CONJUGATE PRIORS FOR GAUSSIAN EMISSION PLSA RECOMMENDER SYSTEMS

Stefan Adalbjörnsson, Johan Swärd, Magnus Örn Berg, Søren Vang Andersen, and Andreas Jakobsson

Centre for Mathematical Sciences, Lund University,
P.O. Box 118, SE-221 00 Lund, Sweden

ABSTRACT

Collaborative filtering for recommender systems seeks to learn and predict user preferences for a collection of items by identifying similarities between users on the basis of their past interest or interaction with the items in question. In this work, we present a conjugate prior regularized extension of Hofmann's Gaussian emission probabilistic latent semantic analysis model, able to overcome the over-fitting problem restricting the performance of the earlier formulation. Furthermore, in experiments using the EachMovie and MovieLens data sets, it is shown that the proposed regularized model achieves significantly improved prediction accuracy of user preferences as compared to the latent semantic analysis model without priors.

***Index Terms*—** Recommender systems, collaborative filtering, probabilistic matrix factorization

1. INTRODUCTION

Recommender systems are of great importance in many forms of web-based applications and services, and are a widely used in applications and services such as, e.g., Facebook, YouTube, Netflix, Spotify, and eBay. In such systems, a user is recommended suitable items based on earlier searches, purchases, and/or other forms of information about the user. In forming the recommendations in these systems, the recommender makes a measure of fit, often by simply using a 2-norm or a mean-squared error (MSE) distortion measure, between the available user information and items, in order to find the most suitable matches in the latter (see, e.g., [1–10]). These recommendations are of great importance to both the user and the company providing the service; according to [11], 35 percent of what consumers purchase on Amazon, and 75 percent of what they watch on Netflix, result from product recommendations. Recommenders can be categorized into two broad groups depending on which kind of data is used as an input to make the prediction, either content based (CB), where meta-data related to the items is used such that new items are recommended that are similar to the previously consumed items, or collaborative filtering (CF), where the prediction may be based on all users interactions with the items, by exploiting the similarities in users consumption behavior to find appropriate recommendations. Earlier works suggest that when historical data of many users' past behavior is available, the CF approach is often superior to the CB filtering approach [12, p. 111]. For this reason, we here focus on the CF approach, and, specifically, on the case where the historical data includes a numerical ranking of the consumption behavior. Originally, nearest neighborhood based methods were often used to create the recommendations in these settings. Recently, the use of matrix factorization have been shown to be more successful, offering improved accuracy and attractive scalability [12]. This factorization is performed on a matrix where unconsumed item-user pairs are represented with missing values, typically using a low-rank approximation or dimensionality reduction approach, with the aim of creating prediction for these missing values. Probabilistic approaches, such as probabilistic Latent Semantic Analysis (pLSA) [2] and Latent Dirichlet Allocation (LDA) [13], perform this factorization using probabilistic frameworks. Due to the numerous parameters required in the pLSA framework, over-fitting is a commonly re-occurring problem, especially for users with only a few observations available, or, similarly, for items with only a few observed consumptions. The LDA approach may in some setups be less prone to overfitting than the pLSA approach. However, the pLSA method has significant merit in its computational simplicity when compared to the LDA method.

In this paper, we propose an extension of the pLSA model to alleviate the overfitting problem, using an explicit formulation of conjugate priors for regularization of the model parameters. The proposed approach forms an extension of the pLSA approach developed in [2], allowing for Gaussian emission distributions with appropriately chosen priors. Furthermore, the method avoids the usage of the generally intractable posterior distribution and the requirement of using approximate inference or computationally demanding simulations methods such as Markov Chain Monte Carlo, as is often used in LDA based methods. Instead, the method proposed here uses prior distributions allowing for a simple maximum a-posteriori expression that can be optimized using the Expectation-Maximization (EM) algorithm and use cross-

This work is supported in part by the Lund Humanities Lab, the Crafoord Foundation, the Swedish Research Council, Carl Trygger's foundation, and the Royal Physiographic Society in Lund

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

Algorithm 1 EM algorithm for pLSA.

