# Approximating the DCM

**Rasmus E. Madsen**
Department of Mathematical Modelling
Technical University of Denmark
Lyngby, DK-2800
rem@imm.dtu.dk

## Abstract

The Dirichlet compound multinomial (DCM), which has recently been shown to be well suited for modeling for word burstiness in documents, is here investigated. A number of conceptual explanations that account for these recent results, are provided. An exponential family approximation of the DCM that is substantially faster to train, while still producing similar probabilities and classification performance is provided.

## 1 Introduction

The Dirichlet compound multinomial has previously been demonstrated capable at modeling word burstiness. The DCM is both qualitatively and quantitatively, better than the multinomial model on standard document collections (Madsen et al., 2005).

The DCM model is however not without problems. Though the Dirichlet distribution and the multinomial distribution both are members of the exponential family, the compound model of the two, the DCM, is not a member of the exponential family. Exponential families have many desirable properties (Banerjee et al., 2005), and it is therefore desirable to use functions within the exponential family. Second, because of the relative complexity of the DCM expression, understanding it's behavior qualitatively is difficult. Third, DCM parameters cannot be estimated quickly, i.e there is no closed form solution. When estimating DCM parameters it is necessary to apply gradient descent methods (Minka, 2003), which are costly and slow. Fast training is important not only for modeling large document collections but also for using DCM distributions in more complex mixtures or hierarchical models, such as LDA (Blei et al., 2003).

We here present an approximation of the DCM that is in the exponential family. An exact solution of the maximum likelihood parameters for the approximate distribution is derived. The approximate distribution can be computed efficiently (more than 100 times faster than the DCM), and has a categorization accuracy similar to that of the DCM.

## 2 Approximation

In this section we derive an exponential family approximation of the DCM distribution that we call the EDCM and investigate the qualitative behavior of the EDCM.

We start by taking a look at the DCM from equation **??** where we define the new variable

$s$ as the sum of the parameters $s = \sum_w \alpha_w$, still keeping in mind the document length $n = \sum_w x_w$.

$$p(x|\alpha) = \frac{n!}{\prod_w^W x_w!} \frac{\Gamma(s)}{\Gamma(s+n)} \prod_{w=1}^{W} \frac{\Gamma(x_w + \alpha_w)}{\Gamma(\alpha_w)}. \quad (1)$$

Where the parameter $s$ is controlling the degree of burstiness in the model. When training the DCM model, we find empirically that $\alpha_w << 1$ for practically all words in the vocabulary. For one class of newsgroup articles, the average $\alpha_w$ is $0.004$ and out of the 59,826 parameters, $99\%$ are below $0.1$, 17 are above 0.5, and only 5 are above 1.0. The $\alpha$ parameters can therefore be regarded as being small.

A useful approximation that can be applied when the $\alpha$'s are small is: $\Gamma(x + \alpha) \approx \Gamma(x)\alpha$. We further use $\Gamma(x) = (x-1)!$ when $x$ is an integer. From these approximations we get the EDCM distribution[1] $q(x|\beta)$.

$$q(x|\beta) = n! \frac{\Gamma(s)}{\Gamma(s+n)} \prod_{w:x_w \geq 1} \frac{\beta_w}{x_w} \quad (2)$$

For clarity, we have denoted the EDCM parameters $\beta_w$. The parameter $s$ is therefore now $s = \sum_w \beta_w$. The $q(x|\beta)$ is not a proper probability distribution, that is it does not sum to one, since we have used the approximation. It is however in principle possible to normalize $q(x|\beta)$ to sum to one, by summing $q(x|\beta)$ over all values of x to get a normalizing constant $Z(\beta)$. This technicality is however not considered here[2]. In practice the values given by Equation 2 are very close to those given by Equation 1. On a sample set of 4000 documents from 20 different classes $q(x|\alpha)$ is highly correlated with $p(x|\alpha)$. On average, $q(x|\alpha)$ only deviates by $2.2\%$ from $p(x|\beta)$. This high correlation is because the $\Gamma(x+\alpha)/\Gamma(\alpha)$ approximation is highly accurate for small $\alpha_w$ values. For a typical $\alpha$- vector trained from 800 documents, of the 69,536 non-zero word counts, the approximation is on average $3.9\%$ off. In Figure 1 the DCM $\alpha$-parameters are compared with the EDCM $\beta$-parameters. The parameter values are close to be forming a straight line, showing high degree of similarity.

