Probable and Improbable Faces

J.P. Lewis, Zhenyao Mo, Ken Anjyo, Taehyun Rhee

Abstract

The multivariate normal is widely used as the expected distribution of face shape. It has been used for face detection and tracking in computer vision, as a prior for facial animation editing in computer graphics, and as a model in psychological theory. In this contribution we consider the character of the multivariate normal in high dimensions, and show that these applications are not justified. While we provide limited evidence that facial proportions are not Gaussian, this is tangential to our conclusion: even if faces are truly "Gaussian", maximum a posteriori and other applications and conclusions that assume that typical faces lie near the mean are not valid.

1 Introduction

In computer vision and graphics research, facial expression and identity are commonly modeled as a high-dimensional vector space, often with a multidimensional Gaussian density. This choice of representation has associated algorithmic approaches such as linear interpolation and maximum a posteriori (MAP) solution of inverse problems.

In this paper we argue several things: 1) the linear and Gaussian assumptions are not strictly correct. 2) existing research that starts from these assumptions has implicitly assumed a low dimensional setting. In high dimensions, common algorithmic approaches such as MAP may not be justified. 3) the problems resulting from these assumptions are not just hypothetical, but are visible in a practical computation, specifically interpolation of faces. Most importantly, we show that consideration of these factors can result in an algorithm with visibly improved results.

2 Linear models

The faces of realistic computer characters in movies are most often generated using the "blendshape" representation [6, 1, 15, 7]. This is a linear representation of the form $\mathbf{f} = \mathbf{B}\mathbf{w}$, where $\mathbf{B}$ is a linear but non-orthogonal basis having semantic meaning. In computer vision, approaches such as active appearance models (AAM) [4] and morphable models [2] use an orthogonal basis generated by principal component analysis (PCA), and assume the multidimensional Gaussian prior. Bilinear (tensor) face models have also been proposed [17]. Psychological research has also employed such linear models with a multivariate Gaussian prior [16].

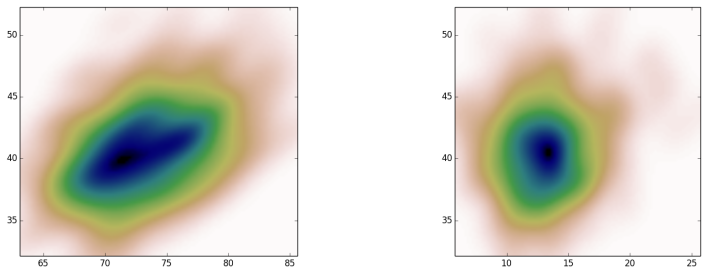

Figure 1: Face proportions are not strictly Gaussian. Kernel density plots of (left) the distance between the eyes versus the width of the mouth, (right) the width of the mouth versus the height of the mouth, measured from a database of 359 faces.

PCA assumes that the data is jointly Gaussian, in that the PCA basis vectors are the eigenvectors of a covariance matrix that does not capture any non-Gaussian statistics. The Gaussian assumption leads to a frequently employed prior or regularizer of the form $\mathbf{c}^T \mathbf{\Lambda}^{-1} \mathbf{c}$ where $\mathbf{c}$ is the vector of PCA coefficients and $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues. The eigenvalues are variances of the data in the directions of the eigenvectors. The Gaussian assumption also naturally leads to the MAP approach to regularising inverse problems. This approach selects model parameters M as the mode of the posterior $P(D|M)P(M)$ given data D. With a Gaussian model the posterior also has a Gaussian form.

The appropriate number of dimensions for a linear facial model of expression or identity has been variously estimated to be in the range 40–100 [10, 13, 9]. High quality blendshape facial models used in movie visual effects sometimes have on the order of 100 dimensions [6]. The main character in the 3D animated movie *Toy Story* had 212 parameters controlling the head [5].