initialize $p(z|u)$ and $\sigma_{i,z}, \mu_{i,z}$
repeat
  E-step for each z:
  $Q(z|\psi;\hat{\theta}) = \frac{\hat{p}(v|i,z)\hat{P}(z|u)}{\sum_{z'}\hat{p}(v|i,z')\hat{P}(z'|u)}$
  M-step for each z:
  $P(z|u) = \frac{\sum_{<u',i,v>u'=u} Q(z|u,i,v;\hat{\theta})}{\sum_{z'}\sum_{<u',i,v>u'=u} Q(z'|u,i,v;\hat{\theta})}$
  $\mu_{i,z} = \frac{\sum_{<u,i',v>i'=i} vQ(z|u,i,v;\hat{\theta})}{\sum_{<u,i',v>i'=i} Q(z|u,i,v;\hat{\theta})}$
  $\sigma^2_{i,z} = \frac{\sum_{<u,i',v>i'=i} (v-\mu_{i,z})^2 Q(z|u,i,v;\hat{\theta})}{\sum_{<u,i',v>i'=i} Q(z|u,i,v;\hat{\theta})}$
until convergence

validation to select the necessary hyper-parameters. The method has the same complexity as pLSA, although requires a one-time pre-calculation to select the hyper-parameters. As shown in the numerical results section, the proposed algorithm shows improved robustness to overfitting and superior prediction accuracy.

2. DATA MODEL

Let $\mathcal{I}$ and $\mathcal{U}$ denote the available sets of items and users, respectively. The task of the recommender system is to predict how the user will rate an unconsumed item, and then use this information to recommend the user an item. In [2], a pLSA model was introduced for this purpose, wherein each user, u, has a probability of belonging to each of K different states, z, here denoted $P(z|u)$. The main idea behind introducing such states is to make a model assumption that each user-item pair is independent when conditioned on a given z, thereby capturing some fundamental behavior of the data, which may be captured by the different states z.

In this work, we focus on the so-called forced prediction case, i.e., when the distribution of each item is modeled conditional on that this item has been rated by the user, although it should be noted that the free prediction case, where one conditions only on the user and not on the item, can be handled in a similar manner. With the introduction of the states, one may form the probability that given an item, i, and a user, u, the probability of the rating, v, may be expressed as

$$P(v|u,i) = \sum_{z} P(v,z|u,i) = \sum_{z} P(v|i,z)P(z|u) \tag{1}$$

where the sum is formed over all available states, z. Typically, a user may rate each item according to a pre-defined set of ratings, here denoted $\mathcal{V}$, such a dislike or like rating, generally represented as $\{-1, 1\}$, or according to some cardinal scale, e.g., $\{1, 2, 3, 4, 5\}$. Depending on the available data, being either *implicit*, such that only the fact that the item has been consumed by the user is available, or *explicit*, for which the user has also provided a numerical rating of the consumed item, the distribution of the emission from each state $P(v|i,z)$ may be suitably chosen, e.g., for a continuous variable, a Gaussian distribution, or for a discrete variable, a multinomial distribution. However, as the data is commonly in need of being rescaled and/or mean adjusted [12, 14] to account for varying user rating behavior, as well as to account for similar item mean differences, the Gaussian distribution is in many cases more useful, and provides a significant improvement in prediction accuracy [2]. Thus, let

$$P(v|u,i) = P(v|u,i;\theta) = \sum_{z} P(z|u) f(v; \mu_{i,z}, \sigma_{i,z}) \tag{2}$$

where θ denotes a vector containing the unknown parameters, and $f(v;\mu,\sigma)$ the Gaussian distribution with mean μ and variance σ^2, i.e.,

$$f(v;\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(v-\mu)^2}{2\sigma^2} \right) \tag{3}$$

Given the parameters, the missing data may be estimated using

$$\begin{aligned} E[v|u,i] &= \int_{\mathcal{V}} v\, P(v|u,i) \mathrm{d}v = \sum_{z} P(z|u) \int_{\mathcal{V}} v\, f(v|z,i) \mathrm{d}v \\ &= \sum_{z} P(z|u) \mu_{i,z} \end{aligned} \tag{4}$$

