Toward Normalization of the Trace Norm Distribution

Jason D. M. Rennie
jrennie@gmail.com*

January 21, 2006

Abstract

Continuing our discussion from [3], we discuss how to obtain the normalization constant for our trace norm distribution.

Consider the task of modeling a set of documents using the framework established in [3]. We treat each document as having its own multinomial distribution. Let θ_i be the multinomial natural parameter vector for document i. The data likelihood is the usual multinomial likelihood, where we assume that the likelihood of a set of documents is simply the product of their individual likelihoods. We use the trace norm distribution to impose a prior on the parameters. Construct a matrix $\Theta = [\theta_1 \theta_2 \ldots \theta_n]^T$, where the multinomial natural parameter vectors serve as the rows. The trace norm distribution on the set of parameter vectors is

$$-\log P(\Theta|\lambda) = \lambda \|\Theta\|_{\mathrm{tr}} + \log Z_\lambda, \tag{1}$$

where $\|\cdot\|_{\mathrm{tr}}$ designates the *trace norm*, or the sum of singular values in a matrix. The single parameter, λ, allows for scaling of the distribution. Let $\sigma_1, \ldots, \sigma_m$ be the singular values of Θ. Then,

$$\|\Theta\|_{\mathrm{tr}} = \sum_i \sigma_i \tag{2}$$

The normalization constant for this distribution is an integral over matrices Θ. Since we are using natural parameters, values can vary anywhere along the real number line, so the integral is over $\mathbb{R}^{n\times m}$. n is the number of rows/documents; m is the number of columns/features. We assume $n > m$ for this discussion, but this assumption is not essential for our analysis.

In comparing different hierarchical models, it is essential that we be able to normalize the distribution. The normalization constant is

$$Z_\lambda = \int_{\mathbb{R}^{n\times m}} \exp\left(-\|\Theta\|_{\mathrm{tr}}\right) d\Theta. \tag{3}$$

*Updated January 23, 2006. Many thanks to John Barnett for helping me understand Edelman's notes and for suggesting helpful directions in this derivation.

But, Θ is an awkward representation for integrating over singular values. In §2.7 of [1], Edelman provides the Jacobian of the singular value decomposition, $\Theta = U\Sigma V^T$,

$$\prod_{i<j\leq m}(\sigma_i^2-\sigma_j^2)\prod_{i=1}^{m}\sigma_i^{n-m}(d\Sigma)^\wedge(H^TdU)^\wedge(V^TdV)^\wedge, \tag{4}$$

where H is an orthogonal $n\times n$ matrix with first m columns identical to U. Assumed is that the singular values are ordered and unique, $\sigma_1 > \sigma_2 > \cdots > \sigma_n$. Even when the singular values are unique, the SVD is not. The sign of columns of U and V may be switched without modifying $U\Sigma V^T$. So, we must divide the integral involving the SVD Jacobian by 2^m. After changing variables, $\Theta = U\Sigma V^T$, our integral becomes,

$$Z_\lambda = \frac{1}{2^m}\int \exp\left(-\lambda\sum_{i=1}^{m}\sigma_i\right)\prod_{i<j\leq m}(\sigma_i^2-\sigma_j^2)\prod_{i=1}^{m}\sigma_i^{n-m}(d\Sigma)^\wedge(H^TdU)^\wedge(V^TdV)^\wedge, \tag{5}$$

where $H\in\mathbb{R}^{n\times n}$ is orthogonal with first m columns identical to U. Note that this calculation can be made as the product of three separate integrals. Note that $\int(H^TdU)^\wedge$ is integration over the Stiefel manifold; $\int(V^TdV)^\wedge$ integrates over a special case of the Stiefel manifold (square matrices), otherwise known as the orthogonal group. In §5 of [2], Edelman provides the volume of the Stiefel manifold, $Q\in\mathbb{R}^{n\times m}$ s.t. $Q^TQ = I_m$, denoted $V_{m,n}$,

$$\mathrm{Vol}(V_{m,n}) = \frac{2^m\pi^{mn/2}}{\Gamma_m(n/2)}, \tag{6}$$

where $\Gamma_m(a) = \pi^{m(m-1)/4}\prod_{i=1}^{m}\Gamma[a-(i-1)/2]$. Note that $\mathrm{Vol}(V_{m,n}) = \prod_{i=m+1}^{n}A_i$, where A_i is the surface of the i-sphere of radius 1. Remaining is the integral over singular values,

$$\int \exp\left(-\sum_{i=1}^{m}\sigma_i\right)\prod_{i<j\leq m}(\sigma_i^2-\sigma_j^2)\prod_{i=1}^{m}\sigma_i^{n-m}(d\Sigma)^\wedge \tag{7}$$