In figure 1 we show that the common multidimensional Gaussian assumption is not strictly accurate. This figure shows a kernel density plot of several simple measurements of facial proportions measured from 359 selected photographs from the facial database [14]. It is also somewhat obvious that a linear model is not entirely appropriate for facial *expression*. For example, the motion of the jaw has a clear rotational component. On the other hand, the widespread use of the blendshape representation in movies (albeit sometimes with nonlinear correction terms [15]) is an argument that linear models suffice even if they are not strictly accurate. It is less clear whether a vector space model of facial *identity* is appropriate, or if a (nonlinear) manifold assumption would be more accurate. While these comments call into question the linear and Gaussian assumptions, existing research does not indicate whether these objections are important in practical computations.

3 High-dimensional Phenomena

High dimensional data is generally subject to a collection of nonintuitive phenomena collectively known as the "curse of dimensionality" [18]. Examples of such phenomena are that a) in high dimensions, "all data is far away" with high proba-

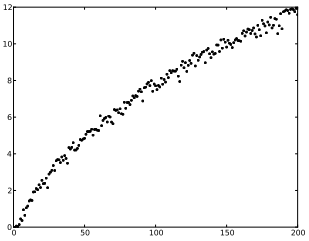

Figure 2: The closest distance to the mean among 1000 unit-variance multidimensional Gaussian random variables (vertical axis) as a function of the dimension (horizontal axis). In 100 dimensions every point in this simulation is more than six standard deviations from the mean.

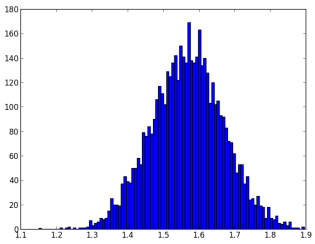

Figure 3: Histogram of the angles between all pairs of 100 randomly chosen isotropic Gaussian random variables in 100 dimensions. The angles cluster around $\pi/2$: in high dimensions, most data are nearly orthogonal.

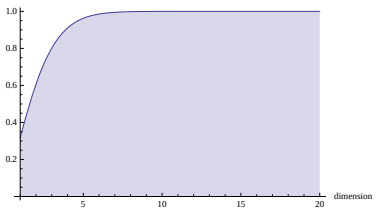

Figure 4: Probability that a sample from a unit variance Gaussian is outside the unit hypersphere for various dimensions.

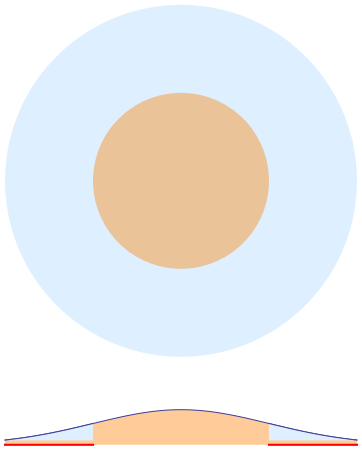

Figure 5: Bottom: schematic one dimensional Gaussian distribution, with the area between one and two deviations indicated in red. This interval is equal to that of the unit radius. Top: In two dimensions, the area between one and two standard deviations (light blue) is relatively larger than the area of the unit standard deviation disc (light orange). Figure is best viewed in the electronic version of this document.

bility (Figure 2), b) randomly chosen vectors are nearly orthogonal (Figure 3), and c) the probability mass of the data is overwhelmingly located near the surface of the hypervolume, with the interior of the volume essentially empty (Figs. 5, 8).

Current face computation approaches generally overlook these phenomena. A notable exception is [12], who described the following apparent paradox: the squared Mahalanobis distance $\mathbf{c}^T \mathbf{\Lambda}^{-1} \mathbf{c}$ follows a χ^2 distribution with n degrees of freedom, since it is the sum of independent, identically distributed (i.i.d.) squared Gaussian variables of variance $\frac{c_i^2}{\lambda_i}$. The expectation of this distribution for d dimensions is d, thus we expect the length of the standardized squared coefficient vector of a typical face to be d. However under the multidimensional Gaussian model, the face at the origin (the mean face) is the most probable, and the length of its squared coefficient vector is zero.

