Supplement: Learning the Irreducible Representations of Commutative Lie Groups

Taco Cohen Max Welling

1 Derivation of Conjugacy Relation

1.1 Regular von-Mises / uncoupled model

We show that a product of von-Mises distributions is a conjugate prior for the φ parameter of a toroidal parameterization of the Gaussian distribution.

$$\begin{aligned}
p(\varphi|\mathbf{x},\mathbf{y}) &\propto p(\mathbf{y}|\mathbf{x},\varphi)p(\varphi) \\
&\propto \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y}-\mathbf{W}\mathbf{R}(\varphi)\mathbf{W}^T\mathbf{x}\|^2\right)\exp\left(\sum_{j=1}^{J}\eta_j^T T(\varphi_j)\right) \\
&= \exp\left(-\frac{1}{2\sigma^2}\left(\|\mathbf{y}\|^2+\|\mathbf{W}\mathbf{R}(\varphi)\mathbf{W}^T\mathbf{x}\|^2-2\mathbf{y}^T\mathbf{W}\mathbf{R}(\varphi)\mathbf{W}^T\mathbf{x}\right)\right)\exp\left(\sum_{j=1}^{J}\eta_j^T T(\varphi_j)\right) \\
&= \exp\left(-\frac{1}{2\sigma^2}\left(\|\mathbf{y}\|^2+\|\mathbf{x}\|^2-2\mathbf{y}^T\mathbf{W}\mathbf{R}(\varphi)\mathbf{W}^T\mathbf{x}\right)\right)\exp\left(\sum_{j=1}^{J}\eta_j^T T(\varphi_j)\right) \\
&\propto \exp\left(\frac{1}{\sigma^2}\mathbf{y}^T\mathbf{W}\mathbf{R}(\varphi)\mathbf{W}^T\mathbf{x}\right)\exp\left(\sum_{j}\eta_j^T T(\varphi_j)\right) \\
&\propto \exp\left(\sum_{j=1}^{J}\frac{\mathbf{v}_j^T\mathbf{R}(\varphi_j)\mathbf{u}_j}{\sigma^2}+\eta_j^T T(\varphi_j)\right)
\end{aligned}$$

Recall that

$$\begin{aligned}
\mathbf{R}(\varphi_j) &= \begin{bmatrix}\cos(\varphi_j) & -\sin(\varphi_j)\\ \sin(\varphi_j) & \cos(\varphi_j)\end{bmatrix} \\
T(\varphi_j) &= [\cos(\varphi_j),\sin(\varphi_j)]^T
\end{aligned} \tag{1}$$

The coefficients in the bilinear form multiplying cosine and sine are $v_{j_1}u_{j_1}+v_{j_2}u_{j_2}$ and $v_{j_2}u_{j_1}-v_{j_1}u_{j_2}$ respectively. We can group these into a vector and dot it with $T(\varphi_j)=(\cos(\varphi_j),\sin(\varphi_j))^T$. Next, factor out the $T(\varphi_j)$ to obtain the result listed in the main paper:

$$p(\varphi|\mathbf{x},\mathbf{y}) \propto \exp\left(\sum_{j=1}^{J}\hat{\eta}_j^T T(\varphi_j)\right) \tag{2}$$

where

$$\hat{\eta}_j = \eta_j + \frac{1}{\sigma^2}[u_{j_1}v_{j_1} + u_{j_2}v_{j_2},\, u_{j_1}v_{j_2} - u_{j_2}v_{j_1}]^T \tag{3}$$

1.2 Generalized von-Mises / coupled model

$$\begin{aligned}
p(\varphi|\mathbf{x},\mathbf{y}) &\propto p(\mathbf{y}|\mathbf{x},\varphi)p(\varphi) \\
&\propto \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{W}\mathbf{R}(s)\mathbf{W}^T\mathbf{x}\|^2\right)\exp\left(\eta^+ \cdot T(s)\right) \\
&\propto \exp\left(\frac{\mathbf{v}^T\mathbf{R}(s)\mathbf{u}}{\sigma^2} + \eta^+ \cdot T(s)\right)
\end{aligned}$$

where now we have:

$$\begin{aligned}
\mathbf{R}(s) &= \begin{bmatrix} \cos(\omega_1 s) & -\sin(\omega_1 s) & & & \\ \sin(\omega_1 s) & \cos(\omega_1 s) & & & \\ & & \ddots & & \\ & & & \cos(\omega_J s) & -\sin(\omega_J s) \\ & & & \sin(\omega_J s) & \cos(\omega_J s) \end{bmatrix} \\
T(s) &= [\cos(s), \sin(s), \ldots, \cos(Ks), \sin(Ks)]^T.
\end{aligned} \tag{4}$$

Expanding the bilinear form, we find:

$$\frac{\mathbf{v}^T\mathbf{R}(s)\mathbf{u}}{\sigma^2} = \sum_{j=1}^{J} \frac{1}{\sigma^2}[(u_{j_1}v_{j_1} + u_{j_2}v_{j_2})\cos(\omega_j s) + (u_{j_1}v_{j_2} - u_{j_2}v_{j_1})\sin(\omega_j s)] \tag{5}$$