Thus, each missing data point can be seen to be estimated via the inner-product of two K-dimensional vectors, indicating the connection to the low-rank matrix factorization approaches. However, it may be noticed that for sparse data sets, wherein only a small percentage of the ratings are observed, the number of estimated parameters might be larger than the number of ratings, making the model prone to overfitting [2, 13, 15]. This is a serious problem for the recommender, occurring for all users who have consumed fewer than K items, and will thus affect all new users to the system as well as the overall model parameters in general. To alleviate this problem, we here introduce the use of a regularizing prior to lessen the overfitting problem, reminiscent to the ideas explored in [16, 17] for the case of a Gaussian mixture model. By using a maximum a-priori approach and an appropriate choice of priors, the resulting optimization algorithm may be formed with the same order of complexity as the algorithm in [2], while avoiding any use of overfitting heuristics, such as early stopping, and allows the biasing effect of the prior to be automatically decreased for users and items which appear more often in the data set. In order to achieve this, the appropriate choice for the prior to the variable $p_{z|u} = P(z|u)$ is the Dirichlet distribution [16, 17]

$$p_{z|u}|\gamma \sim Dir(\gamma_1 + 1, ..., \gamma_K + 1) \tag{5}$$

where γ is a K-dimensional hyper-parameter, such that

$$f\left(p_{z|u}|\gamma\right) = \frac{1}{\mathrm{B}(\gamma)} \prod_{i=1}^{K} p_{z_i|u}^{\gamma_i - 1} \tag{6}$$

Algorithm 2 EM algorithm for conjugate-prior pLSA

1: initialize $p(z|u), \sigma_{i,z}$ and, $\mu_{i,z}$
2: **repeat**
3: E-step for each z:
4: $Q^*(z|\psi;\hat{\theta}) = \frac{\hat{p}(v|i,z)\hat{P}(z|u)}{\sum_{z'}\hat{p}(v|i,z')\hat{P}(z'|u)}$
5: M-step for each z:
6: $P(z|u) = \frac{\sum_{<u',i,v>u'=u} Q(z|u,i,v;\hat{\theta})+(\gamma_{u,z}-1)}{\sum_z \sum_{<u',i,v>u'=u} Q(z|u,i,v;\hat{\theta})+(\gamma_{u,z}-1)}$
7: $\mu_{i,z} = \frac{\sum_{<u,i',v>i'=i} vQ(z|u,i,v;\hat{\theta})}{\sum_{<u,i',v>i'=i} Q(z|u,i,v;\hat{\theta})}$
8: $\sigma^2_{i,z} = \frac{\sum_{<u,i',v>i'=i}(v-\mu_{i,z})^2 Q(z|u,i,v;\hat{\theta})+2\beta_{i,z}}{\sum_{<u,i',v>i'=i} Q(z|u,i,v;\hat{\theta})+2\alpha_{i,z}+1}$
9: **until** convergence

with $\mathrm{B}(\cdot)$ denoting the multinomial Beta function, which may be expressed using the Γ-function [18] as

$$\mathrm{B}(\gamma) = \frac{\prod_{i=1}^K \Gamma(\gamma_i)}{\Gamma\big(\sum_{i=1}^K \gamma_i\big)} \tag{7}$$

It may be noted that the Dirichlet distribution is a conjugate prior distribution for categorical and multinomial distributions, meaning that for such model, using such a prior implies that the posterior distribution will also be Dirichlet distributed. This choice of regularizer also allows for an intuitive interpretation of the hyper-parameter; the choice of γ_i will be equivalent with a corresponding assumption on the number of observations, say n_i, of the ith categorical variable

$$\gamma_i = n_i + 1 \tag{8}$$

Similarly, for the Gaussian emission distribution, we choose priors for μ_i and σ_i^2 using the normal-Gamma distribution

$$\mathrm{N}(\mu|\eta, \frac{\sigma}{\sqrt{\nu}})\Gamma^{-1}(\sigma^2|\alpha,\beta) = \frac{\sqrt{\nu}}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\nu(\mu-\eta)^2}{2\sigma^2}\right)\times \\ \times \frac{\beta^\alpha}{\Gamma(\alpha)}\left(\frac{1}{\sigma^2}\right)^{\alpha+1} \exp\left(-\frac{\beta}{\sigma^2}\right) \tag{9}$$

where, for notational convenience, we have omitted the subindex i from all variables, and

$$\alpha = \frac{n_i}{2} \tag{10}$$

$$\beta = \frac{n_i}{2}\frac{\sum_{y\in x_i}(y-\bar{x}_i)^2}{n_i+1} \tag{11}$$

with x_i denoting the observed sets of means, μ_i, or variances, σ_i^2, respectively, and with $\bar{x}_i$ being the mean of x_i. This allows for a similar pseudo-observation interpretation of the hyper-parameters