[12] also state a hypothesis that faces should lie on the shell of a hyperellipsoid dictated by the squared coefficient length. The resolution to the apparent paradox is simply that it is the difference between the variance and mean. A zero-mean random variable can (and typically does!) have a nonzero variance. Randomly sampling from a multidimensional Gaussian will generate a sequence of samples that have *both* the expected mean and variance of course.

4 The High-Dimensional Gaussian Prior

Next we will verify the statement that high dimensional data is concentrated overwhelmingly near the surface of the hypervolume. In the case of a uniformly distributed random variable in a hypercube, this is easy to see. Consider a unit hypercube in d dimensions, that encloses a smaller hypercube of side $1-\varepsilon$. As $d \to \infty$,

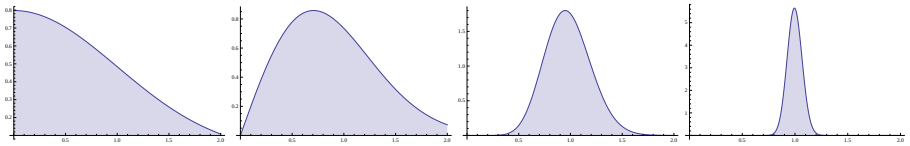

Figure 6: The radially integrated Gaussian $N(0, \mathbf{I}_n)$ in various dimensions. Each subfigure shows the radially integrated Gaussian profile $S_{d-1}(r)G(r)$ (vertical axis) plotted in units of $\sqrt{d}$ (horizontal axis). From left to right: 1, 2, 10, and 100 dimensions. In high dimensions the probability concentrates in a shell centered at radius $\sqrt{d}$.

the volume of the enclosed hypercube is $(1-\varepsilon)^d \to 0$.

The fact that the multivariate Gaussian is a heavy tailed distribution in high dimensions is less obvious. For example, [16] states, "even for a face space of high dimensionality, the assumption of a multivariate normal distribution means that... There will be many typical faces that will be located relatively close to the center". However this phenomenon is at least intuitively suggested by comparing the one- and two-dimensional Gaussian distributions (Figure 5). In one dimension, the "volume" of the interval between one and two standard deviations is equal to the radius of the unit interval. In two dimensions the area of the annulus between one and two standard deviations is relatively larger than the area of the unit disc. In higher dimensions the trend continues, with the available volume overwhelmingly concentrated near the outside.

Discussion of the multivariate Gaussian is simplified by a "whitening" transformation $c_i \to c_i/\sqrt{\lambda_i}$ from the original hyperellipsoidal density to an isotropic density. We can also consider a unit-variance density without loss of generality. In this case the probability that a point is within a hypersphere of radius r is proportional to

$$\int_0^r S_{d-1}(r)G(r) = \frac{2\pi^{d/2}}{\Gamma(d/2)} \int_0^r r^{d-1} G(r) dr$$

where d is the dimension, $G(r) = \frac{1}{\sqrt{(2\pi)^d}} \exp^{-r^2/2}$ is the isotropic unit variance Gaussian density function, $S_{d-1}(r) = \frac{2\pi^{d/2} r^{d-1}}{\Gamma(d/2)}$ is the "surface area" of the d-hypersphere, and Γ is the Gamma function. This can be used to plot the tail probability that a point lies outside the unit hypersphere in various dimensions (Figure 4). While in one dimension the majority of the probability mass is within the unit interval, in 100 dimensions the probability that a point is outside the unit hypersphere is 1. to within machine precision! It may be worth contrasting the mode of the high-dimensional Gaussian with the Dirac delta generalised function familiar in signal processing [3]. The delta function has zero width but unit volume when integrated over. In contrast, the high-dimensional Gaussian has nonzero width near the origin, but negligible volume.

