Bayesian factorization of joint categorical distributions for relational data and classical conditioning models

Ingmar Schuster  
Natural Language Processing Group, Department of Computer Science, University of Leipzig, Augustusplatz 10, 04109 Leipzig, Germany  
INGMAR.SCHUSTER.LINGUISTICS@GMAIL.COM

Patrick Jähnichen  
Natural Language Processing Group, Department of Computer Science, University of Leipzig, Augustusplatz 10, 04109 Leipzig, Germany  
JAEHNICHEN@INFORMATIK.UNI-LEIPZIG.DE

Abstract

We explore the problem of inferring latent structure in the joint probability of categorical variables by factorizing them into a latent representation for categories and a weight matrix that encodes a PMF, mapping these latent representations to the mass associated with seeing categories appear together. The prior for latent category representations is either the Chinese Restaurant Process (CRP) or a multivariate Gaussian with fixed dimensionality. We develop a simple and interpretable evaluation measure and discuss open questions.

1. Introduction

In recent years, several research groups have proposed Bayesian models for undirected graph data that infer latent structure. From this, the probability of edges in a held out dataset is inferred. Some models like (Miller et al., 2009) and (Palla et al., 2012) use the Indian Buffet Process (IBP) (Griffiths & Ghahramani, 2005) to infer latent variables for each node in the network. While (Miller et al., 2009) uses only the IBP, (Palla et al., 2012) proposes a hierarchy of the IBP and the Chinese Restaurant Process (CRP) (Aldous, 1985) to improve performance. (Lloyd et al., 2012) have used a Gaussian latent variable model for the representation of nodes. One of the contributions of this paper is the comparison of a discrete prior like the CRP to a continuous prior in terms of computational and prediction performance. We discuss the merits of both and open questions in terms of evaluation.

The second contribution of this paper is that the developed models apply to a more difficult type of dataset. In the context of networks, we model the probability of two nodes interacting using a Dirichlet compound Multinomial likelihood. Another perspective is that we provide a factorization of the joint probability of two categorical variables into a latent representation for each category and a matrix of weights. Together, these encode a probability mass function (pmf) mapping the latent representations $Z_a$ and $Z_b$ (for categories $a$ and $b$ respectively) to the probability mass associated with seeing $a$ occur together with $b$. Finally, a third perspective is that we model Pavlovian conditioning like the learning rules of Rescorla-Wagner equations (Rescorla & Wagner, 1972). However, while in the original Rescorla-Wagner equations only predictions for previously seen combinations of conditioned stimulus (CS) and unconditioned stimulus (US) are possible, the models presented here can make predictions for previously unseen combinations of CS and US.

We evaluate the mixing properties of our sampler for CRP and multivariate Gaussian latent variable models when applied to the NIPS coauthorship dataset originally compiled for (Globerson et al., 2007) and adjective-noun pairs extracted from the Wacky09 corpus (Baroni et al., 2009).

2. Factorization of Joint Probability of categorical variables

We would like to uncover latent structure in the joint probability of categorical variables. Joint categorical and multinomial distributions of high dimensionality occur frequently. For example, words occurring together in text, coauthors of scientific publications and stimuli occurring together in a real world environment can be interpreted as being generated by a joint distribution of categorical variables. Similarly to (Miller et al., 2009), (Palla et al., 2012), and (Lloyd et al., 2012), we propose a factorization of the joint probability of categorical variables into a latent representation for categories and a weight matrix that encodes a PMF, mapping these latent representations to the mass associated with seeing categories appear together. The prior for latent category representations is either the Chinese Restaurant Process (CRP) or a multivariate Gaussian with fixed dimensionality. We develop a simple and interpretable evaluation measure and discuss open questions in terms of evaluation.

