On Wasserstein Gaussian Barycenters

Report completed by Wessel Bruinsma and Gabriel Arpino at Invenia Labs, 2018

Contents

1 Abstract

In this report, we study the problem of averaging gaussian probabilistic models through Wasserstein Barycenters. We compare the approach to the standard moment matching technique, and prove that the Wasserstein Barycenter of gaussian distributions produce lower entropy averages, hinting at its applicability in averaging distributions with latent bias. We analyze the currently used fixed point approximation solution to the problem, and offer the first closed-form solution by drawing connections to Optimal Control theory. Through preliminary experiments we demonstrate the usefulness of the approach and hint at a phase transition point where the Wasserstein Barycenter becomes more accurate as a method of inference over standard moment matching. [1]

2 Wasserstein Gaussian Barycenters

Let $x_i \sim \mathcal{N}(\mu_i, K_i)$. We wish to compute a Gaussian average of (x_i).

[1] We would like to thank Dr. Matthew Thorpe for helpful discussions and mentorship.

Define the *moment-matched average* to be

$$\overline{x}_{\mathrm{MM}} \mid (x_i), (\alpha_i) \sim \mathcal{N}\left(\sum_i \alpha_i \mu_i, \sum_i \alpha_i K_i\right),$$

which can be interpreted as the Gaussian distribution closest, in the KL sense, to the mixture distribution with components (x_i) and weights (α_i). Further define the 2-*Wasserstein average* to be

$$\overline{x}_{W_2} \mid (x_i), (\alpha_i) = \arg\min_x \sum_i \alpha_i d^2_{W_2}(x, x_i)$$

where d_{W_2} is the 2-Wasserstein distance employing the Euclidean distance, and where the minimisation is understood over all probability measures with finite second moments.

It can be shown in **Mallasto2017LearningFromUncertainCurvesThe** that

$$\overline{x}_{W_2} \mid (x_i), (\alpha_i) \sim \mathcal{N}\left(\sum_i \alpha_i \mu_i, K\right)$$

where K is the unique positive definite matrix satisfying

$$K = \sum_i \alpha_i (K^{\frac{1}{2}} K_i K^{\frac{1}{2}})^{\frac{1}{2}}.$$

2.1 On $XAX = B$

Theorem 2.1 (Shurbet **Shurbet:1974:QuadraticMatrixEquations**). The matrix equation $XAX = B$ has a solution if and only if $(AB)^{\frac{1}{2}}$ exists and

$$AA^-(AB)^{\frac{1}{2}} = (AB)^{\frac{1}{2}},$$
$$B(AB)^{\frac{1}{2}-}(AB)^{\frac{1}{2}} = B,$$

where A^- denotes a generalised inverse of A. If $XAX = B$ has a solution, then it is of the form

$$X = A^-(AB)^{\frac{1}{2}} + (I - A^-A)\{B(AB)^{\frac{1}{2}-} + U[I - (AB)^{\frac{1}{2}}(AB)^{\frac{1}{2}-}]\}$$

where U is arbitrary. ◀

If A and B are positive definite, then the answer to $XAX = B$ simplifies considerably.

Theorem 2.2. Let A and B be positive definite, and consider $XAX = B$.

(1) The unique positive definite solution is

$$X = A^{-\frac{1}{2}} \left(A^{\frac{1}{2}} B A^{\frac{1}{2}} \right)^{\frac{1}{2}} A^{-\frac{1}{2}}.$$

(2) If v and w are eigenvectors of X, then

$$\lambda_v \lambda_w \langle v, Aw \rangle = \langle v, Bw \rangle, \qquad \lambda_v^2 = \frac{\langle v, Bv \rangle}{\langle v, Av \rangle}.$$

◀

Proof. (1): Let $X = A^{-\frac{1}{2}} Y A^{-\frac{1}{2}}$. This eliminates the inner A: $A^{-\frac{1}{2}} Y Y A^{-\frac{1}{2}} = B \iff YY = A^{\frac{1}{2}} B A^{\frac{1}{2}}$. But $A^{\frac{1}{2}} B A^{\frac{1}{2}}$ is positive definite, so it has a unique square root: $Y = \left(A^{\frac{1}{2}} B A^{\frac{1}{2}} \right)^{\frac{1}{2}}$. The expression for X now follows. A similar argument can be found on MathOverflow.[2]

(2): Note that

$$\langle v, XAXw \rangle = \langle v, Bw \rangle \implies \lambda_v \lambda_w \langle v, Aw \rangle = \langle v, Bw \rangle,$$

and for the second statement use that A and B are positive definite. Q.E.D.