High dimensional data can also be tightly concentrated in a shell of relatively narrow thickness. In the case of the multi-dimensional Gaussian, the majority of its mass is concentrated within a shell centered at radius $\sqrt{d}$. Figure 6 plots the radially integrated unit variance Gaussian profile $S_{d-1}(r)G(r)$ relative to the distance $\sqrt{d}$ (i.e. with a change of variable $r \to r\sqrt{d}$). The data is concentrated increasingly around $\sqrt{d}$ (relative to the distance $\sqrt{d}$ itself) in high dimensions.

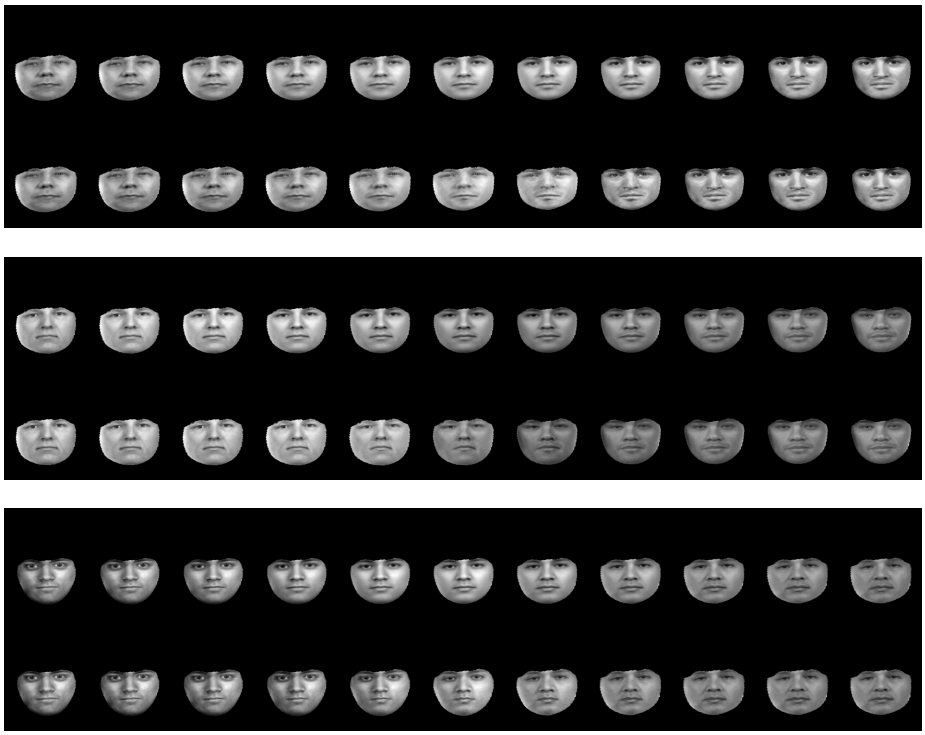

Figure 7: Interpolating between a randomly chosen face (left column) and a second face (right column) nearly on the opposite side of the hyperellipse of coefficients. Top row of each image: linear interpolation of coefficients. The middle images lack distinctiveness. Bottom row of each image: interpolating "around the hyperellipse". Detail is preserved throughout the interpolation. Please enlarge to see details.

> The observations collected above lead to the remarkable conclusion that algorithms such as MAP may be nonsensical in high dimensions! This conclusion is not widely known in the computer vision and graphics community, where MAP is commonly used for face computations with models having 10-100 dimensions.[1] However, our conclusion is supported in [8], where Mackay states "probability density maxima often have very little associated probability mass even though the value of the probability density there may be immense, because they have so little associated volume... the locations of probability density maxima in many dimensions are generally misleading and irrelevant. Probability densities should only be maximized if there is good reason to believe that the location of the maximum conveys useful information about the whole distribution."

[1] In fact many results in statistics focus on the case where increasing amounts of data are available, i.e. $n/d \to \infty$ with n the number of data points. In our problem we may have n/d finite and small, as in the case of a face model with several hundred training examples, each with 100 degrees of freedom.

