# Learning Structure with the Trace Norm Distribution*

Jason D. M. Rennie
jrennie@gmail.com

February 26, 2006

**Abstract**

We consider the problem of learning model structure. We assume that data is generated by one or more trace norm distributions [3]. We find that, as with other unsupervised problems, there are a number of equally-corrrect solutions. Which solution is best depends on the interpretation of distances in the data space.

## 1 Trace Norm Distribution

Let $X \in \mathbb{R}^{n \times m}$. The trace norm distribution is defined as

$$P_\lambda^{n \times m}(X) = \frac{1}{Z_\lambda^{n \times m}} \exp(-\lambda \|X\|_\Sigma), \tag{1}$$

where the superscript is not exponentiation, but rather a designation that the distribution is specific to the matrix size. $\|X\|_\Sigma$ is the trace norm of $X$ (the sum of singular values of $X$), and

$$Z_\lambda^{n \times m} = \int \exp(-\lambda \|X\|_\Sigma) dX \tag{2}$$

is the normalization constant, where the integral is over all matrices of size $n \times m$. See [1] and [2] for discussion of computation of the normalization constant.

## 2 An Example

Consider real data generated in the form of a $2 \times 2$ matrix, $X \in \mathbb{R}^{2 \times 2}$. Let $X_i$ designate the $i^{\text{th}}$ row of $X$. We consider two models: (1) the entire matrix is generated via a trace norm distribution, $X \sim P^{2 \times 2}$, and (2) each row is generated independently via a trace norm distribution, $X \sim P^{1 \times 2}(X_1) P^{1 \times 2}(X_2)$.

*Joint work with John Barnett and Tommi Jaakkola.

Given data $X \in \mathbb{R}^{2\times 2}$, we find the model most likely to have generated the data. This is simply a matter of determining which model yields higher data likelihood.

First, we consider case #1. Here, the likelihood of the data is

$$P^{2\times 2}(X) = \frac{\exp(-\lambda\|X\|_\Sigma)}{Z^{2\times 2}} = \frac{\exp(-\lambda\|X\|_\Sigma)}{\frac{1}{\lambda^4}\frac{1}{4}\left(\frac{3}{2}\right)^2 \mathrm{Vol}(V_{2,2})^2}, \tag{3}$$

where $\mathrm{Vol}(V_{2,2}) = \frac{2\pi}{\Gamma(1)}\frac{2\pi^{1/2}}{\Gamma(1/2)} = 4\pi$ is the volume of the $2 \times 2$ Stiefel manifold [4]. Next, we consider case #2. Here, the likelihood is

$$P^{1\times 2}(X_1)P^{1\times 2}(X_2) = \frac{\exp(-\lambda\|X_1\|_\Sigma - \lambda\|X_2\|_\Sigma)}{\left(\frac{1}{\lambda^2}\frac{1}{2}\mathrm{Vol}(V_{2,1})\mathrm{Vol}(V_{1,1})\right)^2}, \tag{4}$$

where $\mathrm{Vol}(V_{2,1}) = \frac{2\pi}{\Gamma(1)} = 2\pi$ and $\mathrm{Vol}(V_{1,1}) = \frac{2\pi^{1/2}}{\Gamma(1/2)} = 2$. Note that $\|X\|_\Sigma \leq \|X_1\|_\Sigma + \|X_2\|_\Sigma$ and $Z^{2\times 2} = 9\pi^2 \geq 4\pi^2 = \left(Z^{1\times 2}\right)^2$. Model #1 will be preferred if

$$\frac{P^{2\times 2}(X)}{P^{1\times 2}(X_1)P^{1\times 2}(X_2)} = \frac{\exp(\lambda(\|X_1\|_\Sigma + \|X_2\|_\Sigma - \|X\|_\Sigma))}{9/4} > 1. \tag{5}$$

Equivalently, model #1 will be preferred if the difference in singular values is larger than a linear function of $\frac{1}{\lambda}$,

$$\|X_1\|_\Sigma + \|X_2\|_\Sigma - \|X\|_\Sigma > \frac{\log 9/4}{\lambda}. \tag{6}$$

# 3 Discussion

The critical value of $\lambda$—the value at which neither model is preferred—is highly sensitive to a number of factors, including the scale of the data. But, absent information pertaining to the interpretation of distances in the data space, the critical value is unimportant. Given a set of models, what is important is the set of models that are preferred for some value of $\lambda$. In our framework, each partitioning of the data into two sets corresponds to a different model. Most partitionings will *never* be preferred, no matter what the value of $\lambda$. Such partitionings provide a poor fit to the data. This issue is much like that of selecting a width for kernel density estimation. The key question is: how far apart can two data items be yet still be considered similar? A large value of $\lambda$ is like a large kernel width—all data items look similar and look like they all belong to the same cluster/partition. A small value of $\lambda$ is like a small kernel width—all data items look different and look like they belong in their own cluster/partition. Sliding $\lambda$ between 0 and $\infty$, we find a set of models which are each optimal for a particular interpretation of distance in the data space.

# References

[1] J. D. M. Rennie. Approximating the trace norm distribution partition function. http://people.csail.mit.edu/jrennie/writing, January 2006.

[2] J. D. M. Rennie. Computing the trace norm distribution via sampling. http://people.csail.mit.edu/jrennie/writing, January 2006.

[3] J. D. M. Rennie. Toward normalization of the trace norm distribution. http://people.csail.mit.edu/jrennie/writing, January 2006.

[4] J. D. M. Rennie. Volume of the Stiefel manifold. http://people.csail.mit.edu/jrennie/writing, January 2006.