2.2 Results on Gaussian Barycenters

Theorem 2.3. There exist unique positive definite (A_i) such that

$$I = \sum_i \alpha_i A_i, \qquad \text{and} \qquad A_i K A_i = K_i \quad \text{for all } i.$$

◀

Proof. We indeed know that such A_i uniquely exist. Then note that

$$(K^{\frac{1}{2}} A_i K^{\frac{1}{2}})(K^{\frac{1}{2}} A_i K^{\frac{1}{2}}) = K^{\frac{1}{2}} A_i K A_i K^{\frac{1}{2}} = K^{\frac{1}{2}} K_i K^{\frac{1}{2}} \implies (K^{\frac{1}{2}} K_i K^{\frac{1}{2}})^{\frac{1}{2}} = K^{\frac{1}{2}} A_i K^{\frac{1}{2}}$$

[2] https://math.stackexchange.com/questions/1575929/generalized-square-root-of-a-real-positive-definite-symmetric-matrix-with-resp

since a PD matrix has a unique PD square root. Hence,

$$K = K^{\frac{1}{2}} K^{\frac{1}{2}} = \sum_i \alpha_i K^{\frac{1}{2}} A_i K^{\frac{1}{2}} \implies I = \sum_i \alpha_i A_i.$$

Q.E.D.

In the case $n = 2$, one can image the following fixed-point iteration scheme:

- $A_2 \to K$ using $K = A_2^{-1} K_2 A_2^{-1}$;
- $K \to A_1$ using $A_1 K A_1 = A_1 A_2^{-1} K_2 A_2^{-1} A_1 = K_1$;
- $A_1 \to A_2$ using $A_2 = \alpha_2^{-1} I - \alpha_2^{-1} \alpha_1 A_1$.

In formulas,

- $X \leftarrow \left(A_2^{-1} K_2 A_2^{-1}\right)^{\frac{1}{2}}$;
- $A_1 \leftarrow X^{-1} (X K_1 X)^{\frac{1}{2}} X^{-1}$;
- $A_2 \leftarrow \alpha_2^{-1} I - \alpha_2^{-1} \alpha_1 A_1$.

Theorem 2.4. There exist (C_i), given by

$$C_j = A_n A_j^{-1} = (\alpha_n^{-1} I - \sum_{i \neq n} \alpha_i \alpha_n^{-1} A_i) A_j^{-1},$$

such that

$$K = (\alpha_1 I + \sum_{i \neq 1, n} \alpha_i C_i^{-1} C_1 + \alpha_n C_1) K_1 (\alpha_1 I + \sum_{i \neq 1, n} \alpha_i C_1^{\mathsf{T}} C_i^{-\mathsf{T}} + \alpha_n C_1^{\mathsf{T}})$$

and $C_i K_i C_i^{\mathsf{T}} = K_n$ for $i \neq n$. ◀

Proof. Expressing $A_n = \alpha_n^{-1} I - \sum_{i \neq n} \alpha_i \alpha_n^{-1} A_i$, we find

$$K_n = (\alpha_n^{-1} I - \sum_{i \neq n} \alpha_i \alpha_n^{-1} A_i) K (\alpha_n^{-1} I - \sum_{i \neq n} \alpha_i \alpha_n^{-1} A_i).$$

Let $j \neq n$. Then $K = A_j^{-1} K_j A_j^{-1}$, so

$$K_n = \underbrace{(\alpha_n^{-1} I - \sum_{i \neq n} \alpha_i \alpha_n^{-1} A_i) A_j^{-1}}_{C_j} K_j A_j^{-1} (\alpha_n^{-1} I - \sum_{i \neq n} \alpha_i \alpha_n^{-1} A_i) = C_j K_j C_j^{\mathsf{T}},$$

using that (A_i), hence their inverses, are symmetric. Rewriting

$$\alpha_n C_j A_j + \sum_{i\neq n} \alpha_i A_i = I,$$

which, given (C_j), is just a system linear in (A_i):

$$\begin{bmatrix}\alpha_1 I & \cdots & \alpha_{j-1} I & \alpha_n C_j + \alpha_j I & \alpha_{j+1} I & \cdots & \alpha_{n-1} I\end{bmatrix} \underbrace{\begin{bmatrix} A_1^\top & \cdots & A_{n-1}^\top \end{bmatrix}^\top}_{A} = I.$$

