by main effect parameters $\theta_{v_i}, \theta_{h_j}$) not being maintained in the model due to dependence (interaction parameters θ_{ij}) (Kaiser 2007).

- Measure: the magnitude of the difference between
  1. Model expectations, $\mathrm{E}[\boldsymbol{X}|\boldsymbol{\theta}]$, and
  2. Expectations given independence, $\mathrm{E}[\boldsymbol{X}|\boldsymbol{\theta}^*]$, where $\boldsymbol{\theta}^*$ matches $\boldsymbol{\theta}$ for all main effects but otherwise has $\theta_{ij} = 0$ for $i = 1, \dots, n_V, j = 1, \dots, n_H$
- If $\left|\mathrm{E}\left[\boldsymbol{X}|\boldsymbol{\theta}\right] - \mathrm{E}\left[\boldsymbol{X}|\boldsymbol{\theta}^*\right]\right|$ is large then the RBM with parameter vector $\boldsymbol{\theta}$ is *uninterpetable*

RBM quantities to compare

$$\mathrm{E}\left[\boldsymbol{X}|\theta\right] = \sum_{\boldsymbol{x}\in\mathcal{C}^{n_H+n_V}} \boldsymbol{x} \frac{\exp\left(\sum\limits_{i=1}^{n_V}\sum\limits_{j=1}^{n_H}\theta_{ij}v_ih_j + \sum\limits_{i=1}^{n_V}\theta_{v_i}v_i + \sum\limits_{j=1}^{n_H}\theta_{h_j}h_j\right)}{\sum\limits_{\boldsymbol{x}\in\mathcal{C}^{n_H+n_V}} \exp\left(\sum\limits_{i=1}^{n_V}\sum\limits_{j=1}^{n_H}\theta_{ij}v_ih_j + \sum\limits_{i=1}^{n_V}\theta_{v_i}v_i + \sum\limits_{j=1}^{n_H}\theta_{h_j}h_j\right)}$$

$$\mathrm{E}\left[\boldsymbol{X}|\theta^*\right] = \sum_{\boldsymbol{x}\in\mathcal{C}^{n_H+n_V}} \boldsymbol{x} \frac{\exp\left(\sum\limits_{i=1}^{n_V}\theta_{v_i}v_i + \sum\limits_{j=1}^{n_H}\theta_{h_j}h_j\right)}{\sum\limits_{\boldsymbol{x}\in\mathcal{C}^{n_H+V=n_V}} \exp\left(\sum\limits_{i=1}^{n_V}\theta_{v_i}v_i + \sum\limits_{j=1}^{n_H}\theta_{h_j}h_j\right)}$$

Data coding to mitigate degeneracy

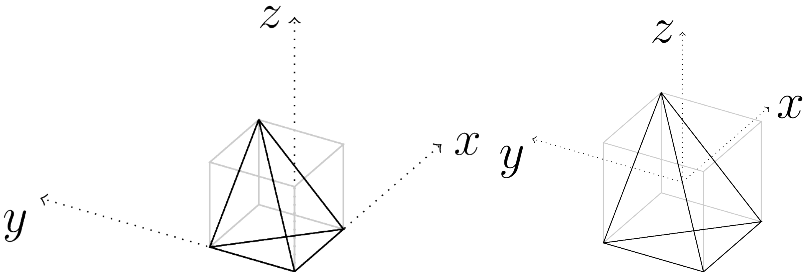

Figure 4: The convex hull of the "statistic space" $\mathcal{T} = \{(v_1, h_1, v_1 h_1) : v_1, h_1 \in \mathcal{C}\}$ in three dimensions for the toy RBM with one visible and one hidden node for $\mathcal{C} = \{0, 1\}$ (left) and $\mathcal{C} = \{-1, 1\}$ (right) data encoding.

The convex hull of $\mathcal{T} \subset \mathcal{C}^3$ does not fill the unit cube $[0, 1]^3$ (left), but does better with $[-1, 1]^3$ (right).

The center of the universe

- For the $\mathcal{C} = \{-1, 1\}$ encoding of hiddens $(H_1, \dots, H_{n_H})$ and visibles $(V_1, \dots, V_{n_V})$, the origin is the center of the parameter space.
- At $\boldsymbol{\theta} = \mathbf{0}$, RBM is equivalent to elements of X being distributed as iid Bernoulli$\left(\frac{1}{2}\right)$ $\Rightarrow$ No *near-degeneracy*, *instability*, or *uninterpretability*!