Copyright 2014 by the author(s).
2012) and (Lloyd et al., 2012), we assume both a latent representation $Z_i$ for each category $i$ as well as a latent function $pmf(\cdot, \cdot)$ mapping latent representations $Z_i$ and $Z_j$ (for categories $i$ and $j$, respectively) to a probability mass. However, (Miller et al., 2009), (Palla et al., 2012) and (Lloyd et al., 2012) all take $pmf$ to represent the probability of a link being present between two network nodes $i$ and $j$ (network nodes correspond to categories in this paper). We take it to be the probability, at any point in time, to see a sample $(i, j)$ as opposed to a sample $(a, b)$ where either $a \neq i$ or $b \neq j$. Then roughly the probability of seeing $(i, j)$ is

$$P((i, j)) \approx \frac{pmf(i, j)}{\sum_{a,b} pmf(a, b)} \tag{1}$$

where $a, b$ range over all categories. While (Lloyd et al., 2012) notably points out that their approach could be used for likelihoods similar to ours, this is true of the approaches by (Miller et al., 2009) and (Palla et al., 2012) as well. However, none of these papers presented experimental results with other likelihoods, even though the datasets used for evaluation are often more naturally modeled as a joint of categorical variables, as we will argue in section 6

3. Model

3.1. General approach

We model the probability mass function $pmf(i, j)$ as

$$pmf(i, j) = \zeta(Z_i^T W Z_j) \tag{2}$$

where $W \in \mathbb{R}^{d \times d}$ is a square weight matrix, all latent category representations $Z_i$ have dimensionality $d$ and $\zeta$ is the element wise softplus introduced as an activation function in Neural Networks (Dugas et al., 2001).

$$\zeta: \mathbb{R} \rightarrow \mathbb{R}^+$$

$$x \mapsto \log(1 + \exp(x))$$

It is the indefinite integral of the logistic function $1/1 + \exp(-x)$ and ensures that we only assign positive probability masses. Because the second derivative of the softplus is positive, $\zeta$ it is strictly convex, leading to favorably convergence properties.

For a data matrix $C$ containing counts of observed samples $(i, j)$ in row $i$ and column $j$, our likelihood model is

$$\text{vec } C \sim DCM \left( \text{vec } \zeta(Z^T W Z) \right) \tag{3}$$

where $\text{vec}$ represents vectorization of a matrix. The columns of $Z$ contain latent variables for categories ($Z = [Z_a Z_b \ldots]$). Note the difference in notation from (Miller et al., 2009) and possibly other papers, where $Z$ contains the latent variables for categories (network nodes) in rows instead of columns. $DCM$ is the Dirichlet compound Multinomial distribution, i.e. a draw from a Dirichlet and a consecutive draw from a multinomial where the Dirichlet draw is integrated out.

$$P(C \mid \alpha) = \frac{\Gamma(A)}{\Gamma(N + A)} \prod_{k=1}^{K} \frac{\Gamma(n_k + \alpha_k)}{\Gamma(\alpha_k)} \tag{4}$$

where $\alpha_k$ is the parameter adjusting the prior probability for seeing category $C_k$, $N$ is the number of draws from the multinomial, $A = \sum_k \alpha_k$ and $n_k$ is the number of draws for category $k$.

3.2. Discrete nonparametric and Continuous parametric Latent Variable Models

To our knowledge, no previous paper has compared the convergence and speed properties of discrete and continuous latent variable spaces for the representation of categories. While (Lloyd et al., 2012) uses Gaussian latent variables to represent categories, (Miller et al., 2009) uses the Indian Buffet Process (IBP) (Griffiths & Ghahramani, 2005). (Palla et al., 2012) uses a hierarchical combination of the IBP and the Chinese Restaurant Process (CRP) (Aldous, 1985) in order to improve prediction performance. The current paper compares continuous (Gaussian) priors with a fixed number of components with the nonparametric CRP, a latent class prior learning the number of classes from data. We show that a continuous prior has advantages over the CRP both in its ability to fit the data as well as computation time. This however comes at the cost of having a fixed number of components which has to be set a priori.