$$\eta = \frac{n_i}{n_i+1}\bar{x}_i \tag{12}$$

$$\nu = n_i + 1 \tag{13}$$

implying that the prior knowledge on μ_i and σ_i^2 may be summarized as

$$\mu_i|\sigma_i, x_i \sim N\left(\frac{n_i}{n_i+1}\bar{x}_i, \frac{\sigma_i^2}{n_i+1}\right) \tag{14}$$

$$\sigma_i^2|x_i \sim \Gamma^{-1}\left(\frac{n_i}{2}, \frac{n_i}{2}\frac{\sum_{y\in x_i}(y-\bar{x}_i)^2}{n_i+1}\right) \tag{15}$$

In [2], Expectation-Maximization (EM) approach was suggested to estimate the parameters of the pLSA model, maximizing the likelihood

$$L\big(\psi,\theta\big) = -\sum_\psi \log f(v|u,i;\theta) \tag{16}$$

where the summation is over all the observed data triplets, $\psi = (u,i,v)$. By the EM method, a variational distribution, $Q(\cdot,\psi)$, is introduced for each observed data triplet, such that

$$\sum_z Q(z|\psi) = 1 \tag{17}$$

allowing the log-likelihood to be majorized using the Jensen's inequality as follows:

$$\begin{aligned} L\big(\psi,\theta\big) &= -\sum_\psi \log\sum_z Q(z|\psi)\frac{P(z|u)f(v;\mu_{i,z},\sigma_{i,z})}{Q(z|\psi)} \\ &\le -\sum_\psi\sum_z Q(z|\psi)\log\frac{P(z|u)f(v;\mu_{i,z},\sigma_{i,z})}{Q(z|\psi)} \\ &= -\sum_\psi\sum_z Q(z|\psi)\log P(z|u)f(v;\mu_{i,z},\sigma_{i,z}) + \\ &\quad +\sum_\psi\sum_z Q(z|\psi)\log Q(z|\psi) \\ &= \tilde{L}(\theta,Q) - \sum_\psi H\big(Q(z|\psi)\big) \end{aligned} \tag{18}$$

with $\tilde{L}(\theta,Q)$ denoting the likelihood of θ expressed in Q, and $H(\cdot)$ is the entropy function. Thus, the log-likelihood in (18) has the same form that lends itself to maximization by the EM algorithm to find a local maximum by iteratively calculating Q and updating the parameters θ. The resulting steps are summarized in Algorithm 1.

We now derive a conjugate-prior version of the above method. We do this by instead seeking to maximize the posterior distribution, implying that

$$\begin{aligned} L_{reg}\big(\psi,\theta\big) &= -\sum_\psi \log P(\theta|\psi) \\ &\le L(\theta,Q) - \sum_\psi H\big(Q(z|\psi)\big) \\ &= -\sum_\psi \log \mathrm{Dir}\big(P(z|u)\big|\gamma\big) - \\ &\quad -\sum_\psi\sum_z\left[\log N(\mu;\mu_0,\frac{\sigma}{\nu}) + \log\Gamma^{-1}(\sigma^2;\alpha,\beta)\right] \end{aligned} \tag{19}$$

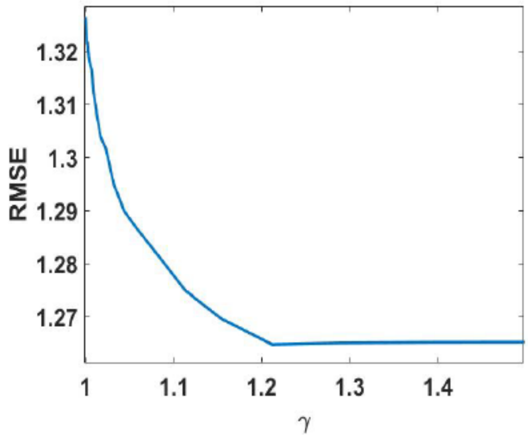

Fig. 1. The RMSE as a function of the used γ parameter.

The resulting steps are summarized in Algorithm 2, from which it may be noted that the prior enters into the optimization steps according to what might be expected from the pseudo-observation comparison above. This also implies that the computational complexity of the algorithms are of the same order of magnitude.

3. NUMERICAL RESULTS

In this section, we examine the performance of the proposed method, comparing it to the pLSA [2] and the Pop-estimator, which simply calculates the average rating for each movie.