From Equation 2 we get some insight about the DCM as well as the EDCM, that was not directly obvious from Equation 1. For fixed $s$ and $n$, the probability of a document is proportional to $\prod_{w:x_w \geq 1} \beta_w / x_w$. This means that the first appearance of a word w reduces the probability of a document by $\beta_w$, a word-specific factor that is almost always much less than 1.0, while the $m$'th appearance of any word reduces the probability by $(m-1)/m$, which tends to 1 as $m$ increases. This behavior reveals how the EDCM, and hence the DCM, allow multiple appearances of the same word to have high probability. In contrast, with a multinomial each appearance of a word reduces the probability by the same factor.

One consideration about the DCM was that it does not belong to the exponential family. We now rewrite Equation 2 to the exponential family form. An exponential family distribution has the form $f(x)g(L)\exp[t(x)h(\beta)]$ where $t(x)$ is a vector of "sufficient statistics" and $\theta = h(\beta)$ is the vector of so-called "natural parameters". We can write $q(x|\beta)$ in this form as:

[1]The EDCM is not a true distribution while the integral over the EDCM is not exactly one.

[2]Since $q(x|\beta)$ is not a proper probability distribution, we cannot calculate perplexity or other probabilistic measures that tell how well the EDCM models the data, but the focus is here categorization.

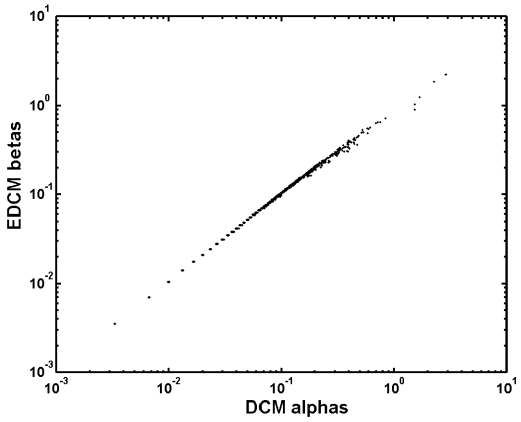



Figure 1: Comparison of the $\alpha$-parameters of the DCM model and the $\beta$-parameters of the EDCM model. The parameter values follow a straight line, showing that the two methods of estimation result in almost the same parameter values. Even for the large parameter values, the approximation is quite accurate, though the approximation equations were conditioned on having small $\alpha$ values.

$$q(x|\beta) = \left( \prod_{w:x_w \geq 1} x_w \right) n! \frac{\Gamma(s)}{\Gamma(s+n)} \exp\left[ \sum_{w:x_w \geq 1} \beta_w \right] \tag{3}$$

For the EDCM distribution, the sufficient statistics for a document $x$ are the normalized data $\langle t_1(x), ..., t_W(x) \rangle$ where $t_w(x) = I(x_w \geq 1)$ and $W$ is the number of words in the vocabulary. The expression also shows that the natural parameters for the EDCM distribution are $\theta_w = \ln \beta_w$.

## 3 Maximum Likelihood Estimation

The maximum likelihood estimate of the EDCM parameters can be determined by taking the derivative of the log-likelihood function. This is in contrast to the complications involved in determining the parameters of the DCM (Minka, 2003). From Equation 2 the log-likelihood function can be determined.

$$\mathcal{L}_\beta(x) = \log(n) + \log\left(\Gamma(s+n)\right) + \sum_{w:x_w \geq 1} \log(\beta_w) - \log(x_w) \tag{4}$$

Given a set of training documents, we can calculate the partial derivative of the log-likelihood.

$$\frac{d\mathcal{L}_\beta(x)}{d\beta_w} = |D|\Psi(s) - \sum_{d=1}^{D} \Psi(s+n_d) + \frac{I(x_{dw} \geq 1)}{\beta_w} \tag{5}$$

Setting the derivative of the log-likelihood equal to zero and solving for the parameters $\beta_w$ we get Equation 6.

$$\beta_w = \frac{\sum_{d=1}^{D} I(x_{dw} \geq 1}{|D|\Psi(s) - \sum_{d=1}^{D} \Psi(s + n_d)} \tag{6}$$