The complete likelihood data model then spells out as follows. In the case of putting a CRP prior on the latent variables representing categories, we have

$$Z^{CRP} \sim CRP(\alpha_{CRP})$$

where $\alpha_{CRP}$ is the CRPs concentration parameter governing the probability of creating a new latent class. Note that we take draws from the CRP to be binary column vectors which have a 1 entry at the latent class corresponding to a category and 0 everywhere else. In the case of a multivariate Gaussian prior on latent variables representing categories we have

$$Z_i^{N} \sim N(0, \sigma^2_Z I)$$

for the latent variable of category $i$. The covariance matrix is diagonal, i.e. there are no dependencies between components of the multivariate Gaussian. Finally, using either of the priors on $Z$, the remaining likelihood model is

$$W \sim N(0, \sigma^2_W I)$$

$$\text{vec } C \sim DCM \left( \text{vec } \zeta(Z^T W Z) \mid \alpha_{DCM} \right)$$
Bayesian factorization of joint categorical distributions

Here, $W$ is a weight matrix. Again, the covariance matrix for the Gaussian prior on $W$ is diagonal. We put a symmetric parameter of $\alpha_{DCM}/K$ on each component of the DCM, where $K$ is the number of seen category pairings. Unseen pairings had a prior probability of $\alpha_{DCM}$, which we updated in sampling with the sum over all probability masses for unseen pairings.

3.3. Parallels of Gaussian LV model with Variational Inference approximations

When approximating the joint posterior of latent variables given the data, two main techniques are prevalent throughout the literature, sampling and variational methods. Looking at variational methods, often the mean field assumption applies (cf. (Wainwright & Jordan, 2008)), i.e. dependencies between latent variables are broken and statistical independence is assumed for the governing variational parameters of the approximating distribution.

Both the CRP and IBP are nonparametric priors on infinite dimensional objects that describe the discrete allocation of categories to classes. In the CRP case each category is assigned to exactly one class whereas in the IBP case, each component can belong to many classes. For both priors, the number of latent classes is unbounded. However, the flexibility of the normal distribution allows our model to fit the data even with a fixed number of dimensions. We break the model induced dependence of components a priori by using a diagonal covariance matrix for the prior on the latent variables corresponding to categories.

This leads to a fairly simple and high-performance sampling algorithm that achieves comparable results as when using more complex prior distributions but is orders of magnitude faster.

4. Related Work

We review relational models that try to predict the probability of a link between two nodes in a network where nodes correspond to categories in our approach. In these models, the data matrix $C$ only contains zeros and ones, depending on whether or not a link exists between two nodes. Then, given the probability $l_{i,j}$ of a link between nodes $i$ and $j$, the likelihood of the complete dataset simply is $P(C \mid Z, W) = \prod_{i,j} \left(C_{i,j}^{l_{i,j}} + (1 - l_{i,j})^{1-C_{i,j}} \right)$ where $C_{i,j}$ is the component in $C$ encoding whether a link exists between $i$ and $j$.

The approach by (Miller et al., 2009) is closest to ours, the difference being that it uses the IBP as the prior on $Z$ while we use the CRP or a Gaussian prior, respectively. The probability of a link between nodes $i$ and $j$ is $l_{i,j} = f_{\sigma}(Z_i^T W Z_j)$, where $f_{\sigma}$ is the logistic or probit function.

(Palla et al., 2012) use a hierarchic combination of the IBP and CRP, called the Infinite Latent Attributes (ILA) model, as latent variables representing network nodes. Each node (category) is assigned a binary feature vector containing ones if a node exhibits a latent attribute and zero else. If a node possesses a latent attribute, it might belong to one of several subclusters of the attribute. (Palla et al., 2012) put a CRP prior on subclusters. The probability of a link between nodes $i$ and $j$ is $l_{i,j} = f_{\sigma}(\sum_m Z_{i,m} Z_{j,m} w_{i,m}^{(m)} + s)$, where $f_{\sigma}$ is the logistic function, $w_{i,m}$ is a weight matrix specifically for the $m$th binary feature, $c_{i,m}^{(m)}$ is the subcluster assignment for node $i$ in feature $m$, $s$ is a bias term and $m$ ranges over all binary features. In the ILA, only features that are set to 1 for both nodes (categories) influence the likelihood of a link between them, because $Z_{i,m} Z_{j,m}$ is zero if either of the features is zero.

Finally, (Lloyd et al., 2012) use a Gaussian Process plus logistic function approach to model the probability of a link $l_{i,j}$ from latent variables $Z_i$ and $Z_j$. They construct a custom kernel function based on the RBF kernel that ensures that the symmetry properties of a network (i.e. an undirected graph) are met. While (Lloyd et al., 2012) use a uniform prior over the unit interval for latent variables $Z_i$ in their theoretical section, during sampling a multivariate Gaussian with diagonal covariance matrix is used just like in our approach.