Note that the singular values were assumed to be ordered, so the limits of integration are not independent of the variables. More explicity, the singular value integral is written as

$$\int_0^\infty\int_0^{\sigma_1}\cdots\int_0^{\sigma_{m-1}}\prod_{i<j\leq m}(\sigma_i^2-\sigma_j^2)\prod_{i=1}^{m}e^{-\sigma_i}\sigma_i^{n-m}d\sigma_m\cdots d\sigma_1 \tag{8}$$

Note that integrals of the form $\int_0^a e^{-x}x^b dx$ are (lower) incomplete gamma functions. Hence, we can write this integral as a sum of incomplete gamma functions. Note that the innermost integral can be written as

$$\int_0^{\sigma_{m-1}} e^{-\sigma_i}\sigma_m^{n-m}\prod_{i=1}^{m-1}(\sigma_i^2-\sigma_m^2)d\sigma_m \tag{9}$$

Expanding the product, we arrive at a polynomial function of σ_m,

$$\prod_{i=1}^{m-1}(\sigma_i^2-\sigma_m^2)=\prod_{i=1}^{m-1}\sigma_i^2-\sigma_m^2\sum_{i=1}^{m-1}\prod_{j\neq i}\sigma_j^2+\cdots-\sigma_m^{2(m-2)}\sum_{i=1}^{m-1}\sigma_i^2+\sigma_m^{2(m-1)}, \tag{10}$$

where the signs of the later terms assume that m is odd. Integrating, we get a sum of lower incomplete gamma functions. Note that the lower incomplete gamma function [4] is defined as

$$\gamma(a,x)=\int_0^x t^{a-1}e^{-t}dt. \tag{11}$$

So, we have

$$\begin{aligned}\int_0^{\sigma_{m-1}} e^{-\sigma_i}\sigma_m^{n-m}\prod_{i=1}^{m-1}(\sigma_i^2-\sigma_m^2)d\sigma_m &= \\ \gamma(n-m,\sigma_{m-1})\prod_{i=1}^{m-1}\sigma_i^2-\gamma(n-m+2,\sigma_{m-1})\sum_{i=1}^{m-1}\prod_{j\neq i}\sigma_j^2+\cdots \\ -\gamma(n-m+2(m-2),\sigma_{m-1})\sum_{i=1}^{m-1}\sigma_i^2+\gamma(n-m+2(m-1),\sigma_{m-1}).\end{aligned} \tag{12}$$

Note that since a is always a positive integer, we can easily evaluate the lower incomplete gamma function. $\gamma(1,x)=1-e^{-x}$. A simple recursion gives us the values for other positive integers ($a\geq 2$):

$$\gamma(a,x)=-x^{a-1}e^{-x}+(a-1)\gamma(a-1,x). \tag{13}$$

Expanding the recursion, we get

$$\gamma(a,x)=(a-1)!-\sum_{i=1}^{a}\frac{(a-1)!}{(i-1)!}x^{i-1}e^{-x}, \tag{14}$$

where $0!\equiv 1$. For the proof, substitute (14) for $\gamma(a-1,x)$ in (13). We can substitute this definition of the lower incomplete gamma function into (12) to obtain a form that is easy to integrate. Continuing in this manner, we can evaluate the full integral. Computational techniques are likely necessary for non-trivial values of m.

References

[1] A. Edelman. Jacobians of matrix transforms (with wedge products). http://web.mit.edu/18.325/www/handouts.html, February 2005. 18.325: Finite Random Matrix Theory, Handout #3.

[2] A. Edelman. Volumes and integration. http://web.mit.edu/18.325/www/handouts.html, March 2005. 18.325: Finite Random Matrix Theory, Handout #4.

[3] J. D. M. Rennie. On trace norm regularization for document modeling. http://people.csail.mit.edu/∼jrennie/writing, November 2005.

[4] E. W. Weisstein. Incomplete gamma function. http://mathworld.wolfram.com/IncompleteGammaFunction.html. From MathWorld–A Wolfram Web Resource.