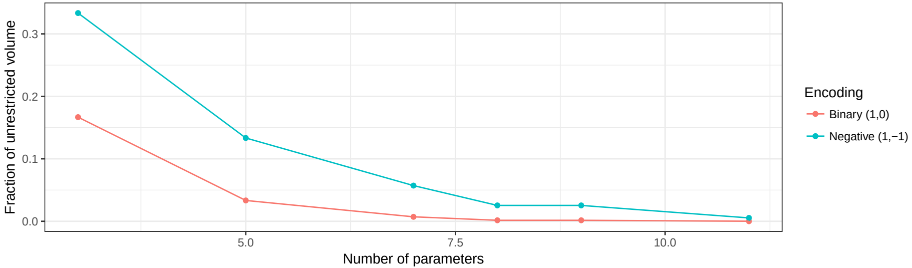

Figure 5: Relationship between volume of the convex hull of possible values of the RBM sufficient statistics and the cube containing it for different size models.

Manageable (a.k.a. small) examples

- To explore the behavior of the RBM parameters $\boldsymbol{\theta}$ as it relates to *near-degeneracy*, *instability*, and *uninterpretability*, consider models of small size
- For $n_H, n_V \in \{1, \dots, 4\}$, sample 100 values of $\boldsymbol{\theta}$
  1. Split $\boldsymbol{\theta}$ into $\boldsymbol{\theta}_{interaction}$ and $\boldsymbol{\theta}_{main}$, in reference to which sufficient statistics the parameters correspond to
  2. Allow the two types of terms to have varying average magnitudes, $||\boldsymbol{\theta}_{main}||/(n_H + n_V)$ and $||\boldsymbol{\theta}_{interaction}||/(n_H * n_V)$
  3. Average magnitudes vary on a grid between 0.001 and 3 with 24 breaks, yielding 576 grid points
- Calculate metrics of model impropriety, $\boldsymbol{\mu}(\boldsymbol{\theta})$, $\text{ELPR}(\boldsymbol{\theta})/n_V$, and the coordinates of $|\text{E}[\boldsymbol{X}|\boldsymbol{\theta}] - \text{E}[\boldsymbol{X}|\boldsymbol{\theta}^*]|$.
- In the case of *near-degeneracy*, classify each model as near-degenerate or "viable" based on the distance of $\boldsymbol{\mu}(\boldsymbol{\theta})$ from the boundary of the convex hull of $\mathcal{T}$

Simulation results

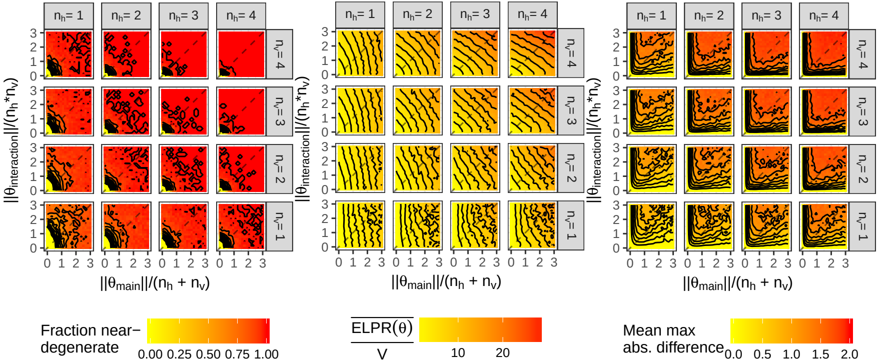

Figure 6: The fraction of models that were near-degenerate (left), the sample mean value of $\text{ELPR}(\boldsymbol{\theta})/n_V$ (middle), and the sample mean of the maximum component of the absolute difference between the model expectation vector, $\text{E}[\boldsymbol{X}|\boldsymbol{\theta}]$, and the expectation vector given independence, $\text{E}[\boldsymbol{X}|\boldsymbol{\theta}^*]$ (right).

Model fitting

1. Computational concerns
   - Fitting a RBM via maximum likelihood (ML) methods infeasible due to the intractibility of the normalizing term $\gamma(\boldsymbol{\theta})$
     - Ad hoc methods are used, aim to avoid this problem by using stochastic ML
     - employ a small number of MCMC draws to approximate $\gamma(\boldsymbol{\theta})$
2. With n_H large enough
   - Potential to re-create any distribution for the data (Le Roux and Bengio 2008; Montufar and Ay 2011; and Montúfar, Rauh, and Ay 2011)
     - The model for the cell probabilities that has the highest likelihood over *all possible model classes* is the empirical distribution
     - The RBM model ensures that this empirical distribution can be arbitrarily well approximated
   - When empirical distribution contains empty cells, fitting will chase parameters to ∞ in order to zero out the corresponding RBM cell probabilities