5. Inference

We used a Markov Chain Monte Carlo approach to do inference in our model (Andrieu et al., 2003). In the CRP case, since we have a finite number of latent classes for finite datasets almost surely, $Z$ can always be stored as a finite matrix. We did not do inference on hyperparameters.

Sample $Z$ given $W$: In every iteration, first the latent variable representation $Z$ for categories is sampled given $W$. In the case of CRP variables, we used Gibbs sampling on $Z$. As (Miller et al., 2009) points out, the main difficulty arises when computing the likelihood of adding a new class is to the representation of $Z$. Like (Miller et al., 2009), we use a Monte Carlo approximation of the likelihood by repeatedly sampling the weights corresponding to the new class and averaging over likelihoods.

In the case of Gaussian variables, each component of $Z$ is slice-sampled (Neal, 2003). This simplistic approach is possible because we assumed a diagonal covariance matrix (i.e. zero covariances).

Sample $W$ given $Z$: Each component of the weight matrix $W$ is slice-sampled because again we assumed a diagonal covariance matrix.
5.1. Sequential Initialization

We found both models to be extremely sensitive to initialization. Here, we adopted and improved an very useful idea described in (Palla et al., 2012): Initially two categories are added to the model and some MCMC iterations (the authors suggest three) are run on the corresponding part of the data. Then all remaining categories are added one after another, running three iterations of MCMC after adding a category. As (Palla et al., 2012) points out, initially the sampler is very fast due to the small number of categories. Also, because fewer categories mean fewer local optima, the sampler reaches a high probability region of the parameter space more easily.

In our sampler, we added more than one category at a time. Also, we ran two instead of three MCMC iterations. The most important change however is the following: in datasets where some categories occur together extremely frequently while other pairs of categories occur less often, adding new categories results in very strong changes to the likelihood. We handle this problem by applying the sequential initialization scheme suggested by (Palla et al., 2012), but with counts of co-occurring categories normalized so that the lowest non-zero count is 1. After all categories have been added, in each following initialization step the counts of category co-occurrences are multiplied by a constant factor until the original data matrix is recovered.

The problem that this scheme solves does not occur in (Palla et al., 2012), because their task is not factorization of joint categorical distributions, but prediction of links between network nodes. Consequently, their data matrix only has entries that are either 0 or 1, while our data matrix can contain any value in \( \mathbb{N}_0^+ \).

6. Experiments

In the relational model literature, many datasets that are actually joint probabilities of categorical variables have been modified to fit into a network model. For example the NIPS dataset (Globerson et al., 2007) containing coauthorship information was used in all of (Palla et al., 2012; Lloyd et al., 2012; Miller et al., 2009) by identifying authors with nodes in an undirected graph. An edge was present (1) if two authors wrote any non-zero number of papers together, otherwise it was taken as not present (0). This is a strong simplification of the coauthorship information in the original dataset.

In a social network setting, it might not be of much interest which person knows which other person, which is the task in inferring network edges. A much more interesting question could be how often two persons interact with each other, for example write messages to each other, because this is a better measure of how closely two people relate. Our models can account for this setting by identifying people with categories and interactions between people (such as sent messages) as observations from a joint categorical (or multinomial) distribution.

We evaluated the models in terms of test log likelihood on a held out dataset. The second evaluation measure used is the mean and absolute differences between the true probability of seeing an observation from the test set and the probability a model assigns to that observation, i.e. the set \( \Delta = \{ | P_{\text{model}}((i,j)) - P((i,j)) | : i, j \in \text{Testset} \} \). We computed mean and standard deviation of these for both the CRP and Gaussian model and conducted Welch’s \( t \)-test to assess whether differences between the models where significant (Welch, 1947). Welch’s test does not assume equal variances for two populations as opposed to the standard \( t \)-test.

6.1. Types of held out data

There are two sensible ways of holding out data depending on the problem.