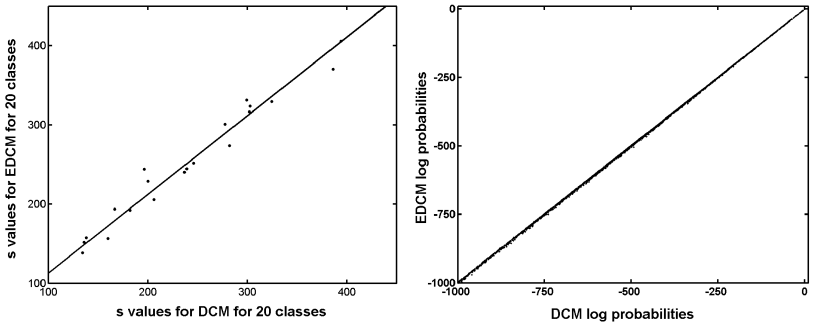
Since $\beta_w$ is part of $s$, Equation 6 is not directly solvable as is. The parameter sum $s$ can be computed though by summing over the words $w$ in Equation 6.

$$s = \sum_{w=1}^{W} \beta_w = \frac{\sum_{w=1}^{W} \sum_{d=1}^{D} I(x_{dw} \geq 1}{|D|\Psi(s) - \sum_{d=1}^{D} \Psi(s + n_d)} \tag{7}$$

Equation 7 can be solved numerically for $s$ efficiently, since it only involves one unknown parameter. Having solved for $s$, Equation 6 is easily solvable.

## 4 Experiments

In Figure 2 we start by comparing the parameter sums $s$ of the DCM and EDCM followed by a comparison of the likelihood for the two models. The $s$-value tells to what extent that burstiness is present in the data. The values are generally close to being the same, revealing that the two models agree on the extent of burstiness. The probability estimates for the two models are also similar, which shows that the approximations used for the EDCM model are accurate.




(a) Sum of DCM $\alpha_w$ versus sum of EDCM $\beta_w$

(b) $\mathcal{L}_\alpha(x)$ (DCM) versus $\mathcal{L}_\beta(x)$ (EDCM)

Figure 2: Comparing the parameter sums (a) for the DCM and EDCM (degree of burstiness) reveals that the two models agree on the burstiness. In (b) the log-likelihood for the two models is compared for 4000 test documents. The log-likelihood estimated by the approximation is close to the real thing.

For classification purposes we have compared the DCM and EDCM with the multinomial model. The multinomial model is smoothed in an optimal way, and the accuracy of the multinomial is therefore higher than in (Rennie et al., 2003).

Table 1: Classification accuracy for the 20 newsgroups and Industry Sector collections, comparing the multinomial, DCM and EDCM. The scores are averages of 10 random splits.

| DATA SET | MULTINOMIAL | DCM | EDCM |
|---|---|---|---|
| 20 NEWSGROUPS | 0.855 | 0.862 | 0.864 |
| INDUSTRY SECTOR | 0.789 | 0.804 | 0.798 |

As we had hoped, the DCM and EDCM model have very similar classification performances. In fact, for 20 newsgroups, the EDCM model actually performs better than the DCM model. This is particularly encouraging considering the EDCM model was almost 150 times faster to train than the DCM model (19 seconds vs. 2788 seconds for a 20 newsgroups split using the fastest DCM fixed point training method). Both the DCM and EDCM use a naive smoothing method and still performs slightly better than the multinomial.

## 5 Conclusion

The approximated DCM model, the EDCM, has added an additional insight by giving an intuitive explanation for how the DCM models burstiness. The proposed approximation to the DCM distribution further belongs to the exponential family of distributions and is orders of magnitude faster to train. The estimated EDCM parameters and EDCM approximated probabilities are close to the true DCM values, resulting in similar classification performance for the two models.

## References


Banerjee, A., Merugu, S., Dhillon, I., & J., G. (2005). Clustering with bregman divergences. *to appear in Journal of Machine Learning Research, 6*.

Blei, D., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Madsen, R., Kauchak, D., & Elkan, C. (2005). Modeling word burstiness using the dirichlet distribution. *to appear in the Proceedings of ICML-05*.

Minka, T. (2003). *Estimating a Dirichlet distribution*. www.stat.cmu.edu/˜minka/papers/dirichlet.

Rennie, J. D. M., Shih, L., Teevan, J., & Karger, D. R. (2003). Tackling the poor assumptions of naive Bayes text classifiers. *Proceedings of the Twentieth International Conference on Machine Learning* (pp. 616–623). Washington, D.C., US: Morgan Kaufmann Publishers, San Francisco, US.