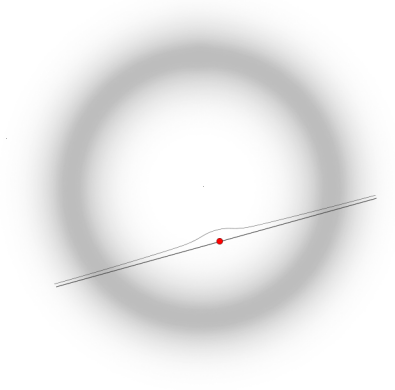

Figure 8: In this schematic illustration the point along the constraint (dark line) that has the highest probability is the red point. In high dimensions however, the interior of the Gaussian is empty and the probability mass is concentrated toward the outside.

5 Example Computation: Interpolating in Face Space

Figure 7 contrasts two approaches to interpolating facial identity. The images are not photographs but are synthesized with an AAM [11]. The face on the far left is generated from a coefficient vector $\mathbf{c}_l$ sampled from a multivariate Gaussian with the appropriate variances (eigenvalues). The face on the far right is also randomly chosen, but its coefficient vector $\mathbf{c}_r$ is modified to constrain it to having a specified inner product $\langle \mathbf{c}_l, \mathbf{c}_r \rangle_{\mathbf{\Lambda}^{-1}} = -0.8$ so as to place it on the opposite side of the coefficient volume. The inner product uses the inverse eigenvalue-weighted norm $\langle \mathbf{c}_l, \mathbf{c}_r \rangle_{\mathbf{\Lambda}^{-1}} = \mathbf{c}_l^T \mathbf{\Lambda}^{-1} \mathbf{c}_r$. The dimensionality of the space (length of the coefficient vector) is 181.

The top rows in figure 7 shows linear interpolation through the Gaussian coefficient space. The midpoint of this interpolation passes closer to the center (mean) face than either end. This results in a somewhat "ghostly" face that lacks detail. The linear interpolation also has the undesired result that (for example) interpolating from a person of age 40 to a person of age 45 might pass through an intermediate face of apparent age 25, if that is the mean age of the database underlying the AAM.

In the lower panels of figure 7 we interpolate "around" a hyperellipsoidal shell in the coefficient space rather than across the volume. Given initial and final coefficient vectors $\mathbf{c}_l, \mathbf{c}_r$, at each step a coefficient vector is generated that interpolates the norm of these vectors (although in fact the difference in norm is expected to be small due to phenomena mentioned above). This interpolation remains inside the high probability shell of the hyperGaussian and generates distinctive faces throughout the interpolation.

6 Conclusion

This paper describes known high-dimensional phenomena that call into question common assumptions underlying much computer vision, graphics, and psychological research on face computation. In particular, we question approaches that assume that typical faces lie in the interior of a high-dimensional Gaussian density (Figure 8). The issue is not due to a bi- or multimodal distribution (as with a combined distribution containing both women and men) but rather is a consequence of high dimensionality. These objections are not merely hypothetical, but are visible in a simple face computation. Our conclusion highlights the need to develop new algorithms that address the intrinsically high-dimensional nature of facial identity and expression.

Appendix: Hyperellipsoidal Angle Calculation

The interpolations in Figure 7 start with a randomly chosen coefficient vector $\mathbf{y}$ with $y_i \sim N(0, \sqrt{\lambda_i})$. This produces the first face. For the second face, we select a coefficient vector $\mathbf{x}$ that has a specified Mahalanobis inner product with that of the first face, $\mathbf{x}\mathbf{\Lambda}^{-1}\mathbf{y} = c$ with $c = -0.8$ for example. To find $\mathbf{x}$ we solve a sequence of problems

$$\mathbf{x} \leftarrow \arg\min_{\mathbf{x}} \quad (\mathbf{x}-\mathbf{r})^T\mathbf{\Lambda}^{-1}(\mathbf{x}-\mathbf{r}) \quad + \quad \lambda(\mathbf{x}^T\mathbf{\Lambda}^{-1}\mathbf{y} - c)$$

$$\mathbf{r} \leftarrow \frac{\mathbf{x}}{\mathbf{x}^T\mathbf{\Lambda}^{-1}\mathbf{x}}$$

with $\mathbf{r}$ initialized to a random vector, in other words, find the vector that is closest to $\mathbf{r}$ and has the desired Mahalanobis angle with $\mathbf{y}$.