Bayesian methods

- Consider what might be done in a principled manner, testing on $n_V = n_H = 4$
- To avoid model impropriety, we want to avoid parts of the parameter space $\mathbb{R}^{n_V + n_H + n_V * n_H}$ that lead to *near-degeneracy*, *instability*, and *uninterpretability*.
  - Shrink $\boldsymbol{\theta}$ toward $\mathbf{0}$
    1. Specify priors that place low probability on large values of $||\boldsymbol{\theta}||$
    2. Shrink $\boldsymbol{\theta}_{interaction}$ more than $\boldsymbol{\theta}_{main}$
- Consider a test case with $n_V = n_H = 4$ and parameters given in in appendix
  - $\boldsymbol{\theta}$ chosen as a sampled value from a grid point in figure 6 with $< 5\%$ near-degeneracy (not near the convex hull of the sufficient statistics)
  - simulate $n = 5,000$ as a training set and fit the RBM using three Bayes methodologies

Fitting methodologies

1. *A "trick" prior (BwTPLV)*
   - Cancel out normalizing term in the likelihood
   - Resulting full conditionals of $\boldsymbol{\theta}$ are multivariate Normal
   - h_j are carried along as latent variables

$$\pi(\boldsymbol{\theta}) \propto \gamma(\boldsymbol{\theta})^n \exp\left(-\frac{1}{2C_1}\boldsymbol{\theta}'_{main}\boldsymbol{\theta}_{main} - \frac{1}{2C_2}\boldsymbol{\theta}'_{interaction}\boldsymbol{\theta}_{interaction}\right),$$

where $C_2 < C_1$ (Li 2014)

Fitting methodologies (cont'd)

2. *A truncated Normal prior (BwTNLV)*
   - Independent spherical normal distributions as priors for $\boldsymbol{\theta}_{main}$ and $\boldsymbol{\theta}_{interaction}$
     - $\sigma_{interaction} < \sigma_{main}$
     - *truncated* at $3\sigma_{main}$ and $3\sigma_{interaction}$, respectively
   - Simulation from the posterior using a geometric adaptive MH step (Zhou 2014)
   - h_j are carried along in the MCMC implementation as latent variables
3. *A truncated Normal prior and marginalized likelihood (BwTNML)*
   - Marginalize out $\boldsymbol{h}$ in $f_{\boldsymbol{\theta}}(\boldsymbol{x})$
   - Use the truncated Normal priors applied to the marginal probabilities for visible variables

Hyperparameters

Table 1: The values used for the hyperparameters for all three fitting methods. A rule of thumb is imposed which decreases prior variances for the model parameters as the size of the model increases and also shrinks $\boldsymbol{\theta}_{interaction}$ more than $\boldsymbol{\theta}_{main}$. The common C defining C_1 and C_2 in the BwTPLV method is chosen by tuning.

Method	Hyperparameter	Value
BwTPLV	C_1	$\frac{C}{n}\frac{1}{n_H+n_V}$
	C_2	$\frac{C}{n}\frac{1}{n_H*n_V}$
BwTNLV	σ^2_{main}	$\frac{1}{n_H+n_V}$
	$\sigma^2_{interaction}$	$\frac{1}{n_H*n_V}$
BwTNML	σ^2_{main}	$\frac{1}{n_H+n_V}$
	$\sigma^2_{interaction}$	$\frac{1}{n_H*n_V}$

Mixing

The truncated Normal method (2) and the marginalized likelihood method (3) are drawing from the same stationary posterior distribution for images.

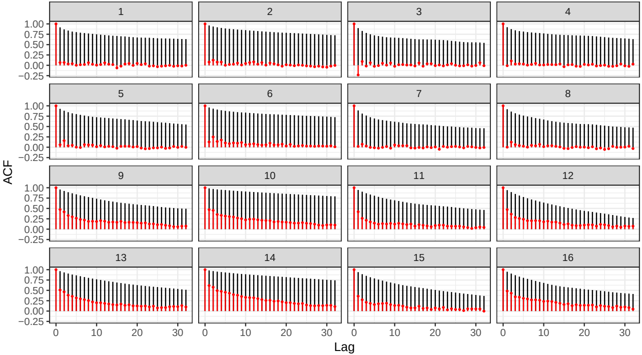

Figure 7: The autocorrelation functions (ACF) for the posterior probabilityies of all $2^4 = 16$ possible outcomes for the vector of 4 visibles assessed at multiple lags for each method with BwTNLV in black and BwTNML in red.