Let B be the block column matrix with $\alpha_i I$ as row element i; let D be the block diagonal matrix with $\alpha_n C_i$ as diagonal element i; and let J be I vertically concatenated n times. Then

$$(JB^\top + D)A = J \implies A = (JB^\top + D)^{-1} J,$$

so by application of the matrix lemma,

$$A = D^{-1}J(I - (I + B^\top D^{-1} J)^{-1} B^\top D^{-1} J) = D^{-1} J (I + B^\top D^{-1} J)^{-1}.$$

Hence

$$\begin{aligned} &A_j = \alpha_n^{-1} C_j^{-1} (I + B^\top D^{-1} J)^{-1} \\ \implies &A_j^{-1} = (\alpha_n I + \sum_{i\neq n} \alpha_i C_i^{-1}) C_j = \alpha_j I + \sum_{i \neq j,n} \alpha_i C_i^{-1} C_j + \alpha_n C_j \end{aligned}$$

and, by symmetry of A_1 and letting $j = 1$,

$$\begin{aligned} K &= A_1^{-1} K_1 A_1^{-\top} \\ &= (\alpha_1 I + \sum_{i\neq 1,n} \alpha_i C_i^{-1} C_1 + \alpha_n C_1) K_1 (\alpha_1 I + \sum_{i \neq 1,n} \alpha_i C_1^\top C_i^{-\top} + \alpha_n C_1^\top). \end{aligned}$$

Q.E.D.

Corollary 2.1. If $n = 2$, then

$$K = \alpha_1^2 K_1 + \alpha_2^2 K_2 + \alpha_1 \alpha_2 \left(K_1^{\frac{1}{2}} \left(K_1^{\frac{1}{2}} K_2 K_1^{\frac{1}{2}} \right)^{\frac{1}{2}} K_1^{-\frac{1}{2}} + K_1^{-\frac{1}{2}} \left(K_1^{\frac{1}{2}} K_2 K_1^{\frac{1}{2}} \right)^{\frac{1}{2}} K_1^{\frac{1}{2}} \right).$$

◀

Proof. If $n = 2$, then

$$C_1 = \alpha_2^{-1} A_1^{-1} - \alpha_1 \alpha_2^{-1} I = A_2 A_1^{-1},$$

so C_1 is symmetric, and we further show that C_1 has positive eigenvalues:

$$C_1 A_2^{\frac{1}{2}} = A_2^{\frac{1}{2}} (A_2^{\frac{1}{2}} A_1 A_2^{\frac{1}{2}}).$$

Thus C_1 is positive definite, satisfies $C_1 K_1 C_1 = K_2$, hence must be given by

$$C_1 = K_1^{-\frac{1}{2}} \left(K_1^{\frac{1}{2}} K_2 K_1^{\frac{1}{2}} \right)^{\frac{1}{2}} K_1^{-\frac{1}{2}}.$$

Plugging back into the expression for K, we find

$$\begin{aligned} K &= (\alpha_2 C_1 + \alpha_1 I) K_1 (\alpha_2 C_1 + \alpha_1 I) \\ &= \alpha_1^2 K_1 + \alpha_2^2 C_1 K_1 C_1 + \alpha_1 \alpha_2 (C_1 K_1 + K_1 C_1) \\ &= \alpha_1^2 K_1 + \alpha_2^2 K_2 + \alpha_1 \alpha_2 \left(K_1^{\frac{1}{2}} \left(K_1^{\frac{1}{2}} K_2 K_1^{\frac{1}{2}} \right)^{\frac{1}{2}} K_1^{-\frac{1}{2}} + K_1^{-\frac{1}{2}} \left(K_1^{\frac{1}{2}} K_2 K_1^{\frac{1}{2}} \right)^{\frac{1}{2}} K_1^{\frac{1}{2}} \right). \end{aligned}$$

Q.E.D.

Corollary 2.2. If v is an eigenvector of A_1—recall that $A_1 K A_1 = K_1$—then

$$\frac{\langle v, K_1 v \rangle}{\langle v, K v \rangle} \leq \alpha_1^2.$$

◀

Proof. From $\alpha_2 C_1 = A_1^{-1} - \alpha_1 I \geq 0$ we find that

$$\lambda_{\min}(A_1^{-1}) \geq \alpha_1 \implies \lambda_{\max}(A_1) \leq \alpha_1.$$

Therefore, if v is an eigenvector of A_1 with eigenvalue λ, then

$$\lambda^2 = \frac{\langle v, K_1 v \rangle}{\langle v, K v \rangle} \leq \alpha_1^2.$$

Q.E.D.