Acknowledgements

This research is partially supported by the Japan Science and Technology Agency, CREST project. JPL acknowledges helpful discussions with Marcus Frean and Alex Ma.

References

[1] Anjyo, K., Todo, H., Lewis, J.: A practical approach to direct manipulation blendshapes. J. Graphics Tools **16**(3), 160–176 (2012)

[2] Blanz, T., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proceedings of ACM SIGGRAPH, pp. 187–194. ACM SIGGRAPH (1999)

[3] Bracewell, R.: The Fourier Transform and Its Applications. McGraw Hill (2000)

[4] Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active Appearance Models, *Lecture Notes in Computer Science*, vol. 1407. Springer (1998)

[5] Henne, M., Hickel, H.: The making of "toy story". In: Proceedings of the 41st IEEE International Computer Conference, COMPCON '96, pp. 463–. IEEE Computer Society, Washington, DC, USA (1996)

[6] Lewis, J., Anjyo, K.: Direct manipulation blendshapes. Computer Graphics and Applications (special issue: Digital Human Faces) **30**(4), 42–50 (2010)

[7] Li, H., Yu, J., Ye, Y., Bregler, C.: Realtime facial animation with on-the-fly correctives. ACM Transactions on Graphics **32**(4), 42:1–42:10 (2013)

[8] MacKay, D.J.: Hyperparameters: Optimize, or integrate out? In: Maximum entropy and Bayesian methods, pp. 43–59. Springer (1996)

[9] Matthews, I., Xiao, J., Baker, S.: On the dimensionality of deformable face models. CMU-RI-TR-06-12 (2006)

[10] Meytlis, M., Sirovich, L.: On the dimensionality of face space. IEEE Trans. Pattern Anal. Mach. Intell. **29**(7), 1262–1267 (2007)

[11] Mo, Z., Lewis, J., Neumann, U.: Face inpainting with local linear representations. In: BMVC, pp. 347–356. BMVA (2004)

[12] Patel, A., Smith, W.A.P.: 3D morphable face models revisited. In: Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1327–1334. IEEE Computer Society, Los Alamitos, CA, USA (2009)

[13] Penev, P.S., Sirovich, L.: The global dimensionality of face space. In: Proc. 4th Int'l Conf. Automatic Face and. Gesture Recognition, pp. 264–270 (2000)

[14] Phillips, P.J., Wechsler, H., Huang, J., Rauss, P.: The feret database and evaluation procedure for face recognition algorithms. Image and Vision Computing J. **16**(5), 295–306 (1998)

[15] Seo, J., Irving, G., Lewis, J.P., Noh, J.: Compression and direct manipulation of complex blendshape models. ACM Trans. Graph. **30**(6), 164:1–164:10 (2011)

[16] Valentine, T.: Face-space models of face recognition. In: M. Wenger, J. Townsend (eds.) Computational, Geometric, and Process Perspectives on Facial Cognition: Contexts and Challenges, Scientific Psychology Series. Taylor & Francis (2012)

[17] Vlasic, D., Brand, M., Pfister, H., Popović, J.: Face transfer with multilinear models. In: ACM Transactions on Graphics (TOG), vol. 24, pp. 426–433. ACM Press, New York, NY, USA (2005)

[18] Wang, J.: Geometric Structure of High-Dimensional Data and Dimensionality Reduction. Springer (2011)