**Held out observations** When using the model to predict future observations from the joint categorical or just to build a more compact representation of it, test data might consist of some observations withheld from the count matrix (i.e. if your dataset contains 10 co-occurrences of “aggressive” and “dog”, 8 of these go into the training data, 2 into the test dataset)

![Figure 1. Holding out observations](image)

**Held out pairings** Some set \( H \) of pairings of categories are held out. The training dataset does not contain counts for these, the test dataset contains only counts for these (treat as missing data). Obviously, this is the more difficult of the two schemes of holding out data. However it is also the more interesting scheme, as it can be interpreted as the transfer of knowledge from the training set to unseen parts of the joint categorical. When using this evaluation scheme, the models developed in this paper can be seen as an instance of a Rescorla-Wagner-type classical conditioning model with the new capability of making predictions in situations that the model has not been trained on (Rescorla

```latex
\begin{align*}
\text{Training} & \quad \text{Test} \\
\text{Training} & \quad \text{Test}
\end{align*}
```

Figure 1. Holding out observations
Bayesian factorization of joint categorical distributions

& Wagner, 1972).

6.2. NIPS coauthorship dataset

The NIPS coauthorship dataset was originally compiled for (Globerson et al., 2007). This dataset contains a list of all papers and authors from the NIPS conferences 1-17. Like (Miller et al., 2009) we used the 234 authors who had published with the most other people. Instead of reducing the dataset to the information whether or not two authors had written a paper together (binary values), we looked at the categorical distribution where authors are identified with categories. Each paper two authors \(i\) and \(j\) published together was taken as an observation from the underlying joint categorical distribution.

The Gaussian model was very fast to mix when evaluating it with held out observations, see Figure 3. This is no surprise when considering that for the Gaussian model we used slice sampling of each component of a latent variable, whereas in the CRP model we had to fall back to Gibbs sampling. Another aspect here is that computation time per sample was much better for the Gaussian model (see Table 1). However, a possible reason for this might not be that slice sampling was much faster, but simply that in the CRP model for each category the likelihood of having a previously unseen latent class had to be estimated using Monte Carlo integration over the corresponding entries in \(W\). This was not the case for the Gaussian model which had fixed dimensionality.

The differences between the true and model probability of seeing a category pairing where significantly better for the Gaussian model than for the CRP model in the case of held out observations (see Table 2). For this dataset, as few as three latent classes where inferred for the CRP model. Considering that there was much more data (and even a slightly larger number of categories) this seems to hint at the CRP model being a misfit for the task. The small number of classes lead to computation time per sample that was close to or better than for the Gaussian model.

Especially interesting is the difference in training and test log likelihood for this dataset. The training log likelihood of the Gaussian model quickly outperforms that of the CRP model in both schemes of held out data. Still, for held out pairings of categories, the test log likelihood is better for the CRP model in every single sample (see Figure 4). One possible reason for this is that the dataset contains only about 1600 observations for \(234^2\) potential category pairings. Since the CRP model, as opposed to the Gaussian model, seems to gravitate towards a distribution that is close to the uniform distribution, it can closely approximate the little data available. The situation is slightly different for the next dataset.

6.3. WaCky adjective-noun cooccurrence dataset

Adjective-noun pairs extracted from the WaCky09 corpus (Baroni et al., 2009).

Again, the differences between the true and model probability of seeing a category pairing where significantly better for the Gaussian model than for the CRP model in the case of held out observations (see Table 2). For this dataset, as few as three latent classes where inferred for the CRP model. Considering that there was much more data (and even a slightly larger number of categories) this seems to hint at the CRP model being a misfit for the task. The small number of classes lead to computation time per sample that was close to or better than for the Gaussian model.

Particularly puzzling for this dataset is the discrepancy in
the intuitive quality of model fits and the data log likelihoods. For both training and test log likelihood, the CRP model is clearly better. However, the intuition when comparing true and model probability of a category pair in Figure 8 is that the Gaussian model approximates the data much better. Like for the NIPS dataset there are clear artifacts for the CRP model.

Some open questions have been posed by this paper. For one, there is a discrepancy in intuitive quality of fit and evaluation measures which became apparent when comparing CRP and Gaussian models of latent variables for categories. We are not sure how this mismatch is to be resolved. Also, better test log likelihood did not completely coincide with absolute differences in category pair probability (although these two evaluation measures do not contradict each other).