Corollary 2.3.

$$\begin{aligned}
&\alpha_1 K_1 + \alpha_2 K_2 - K \\
&\quad = \alpha_1\alpha_2 \left(K_1 + K_2 - K_1^{\frac{1}{2}} \left(K_1^{\frac{1}{2}} K_2 K_1^{\frac{1}{2}} \right)^{\frac{1}{2}} K_1^{-\frac{1}{2}} - K_1^{-\frac{1}{2}} \left(K_1^{\frac{1}{2}} K_2 K_1^{\frac{1}{2}} \right)^{\frac{1}{2}} K_1^{\frac{1}{2}} \right) \\
&\quad = \alpha_1\alpha_2 \left(K_1^{-\frac{1}{2}} \left(K_1^{\frac{1}{2}} K_2 K_1^{\frac{1}{2}} \right)^{\frac{1}{2}} K_1^{-\frac{1}{2}} - I \right) K_1 \left(K_1^{-\frac{1}{2}} \left(K_1^{\frac{1}{2}} K_2 K_1^{\frac{1}{2}} \right)^{\frac{1}{2}} K_1^{-\frac{1}{2}} - I \right) \\
&\quad \geq 0.
\end{aligned}$$

◀

And this result neatly generalises in a slightly different form.

Theorem 2.5.

$$\begin{aligned}
&\sum_i \alpha_i K_i - K \\
&\quad = \sum_{i \neq n} \alpha_i \alpha_n (I - C_i^{-1}) K_n (I - C_i^{-\mathsf{T}}) + \sum_{i \neq n} \sum_{j=i+1}^{n-1} \alpha_i \alpha_j (C_i^{-1} - C_j^{-1}) K_n (C_i^{-\mathsf{T}} - C_j^{-\mathsf{T}}) \\
&\quad \geq 0.
\end{aligned}$$

◀

Proof.

$$\begin{aligned}
&\sum_i \alpha_i K_i - K \\
&\quad= \sum_i \alpha_i K_i - (\alpha_n I + \sum_{i\neq n} \alpha_i C_i^{-1}) C_1 K_1 C_1^\top (\alpha_n I + \sum_{i\neq n} \alpha_i C_i^{-\top}) \\
&\quad= \sum_i \alpha_i K_i - (\alpha_n I + \sum_{i\neq n} \alpha_i C_i^{-1}) K_n (\alpha_n I + \sum_{i\neq n} \alpha_i C_i^{-\top}) \\
&\quad= \sum_i (\alpha_i - \alpha_i^2) K_i - \sum_{i\neq n} \alpha_i \alpha_n (C_i^{-1} K_n + K_n C_i^{-\top}) - \sum_{i\neq n} \sum_{j\neq n,\, j\neq i} \alpha_i \alpha_j C_i^{-1} K_n C_j^{-\top} \\
&\quad= \sum_i \sum_{j\neq i} \alpha_i \alpha_j K_i \\
&\qquad - \sum_{i\neq n} \alpha_i \alpha_n (C_i^{-1} K_n + K_n C_i^{-\top}) \\
&\qquad - \sum_{i\neq n} \sum_{j=i+1}^{n-1} \alpha_i \alpha_j (C_i^{-1} K_n C_j^{-\top} + C_j^{-1} K_n C_i^{-\top}) \\
&\quad= \sum_{i\neq n} \alpha_i \alpha_n (K_i + K_n - C_i^{-1} K_n - K_n C_i^{-\top}) \\
&\qquad + \sum_{i\neq n} \sum_{j=i+1}^{n-1} \alpha_i \alpha_j (K_i + K_j - C_i^{-1} K_n C_j^{-\top} - C_j^{-1} K_n C_i^{-\top}) \\
&\quad= \sum_{i\neq n} \alpha_i \alpha_n (C_i^{-1} K_n C_i^{-\top} + K_n - C_i^{-1} K_n - K_n C_i^{-\top}) \\
&\qquad + \sum_{i\neq n} \sum_{j=i+1}^{n-1} \alpha_i \alpha_j (C_i^{-1} K_n C_i^{-\top} + C_j^{-1} K_n C_j^{-\top} - C_i^{-1} K_n C_j^{-\top} - C_j^{-1} K_n C_i^{-\top}) \\
&\quad= \sum_{i\neq n} \alpha_i \alpha_n (I - C_i^{-1}) K_n (I - C_i^{-\top}) + \sum_{i\neq n} \sum_{j=i+1}^{n-1} \alpha_i \alpha_j (C_i^{-1} - C_j^{-1}) K_n (C_i^{-\top} - C_j^{-\top}).
\end{aligned}$$

Q.E.D.