Effective sample size

- Overlapping blockmeans approach (Gelman, Shirley, and others 2011)
  - Crude estimate for the aysmptotic variance of the probability of each image
  - Compare it to an estimate of the asymptotic variance assuming IID draws from the target distribution

Table 2: The effective sample sizes for a chain of length $M = 1000$ regarding all 16 probabilities for possible vector outcomes of visibles. BwTNLV would require at least 4.7 times as many MCMC iterations to achieve the same amount of effective information about the posterior distribution.

Outcome	BwTNLV	BwTNML	Outcome	BwTNLV	BwTNML
1	73.00	509.43	9	83.47	394.90
2	65.05	472.51	10	95.39	327.35
3	87.10	1229.39	11	70.74	356.56
4	72.64	577.73	12	81.40	338.30
5	71.67	452.01	13	105.98	373.59
6	66.49	389.78	14	132.61	306.91
7	84.30	660.37	15	82.15	365.30
8	75.46	515.09	16	98.05	304.57

Posterior distributions of images

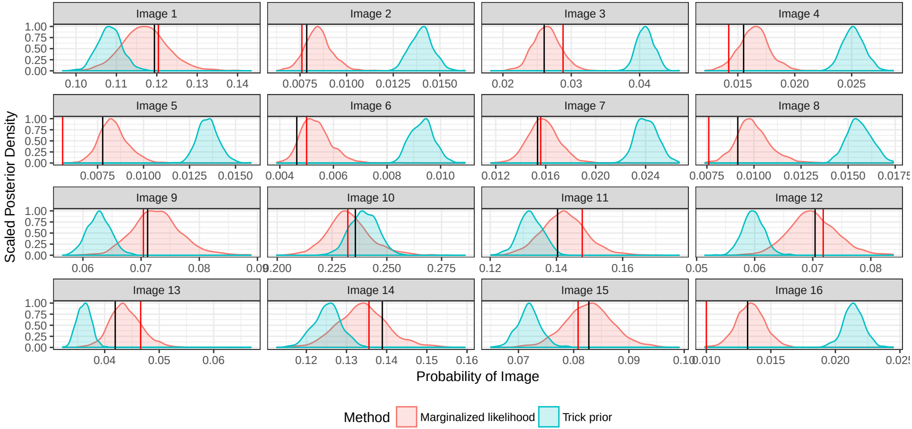

Figure 8: Posterior probabilities of $16 = 2^4$ possible realizations of 4 visibles using two of the three Bayesian fitting techniques, BwTPLV and BwTNML. Black lines show true probabilities of each vector of visibles based on the parameters used to generate the training data while red lines show the empirical distribution.

Wrapping up

- RBMs are thought to be useful for classification, but in the context of generative statistical models, poor fit due to *near-degeneracy*, *S-instability*, and *uninterpretability*
- Rigorous fitting methodology is difficult due to the dimension of the parameter space & size of the latent variable space
- For a RBM model with enough hidden variables, any distribution for the visibles can be approximated arbitrarily well (Le Roux and Bengio 2008; Montufar and Ay 2011; and Montúfar, Rauh, and Ay 2011)
  - The empirical distribution of a training set is the best fitting model for observed cell data
  - There can be no "smoothed distribution" achieved in a RBM model of sufficient size with a rigorous likelihood-based method

Skeptical that any model built using RBMs (i.e. deep Boltzmann machine) can achieve useful **prediction** or **inference** in a principled way without limiting the flexibility of the fitted model

Future work

1. Generalization of instability results (ongoing, see Kaplan, Nordman, and Vardeman 2016)
2. Image classification
   - Ensemble methods (super learners) using AdaBoost (Freund and Schapire 1995)
   - Decision theoretic based approach to approximating the likelihood ratio test for classification
3. Markov chain Monte Carlo methods for data with Markovian dependence
   - Spatial data
   - Network data

Thank you

- Slides – `http://bit.ly/kaplan-umass`
- Contact
    - Email – `ajkaplan@iastate.edu`
    - Twitter – `http://twitter.com/andeekaplan`
    - GitHub – `http://github.com/andeek`

Appendices

Appendix: Parameters used

Table 3: Parameters used to fit a test case with $n_V = n_H = 4$. This parameter vector was chosen as a sampled value of $\boldsymbol{\theta}$ that was not near the convex hull of the sufficient statistics for a grid point in figure 6 with $< 5\%$ near-degeneracy.