We initially examine the so-called EachMovie data set [19], containing 2811983 ratings made by 72916 users on 1628 movies. The data set was divided into three sets: training, validation, and test sets. Users having rated less than 3 movies, and movies that had been rated fewer than 1 times where removed from both the validation and test sets. Furthermore, the data sets were centered and rescaled, such that each users rating has zero mean and unit variance, i.e., the centered and rescaled ratings were formed as

$$v' = \frac{v - \mu_u}{\sigma_u} \tag{20}$$

where

$$\sigma_u^2 = \frac{\sum_{<u',v,i>:u'=u} (v - \mu_u)^2 + q\tilde{\sigma}^2}{N_u + q} \tag{21}$$

$$\mu_u = \frac{\sum_{<u',v,i>:u'=u} v + q\tilde{\mu}}{N_u + q} \tag{22}$$

Here q is chosen as in [2], were it was found to yield a performance increase, since it removes some of the user biases connected to each users unique interpretation of the rating scale. After estimating the parameters using the centered and

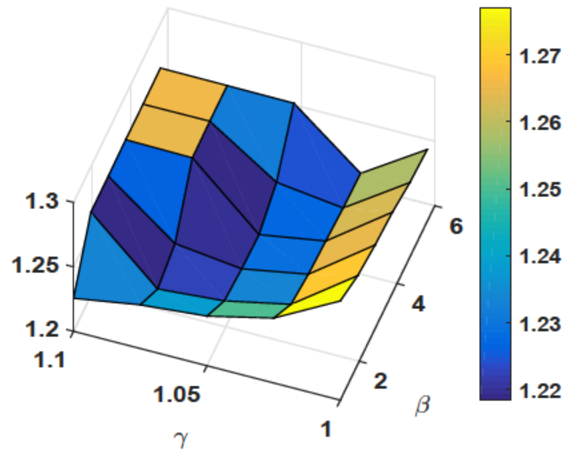

Fig. 2. The final grid search over the three prior parameters, showing the γ-β plane, with α being fixed at 0.5. The minimum being at $\hat{\alpha} = 0.5$, $\hat{\beta} = 1.92$, and $\hat{\gamma} = 1.075$.

rescaled ratings, the predictions were formed as

$$E[v'|u,i] = \mu_u + \sigma_u \sum_z P(z|u)\mu_{i,z} \tag{23}$$

Throughout this evaluation, the EM-iterations of the proposed method were terminated when the log-likelihood did not improve more than 10^{-3} between two consecutive iterations, whereas the pLSA algorithm was terminated using early stopping, as advised in [2], meaning that the algorithm was terminated when the RMSE of the validation set was increasing, whereafter a final iteration was done on the combined data of the training and the validation set. The estimated parameters where then applied to the test data set, yielding the final root mean square error (RMSE), formed as

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(v_i - \hat{v}_i\right)^2} \tag{24}$$

where $\hat{v}_i$ and v_i denotes the predicted and the true rating, respectively. In order to find appropriate values of the prior parameters α, β, and γ, the recommender was trained on the training data set, and then evaluated over the priors on the validation set. This was done by initially finding each of the γ parameters minimizing the RMSE over the validation set, as illustrated in Figure 1. Using each initial γ parameter, the RMSE was minimized over α, β, and γ, by initially evaluating the RMSE over a coarse grid structure, and then zoom in on the most promising parts on the grid. Figure 2 shows the result for the final grid search; the final estimate was found to be $\hat{\alpha} = 0.5$, $\hat{\beta} = 1.92$, and $\hat{\gamma} = 1.075$ for the EachMovie database. Given these priors, the performance of the proposed method was evaluated on the test data set. Table 3 summarize

	Proposed	Hofmann	Pop
RMSE	1.2191	1.2690	1.3715
MAE	0.9394	0.9898	1.0914

Table 1. Results from the EachMovie data set.

	Proposed	Hofmann	Pop
RMSE	0.9052	0.9471	0.9877
MAE	0.7094	0.7395	0.7892

Table 2. Results from the MovieLens data set.

the RMSE and mean absolute error (MAE), defined as

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|v_i - \hat{v}_i\right| \tag{25}$$

as compared to the corresponding results for the pLSA and Pop methods, clearly indicating the achievable improvement when including the priors.