Theorem 2.6. If (K_i) are all simultaneously diagonalisable with eigenvectors U and eigenvalues Λ_i, then

$$K = U \left(\sum_i \alpha_i \Lambda_i^{\frac{1}{2}} \right)^2 U^\top.$$

◀

Proof. Set $\Lambda = U^\mathsf{T} K U$. Then

$$U \Lambda U^\mathsf{T} = \sum_i \alpha_i \left(U \Lambda^{\frac{1}{2}} U^\mathsf{T} U \Lambda_i U^\mathsf{T} U \Lambda^{\frac{1}{2}} U^\mathsf{T} \right)^{\frac{1}{2}} \iff \Lambda = \sum_i \alpha_i \Lambda^{\frac{1}{2}} \Lambda_i^{\frac{1}{2}},$$

from which the result directly follows. Q.E.D.

2.3 A Generative Model for Barycenters

$$\begin{aligned} K &\sim \mathcal{W}^{-1}, \\ (\alpha_1, \ldots, \alpha_n) &\sim \mathrm{Dir}, \\ (\hat{A}_1, \ldots, \hat{A}_n) &\sim \mathrm{Dir}, \\ A_i \,|\, \hat{A}_i, \alpha_i &= \alpha_i^{-1} \hat{A}_i \\ K_i \,|\, K, A_i &= A_i K A_i \\ x_j^i \,|\, K_i &\sim \mathcal{N}(0, K_i). \end{aligned}$$

Then, for every sample, K is the (α_i)-weighted barycenter of (K_i).

2.4 Preliminary Experiments

Experiment 1

We create toy data points by sampling from high dimensional gaussian processes, corrupt these with fat tailed noise (Cauchy), and fit each sample with another gaussian process. The objective is to observe which method, the Wasserstein or the Euclidean Barycenter, best approximates the true covariance matrix of the originally sampled gaussian process. In figure 1, we observe that as the magnitude of the fat tailed noise increases, the norm of the difference between the true covariance matrix and the estimated covariance matrix of the Euclidean Barycenter becomes large, and in the limit of the fat tailed noise increasing, the covariance matrix obtained by the Wasserstein Barycenter becomes closer in norm to the true matrix relative to the Euclidean Barycenter.

This indicates that there is a phase transition point where the Wasserstein Barycenter better rejects fat tailed noise and conducts more accurate inference over the parameters of the true covariance matrix.

Experiment 2

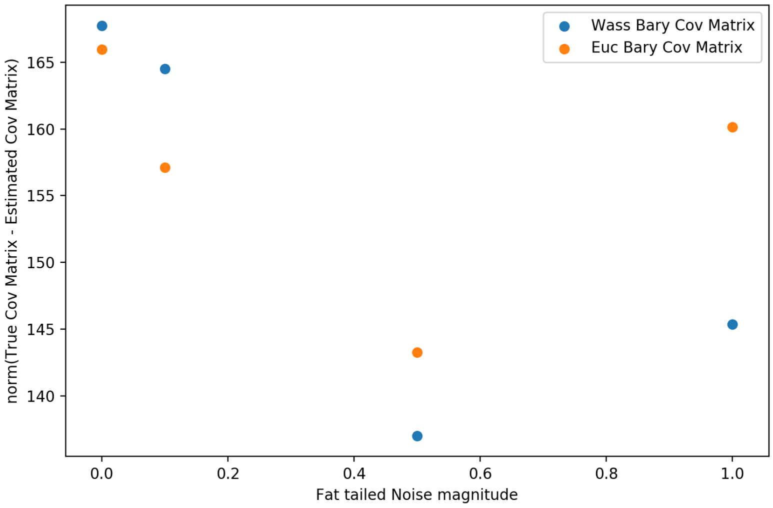

Figure 1: Experiment 1

In this experiment, we create 3 hypothetical "sensors", each with a gaussian distribution over the incoming signal, parametrized by zero mean and a common covariance matrix. To each sensor covariance matrix, we add small components of noise kernels of different parameters (Matern32 kernel, $\alpha = 0.5, 1.0, 5.0, 25.0$). The experiment consisted of observing whether the Wasserstein Barycenter of these sensor signal distributions produced a covariance matrix closer to the true common covariance matrix than the Euclidean Barycenter. Results showed that the Wasserstein Barycenter produced covaraince matrices that were always closer to the true common covariance matrix than the Euclidean Barycenter. Figures to be added.