Factoring out the trigonometric functions of each unique frequency j yields the result in the paper:

$$\begin{aligned}
\hat{\eta}_j^+ &= \eta_j^+ + \sum_{k:\omega_k = j} \hat{\eta}_k \\
&= \eta_j^+ + \sum_{k:\omega_k = j} \frac{1}{\sigma^2}[u_{k_1}v_{k_1} + u_{k_2}v_{k_2},\, u_{k_1}v_{k_2} - u_{k_2}v_{k_1}]^T
\end{aligned} \tag{6}$$

2 Marginal likelihood

$$
\begin{aligned}
p(\mathbf{y}|\mathbf{x}) &= \int_{\varphi\in\mathbb{T}^J} \mathcal{N}(\mathbf{y}|\mathbf{W}\mathbf{R}(\varphi)\mathbf{W}^T\mathbf{x}) \prod_j \mathcal{M}(\varphi_j|\eta_j) d\varphi \\
&= \int_{\varphi\in\mathbb{T}^J} \frac{1}{\sqrt{(2\pi\sigma)^D}} \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{y}-\mathbf{W}\mathbf{R}(\varphi)\mathbf{W}^T\mathbf{x}\|^2\right) \prod_j \frac{\exp \eta_j^T T(\varphi_j)}{2\pi I_0(\|\eta_j\|)} d\varphi \\
&= \frac{\exp\left(-\frac{1}{2\sigma^2}(\|\mathbf{x}\|^2+\|\mathbf{y}\|^2)\right)}{\sqrt{(2\pi\sigma)^D}} \int_{\varphi\in\mathbb{T}^J} \exp\left(\sum_{j=1}^{J} \frac{\mathbf{v}_j^T\mathbf{R}(\varphi_j)\mathbf{u}_j}{\sigma^2} + \eta_j^T T(\varphi_j)\right) d\varphi \prod_j \frac{1}{2\pi I_0(\|\eta_j\|)} \\
&= \frac{\exp\left(-\frac{1}{2\sigma^2}(\|\mathbf{x}\|^2+\|\mathbf{y}\|^2)\right)}{\sqrt{(2\pi\sigma)^D}} \prod_j \int_0^{2\pi} \exp\left(\hat{\eta}_j^T T(\varphi_j)\right) d\varphi_j \frac{1}{2\pi I_0(\|\eta_j\|)} \\
&= \frac{\exp\left(-\frac{1}{2\sigma^2}(\|\mathbf{x}\|^2+\|\mathbf{y}\|^2)\right)}{\sqrt{(2\pi\sigma)^D}} \prod_j \frac{I_0(\|\hat{\eta}_j\|)}{I_0(\|\eta_j\|)}
\end{aligned}
\tag{7}
$$

The derivation for the coupled model is completely analogous, except that we now have to work towards a form where the GvM conjugacy can be used. Then the integral is simply the computation of the normalization constant of the GvM. That this normalization constant is equal to a Generalized Bessel Function (GBF) can be seen in the paper by Dattoli et al. cited in the main article, and is easily derived.

3 MAP inference

In order to perform MAP inference in the coupled model, we recast our real-valued model into complex form, and then apply the algorithm of Sohl-Dickstein et al directly. Recall that the transformations in our model are parameterized as $\mathbf{Q}(s) = \mathbf{W}\mathbf{R}(s)\mathbf{W}^T$, where $\mathbf{R}(s)$ is *block*-diagonal. Sohl-Dickstein et al. use a fully diagonal form, $\mathbf{Q}(s) = \mathbf{U}\mathbf{D}\mathbf{U}^{-1}$, where $\mathbf{U}$ is complex and not constrained in any way (except for the requirement of invertibility).

We can diagonalize our model by performing an eigendecomposition of the generator matrix $\mathbf{A} = \sum_j \mathbf{A}_j$. This matrix is zero everywhere, except for the entries $a_{2j,\,2j-1} = 1$ and $a_{2j-1,\,2j} = -1$. Let $\mathbf{E}$ be the eigenvectors of this matrix, then $\mathbf{U} = \mathbf{W}\mathbf{E}$.

The diagonal matrix $\mathbf{D}$ is given by the matrix exponential of a diagonal matrix, $\mathbf{D} = \exp(s\Lambda)$. To get a model that is equivalent to TSA, set $\Lambda_{2j-1} = i\omega_j$ and $\Lambda_{2j-1} = -i\omega_j$. One can verify that the matrices $\mathbf{U}\mathbf{D}\mathbf{U}^{-1}$ and $\mathbf{W}\mathbf{R}(s)\mathbf{W}^T$ are indeed the same, numerically. Now, we can use Sohl-Dicksteins method to obtain a MAP estimate of s. It is most likely not too hard to apply the basic idea of S-D et al. directly to TSA, but we have not tried this yet.