Parameter	Value	Parameter	Value	Parameter	Value
θ_{v1}	-1.1043760	θ_{11}	-0.0006334	θ_{31}	-0.0038301
θ_{v2}	-0.2630044	θ_{12}	-0.0021401	θ_{32}	0.0032237
θ_{v3}	0.3411915	θ_{13}	0.0047799	θ_{33}	0.0020681
θ_{v4}	-0.2583769	θ_{14}	0.0025282	θ_{34}	0.0041429
θ_{h1}	-0.1939302	θ_{21}	0.0012975	θ_{41}	0.0089533
θ_{h2}	-0.0572858	θ_{22}	0.0000253	θ_{42}	-0.0042403
θ_{h3}	-0.2101802	θ_{23}	-0.0004352	θ_{43}	-0.0000480
θ_{h4}	0.2402456	θ_{24}	-0.0086621	θ_{44}	0.0004767

References I

Freund, Yoav, and Robert E Schapire. 1995. "A Desicion-Theoretic Generalization of on-Line Learning and an Application to Boosting." In *European Conference on Computational Learning Theory*, 23–37. Springer.

G. E. P. Box, W. J. Hill. 1967. "Discrimination Among Mechanistic Models." *Technometrics* 9 (1). [Taylor & Francis, Ltd., American Statistical Association, American Society for Quality]: 57–71.

Gelman, Andrew, Kenneth Shirley, and others. 2011. "Inference from Simulations and Monitoring Convergence." *Handbook of Markov Chain Monte Carlo*, 163–74.

Handcock, Mark S. 2003. "Assessing Degeneracy in Statistical Models of Social Networks." Center for Statistics; the Social Sciences, University of Washington. `http://www.csss.washington.edu/`.

Hinton, Geoffrey E, Simon Osindero, and Yee-Whye Teh. 2006. "A Fast Learning Algorithm for Deep Belief Nets." *Neural Computation* 18 (7). MIT Press: 1527–54.

Kaiser, Mark S. 2007. "Statistical Dependence in Markov Random Field Models." *Statistics Preprints* Paper 57. Digital Repository @ Iowa State University. `http://lib.dr.iastate.edu/stat_las_preprints/57/`.

Kaplan, Andee, Daniel Nordman, and Stephen Vardeman. 2016. "A Note on the Instability and Degeneracy of Deep Learning Models." *Under Review*.

Le Roux, Nicolas, and Yoshua Bengio. 2008. "Representational Power of Restricted Boltzmann Machines and Deep Belief Networks." *Neural Computation* 20 (6). MIT Press: 1631–49.

Lehmann, E. L. 1990. "Model Specification: The Views of Fisher and Neyman, and Later Developments." *Statistical Science* 5 (2). Institute of Mathematical Statistics: 160–68.

Li, Jing. 2014. "Biclustering Methods and a Bayesian Approach to Fitting Boltzmann Machines in Statistical Learning." PhD thesis, Iowa State University; Graduate Theses; Dissertations. `http://lib.dr.iastate.edu/etd/14173/`.

Montufar, Guido, and Nihat Ay. 2011. "Refinements of Universal Approximation Results for Deep Belief Networks and

References II

Restricted Boltzmann Machines." *Neural Computation* 23 (5). MIT Press: 1306–19.

Montúfar, Guido F, Johannes Rauh, and Nihat Ay. 2011. "Expressive Power and Approximation Errors of Restricted Boltzmann Machines." In *Advances in Neural Information Processing Systems*, 415–23. NIPS.

Nguyen, Anh Mai, Jason Yosinski, and Jeff Clune. 2014. "Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images." *ArXiv Preprint ArXiv:1412.1897*. http://arxiv.org/abs/1412.1897.

Salakhutdinov, Ruslan, and Geoffrey E Hinton. 2009. "Deep Boltzmann Machines." In *International Conference on Artificial Intelligence and Statistics*, 448–55. AI & Statistics.

Schweinberger, Michael. 2011. "Instability, Sensitivity, and Degeneracy of Discrete Exponential Families." *Journal of the American Statistical Association* 106 (496). Taylor & Francis: 1361–70.

Smolensky, Paul. 1986. "Information Processing in Dynamical Systems: Foundations of Harmony Theory." DTIC Document.

Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. 2013. "Intriguing Properties of Neural Networks." *ArXiv Preprint ArXiv:1312.6199*. http://arxiv.org/abs/1312.6199.

Zhou, Wen. 2014. "Some Bayesian and Multivariate Analysis Methods in Statistical Machine Learning and Applications." PhD thesis, Iowa State University; Graduate Theses; Dissertations. http://lib.dr.iastate.edu/etd/13816/.