For future work, several opportunities exist. One would be to enable fits of our Gaussian model for the streaming data setting. (Welling & Teh, 2011) explores a sampling scheme using stochastic gradient langevin dynamics with artificial noise in order to achieve a streaming data fit for continuous distributions that could be adaptable to our Gaussian model. This would be especially attractive for the interpretation of our model as a classical conditioning with transfer learning. If it is possible to fit our model with artificial noise in order to achieve a streaming data setting, this closely resembles the learning environment present in classical conditioning, where combinations of conditioned and unconditioned stimuli are presented incrementally over time.

Another interesting area for future work is to remove the fixed dimensionality constraint of our Gaussian model. This would be especially attractive for continuous distributions that could be adaptable to our type of data in a straightforward way.

We provided interpretations of our models as a factorization of the probability of interaction for two nodes in a network as well as a generalization of Rescorla-Wagner-type models of classical conditioning (Rescorla & Wagner, 1972). While in the original Rescorla-Wagner equations only allow predictions for previously seen combinations of stimuli, our models can make predictions for previously unseen combinations. Finally, we developed a simple and interpretable evaluation measure (absolute differences in category pair probability).

We evaluated our models on the NIPS coauthorship dataset originally compiled for (Globerson et al., 2007) and adjective-noun pairs extracted from the Wacky09 corpus (Baroni et al., 2009).

7. Discussion and Future Work

In this paper, we explored the problem of inferring latent structure in the joint probability of categorical variables. Our model factorized the original joint distribution into a latent representation for categories and a weight matrix that encodes a PMF. The prior for latent category representations was either the Chinese Restaurant Process (CRP) or a multivariate Gaussian with fixed dimensionality. Our models apply to more complex datasets than the models developed for predicting edges in an undirected graph. However, models developed for the latter scenario can often be adapted to our type of data in a straightforward way.

Table 1. Statistics of inference on the NIPS dataset

| Held out type | Model  | Dimens. | sec/sample | probabil. dist. | dist | test ll | Pll |
|---------------|--------|---------|------------|-----------------|------|---------|-----|
| Observations  | CRP    | 21.7    | 996        | 0.0025 ± 0.0017 | 1.82 e − 07 | −5625.08 | 6.55 e − 39 |
|               | Gaussian | 6       | 428        | 0.0019 ± 0.0014 | 1.82 e − 07 | −5625.08 | 6.55 e − 39 |
| Pairings      | CRP    | 15.2    | 581        | 0.0043 ± 0.0027 | 0.99 | −2100.53 | 0.99 |
|               | Gaussian | 6       | 302        | 0.0043 ± 0.0028 | 0.99 | −2100.53 | 0.99 |
Figure 5. NIPS joint probability. Darker means higher probability. Pairings sumsampling for clarity.

Figure 8. WaCky Joint probability. Darker means higher probability. Pairings sumsampling for clarity.
Bayesian factorization of joint categorical distributions

| Held out type | Model | Dimens. | sec/sample | probab. dist. | $p_{\text{dist}}$ | test ll | $p_{\text{ll}}$ |
|---------------|-------|---------|------------|---------------|-----------------|---------|-------------|
| Observations  | CRP   | 3.4     | 439        | 0.0033 ± 0.0036 | 1.44 e − 07    | −47840.54 | 9.61 e − 13 |
|               | Gaussian | 8       | 392        | 0.0006 ± 0.0079 | −47840.54      | 9.61 e − 13 |
| Pairings      | CRP   | 3       | 261        | 0.0041 ± 0.0037 | 0.96            | −73474.67 | 5.98 e − 267 |
|               | Gaussian | 8       | 472        | 0.0042 ± 0.0037 | −73474.67      | 5.98 e − 267 |

One viable and simple way might be to put a prior on the number of dimensions, like done for the alternative for the Dirichlet Process in (Miller & Harrison, 2013). This approach might have the advantage of not overestimating the number of components. Another path would be to explore the applicability of the weight space view of Gaussian Processes (Seeger, 2004) to the problem of inferring nonparametric latent variables. Interestingly, one view of the inference of numeric latent variables for categories can be interpreted as providing a feature map from categories to higher dimensional vector space (Schoelkopf & Smola, 2001).