To get good results using the method of S-D et al., we found that it is necesary to initialize the smoothing parameter σ at a value of around 3. Much higher and we run into numerical issues due to the exponentiation, and much lower (e.g. 0) and the algorithm will not find a good minimum.

4 Computation of Generalized Bessel Functions

One way to define the modified Generalized Bessel functions (GBF) is as follows:

$$I_n(x_1, \ldots, x_M) = \sum_{l=-\infty}^{\infty} I_{n-Ml}(x_1, \ldots, x_{M-1}) I_l(x_M) \tag{8}$$

where the scalar function $I_n(x_i)$ is the modified Bessel function or order n. Analogous functions can be defined by replacing some or all of the I_l in this recursion by the ordinary Bessel function J_l, but here we will only focus on the pure-I form.

A set of complex *parameters* τ_m $(m = 1, \ldots M - 1)$ may be included in the definition:

$$I_n(x_1, \ldots, x_M; \tau_1, \ldots, \tau_{M-1}) = \sum_{l=-\infty}^{\infty} I_{n-Ml}(x_1, \ldots, x_{M-1}; \tau_1, \ldots, \tau_{M-2}) I_l(x_M) \tau_{M-1}^l \tag{9}$$

In the main paper we use a slightly different convention where we use M parameters:

$$I_n(x_1, \ldots, x_M; \tau_1, \ldots, \tau_M) = \sum_{l=-\infty}^{\infty} I_{n-Ml}(x_1, \ldots, x_{M-1}; \tau_1, \ldots, \tau_{M-1}) I_l(x_M) \tau_M^l \tag{10}$$

so that the two-variable two-parameter GBF takes the form

$$I_n(x_1, x_2; \tau_1, \tau_2) = \sum_{l=-\infty}^{\infty} I_{n-2l}(x_1) \tau_1^{n-2l} I_l(x_M) \tau_M^l$$

The GBF plays an important role in many areas of physics, and many analytical results are known. However, at this point not much computational work has been done. The only works we have found, [3] and [2], focus on 2-variable GBF, which is not nearly enough for our purposes. The authors of [2] derive a fast and accurate but rather complicated method that works only for 2-variable GBFs. According to [1], the computation of GBF is considered difficult: "[..] gives rise to so-called generalized Bessel functions which are infinite sums of ordinary Bessel functions and have proven to be notoriously difficult to evaluate."

5 Algorithm for the computation of GBF

The modified Bessel function $I_n(x_i)$ has the property that it tends towards zero as $n \to \infty$ and as $n \to -\infty$, so we can truncate the sum:

$$I_n(x_1, \ldots, x_M) \approx \sum_{l=-L}^{L} I_{n-Ml}(x_1, \ldots, x_{M-1}) I_l(x_M) \tag{11}$$

However, a naive implementation of this function using a recursion or nested summation has a runtime complexity that is exponential in the number of variables M, because each of the $2L + 1$ values $l \in [-L, L]$ will call a GBF of order

$M-1$, which itself makes $2L+1$ calls, and so on. This yields a complexity of $O(L^M)$, which is intractable as the number of variables M grows large.

A much faster algorithm can be obtained by creating a function that uses convolutions to compute the GBF, and can return an array of GBFs of different orders n at once. The algorithm proceeds as follows. First obtain an array of $(M-1)$-variable bessel functions of different orders using a single recursive call:

$$I_{M-1} = \{I_{-L}(x_1, \dots, x_{M-1}), \dots, I_L(x_1, \dots, x_{M-1})\}. \tag{12}$$

Next, create an array

$$I_M = \{I_{-L/M}(x_M), 0, \dots, 0, I_{-L/M+1}(x_M), 0, \dots, 0, I_{L/M+1}(x_M)\} \tag{13}$$

where each gap of zeros has length $M-1$, and we have assumed that M divides L evenly. The function $I_n(x_1, \dots, x_M)$ can now be obtained as the n-th element away from the center of the convolution $I_M * I_{M-1}$. In other words, by computing the convolution, we get the GBF for all orders at once. GBF with parameters can be computed simply by multplying each element of the array I_M by the appropriate power of by τ_{M-1} or τ_M (depending on the convention).

The convolutions can be performed in time $L \log L$ by doing a fast Fourier transform of the arrays, then computing their product and then computing the inverse transform on that. The computational complexity of the full method is then $O(ML \log L)$: for M variables there are $O(M)$ convolutions to be performed, each one taking $O(L \log L)$.

References

[1] Ulrich D. Jentschura, High-Intensity Lasers and the Furry-Volkov Picture http://web.mst.edu/ jentschurau/activities.html

[2] E. Lotstedt, U. D. Jentschura, Phys. Rev. E 79, 026707 (2009)

[3] Analytical and numerical results on M-variable generalized Bessel functions Dattoli, G. Mari, C. Torre, A. Chiccoli, C. Lorenzutta, S. Maino, G. Journal of Scientific Computing, 1992.