Finally, we examine the so-called MovieLens data set [20]. This data set contains 1,000,209 ratings of about 3900 movies, made by 6040 users. Using a similar cross-validation as detailed for the EachMovie data set, we find the priors to be $\hat{\alpha} = 0.5$, $\hat{\beta} = 3$, and $\hat{\gamma} = 1.25$ for the MovieLens data set. Table 3 summarize the RMSE and MAE for the resulting predictions, again showing the benefit of introducing priors in the recommender.

REFERENCES

[1] Y. Koren and R. Bell, *Recommender Systems Handbook*, chapter Chapter 5 Advances in Collaborative Filtering, pp. 145–186, Springer, 2011.

[2] T. Hofmann, "Latent Semantic Models for Collaborative Filtering," *ACM Trans. Inf. Sys.*, vol. 22, no. 1, pp. 89–115, Jan. 2004.

[3] A. Theodoridis, C. Kotropoulos, and Y. Panagakis, "Music Recommendation using Hypergraphs and Group Sparsity," in *IEEE Int. Conf. Acoust., Speech, and Signal Processing*, Vancouver, Canada, May 26-31. 2013, pp. 56–60.

[4] K. Alodhaibi, A. Brodsky, and G. A. Mihaila, "A Confidence-Based Recommender with Adaptive Diversity," in *IEEE Symp. on Computational Intelligence in Multicriteria Decision-Making*, Paris, France, April 11–15 2011.

[5] G. Dror, N. Koenigstein, and Y. Koren, "Web-Scale Media Recommendation Systems," *Proceedings of the IEEE*, vol. 100, no. 9, pp. 2722–2736, Sept. 2012.

[6] Y. Koren, "Factorization Meets the Neighborhood: a Multifaceted Collaborative Filtering Model," in *KDD*, Las Vegas, Nevada, USA, Aug. 24–27 2008.

[7] X. Li and T. Murata, "Using Multidimensional Clustering based Collaborative Filtering Approach Improving Recommendation Diversity," in *IEEE/WIC/ACM Int. Conf. on Web Intelligence and Intelligent Agent Technology*, Macau, Dec. 4–7 2012.

[8] M. Nazim, J. Shrestha, and G. S. Jo, "Enhanced Content-based Filtering using Diverse Collaborative Prediction for Movie Recommendation," in *First Asian Conf. on Intelligent Information and Database Systems*, Dong Hoi, Quang Binh, Vietnam, April 1–3 2009.

[9] Y. Song, S. Dixon, and M. Pearce, "A Survey of Music Recommendation Systems and Future Perspectives," in *Proc. 9'th Int. Symp. Computer Music Modelling and Retrieval*, London, United Kingdom, June 19–22. 2012.

[10] Y. X. Zhu, W. Zeng, and Q. M. Zhang, "The Effect of Rating Variance on Personalized Recommendation," in *5th Int. Conf. on Computer Science & Education*, Hefei, China, Aug. 24–27 2010.

[11] I. MacKenzie, C. Meyer, and S. Noble, "How retailers can keep up with consumers," "{`http://www.mckinsey.com/insights/consumer\_and\_retail/how\_retailers\_can\_keep\_up\_with\_consumers`}", October 2013.

[12] F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, Eds., *Recommender Systems Handbook*, Springer, 2011.

[13] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[14] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical Learning with Sparsity: The Lasso and Generalizations*, Chapman and Hall/CRC, 2015.

[15] N. Barbierir, G. Manco, and E. Ritacco, *Probabilistic Approaches to Recommendations*, Morgan & Claypool, 2014.

[16] J. Chu and Y. Lee, "Conjugate Prior Penalized Learning of Gaussian Mixture Models for EMG Pattern Recognition," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, San Diego, CA, 2007, pp. 1093–1098.

[17] J. Chu and Y. Lee, "Conjugate-Prior-Penalized Learning of Gaussian Mixture Models for Multifunction Myoelectric Hand Control," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 17, no. 3, pp. 287–297, June 2009.

[18] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis, Second Edition (Chapman & Hall/CRC Texts in Statistical Science)*, Chapman and Hall/CRC, 2nd edition, 2003.

[19] Digital Equipment Corporation, "EachMovie recommendation dataset," `http://www.cs.cmu.edu/~lebanon/IR-lab/data.html`, 2004.

[20] GroupLens Research Group, "MovieLens dataset," "`http://grouplens.org/datasets/movielens/1m/`", 2003.