References

Aldous, David J. Exchangeability and Related Topics. In École d’Été de Probabilités de Saint-Flour XIII1983, pp. 1–198. Springer, 1985. ISBN 9783540152033. doi: 10.1007/BFb0099421. URL http://www.springerlink.com/index/c31v17440871210x.pdf.

Andrieu, Christophe, de Freitas, Nando, Doucet, Arnaud, and Jordan, Michael I. An introduction to MCMC for machine learning. Machine learning, pp. 5–43, 2003. URL http://link.springer.com/article/10.1023/A:1020281327116.

Baroni, Marco, Bernardini, Silvia, Ferraresi, Adriano, and Zanchetta, Eros. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora, 2009. ISSN 1574-020X.

Dugas, Charles, Bengio, Yoshua, Belisle, Francois, Nadeau, Claude, and Garcia, Rene. Incorporating second-order functional knowledge for better option pricing. In Advances in Neural Information Processing Systems, 2001. URL http://books.google.com/books?id=en&lr=&id=Mgs2FwtgNxwC&oi=fnd&pg=PA472&dq=Incorporating+Second-Order+Functional+Knowledge+for+Better+Option+Pricing&ots=EJ3J2kFq8u&sig=vazK0p1n86e7xtyjGr1Gx4h4i_8.

Globerson, Amir, Chechik, Gal, Pereira, Fernando, and Tishby, Naftali. Euclidean embedding of co-occurrence data. Journal of Machine Learning Research, 8:2047–2076, 2007. URL http://eprints.pascal-network.org/archive/00004085/.

Griffiths, T and Ghahramani, Z. Infinite latent feature models and the Indian buffet process. 2005. URL http://eprints.pascal-network.org/archive/00001359/.

Lloyd, James Robert. Orbanz, Peter, and Roy, Daniel M. Random function priors for exchangeable arrays with applications to graphs and relational data. In Advances in Neural Information Processing Systems, pp. 1–9, 2012.

Miller, Jeffrey W and Harrison, Matthew T. A simple example of Dirichlet process mixture inconsistency for the number of components. In Advances in Neural Information Processing Systems, 2013. URL http://arxiv.org/abs/1301.2708.

Miller, Kurt T, Griffiths, Thomas L, and Jordan, Michael I. Nonparametric latent feature models for link prediction. In Advances in Neural Information Processing Systems, pp. 1–9, 2009. URL http://books.nips.cc/papers/files/nips22/NIPS2009_0960.pdf.

Neal, Radford M. Slice sampling. Annals of statistics, 31(3):705–767, 2003. URL http://www.jstor.org/stable/10.2307/3484813.

Palla, Konstantina, Knowles, D, and Ghahramani, Z. An infinite latent attribute model for network data. arXiv preprint, 2012. URL http://arxiv.org/abs/1206.6416.

Rescorla, R A and Wagner, A R. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In Black, A H and Prokasy, W F (eds.), Classical Conditioning II Current Research and Theory, volume 21, chapter 3, pp. 64–99. Appleton-Century-Crofts, 1972. ISBN 0390718017. doi: 10.1101/gr.110528.110. URL http://homepage.mac.com/sanagnos/rescorlawagner1972.pdf.
Schoelkopf, Bernhard and Smola, Alexander J. *Learning with Kernels*, volume 180. 2001. ISBN 0262194759. doi: 10.1198/jasa.2003.s269. URL http://www.ncbi.nlm.nih.gov/pubmed/16113762.

Seeger, Matthias. Gaussian processes for machine learning. *International journal of neural systems*, 14(2): 69–106, 2004. ISSN 0129-0657. doi: 10.1142/S0129065704001899.

Wainwright, MJ and Jordan, MI. *Graphical models, exponential families, and variational inference*, volume 1. 2008. doi: 10.1561/2200000001. URL http://www.nowpublishers.com/product.aspx?product=MAL&doi=2200000001http://dl.acm.org/citation.cfm?id=1498841.

Welch, BL. The generalization of "student’s" problem when several different population variances are involved. *Biometrika*, 1947. URL http://www.jstor.org/stable/10.2307/2332510.

Welling, Max and Teh, YW. Bayesian learning via stochastic gradient langevin dynamics. In *International Conference on Machine Learning*, 2011. URL http://machinelearning.wustl.edu/mlpapers/paper_files/ICML2011Welling_398.pdf.