2016

Information Geometric Probability Models in Statistical Signal Processing

Russell Costa
University of Rhode Island, costar@ele.uri.edu

Follow this and additional works at: https://digitalcommons.uri.edu/oa_diss

Recommended Citation

Costa, Russell, "Information Geometric Probability Models in Statistical Signal Processing" (2016). *Open Access Dissertations*. Paper 502.
https://digitalcommons.uri.edu/oa_diss/502

This Dissertation is brought to you for free and open access by DigitalCommons@URI. It has been accepted for inclusion in Open Access Dissertations by an authorized administrator of DigitalCommons@URI. For more information, please contact digitalcommons-group@uri.edu.

INFORMATION GEOMETRIC PROBABILITY MODELS IN STATISTICAL
SIGNAL PROCESSING

BY
RUSSELL COSTA

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN
ELECTRICAL ENGINEERING

UNIVERSITY OF RHODE ISLAND

2016

DOCTOR OF PHILOSOPHY DISSERTATION
OF
RUSSELL COSTA

APPROVED:

Dissertation Committee:

Major Professor Steven Kay

Peter Swaszek

Thomas Bella

Nasser H. Zawia

DEAN OF THE GRADUATE SCHOOL

UNIVERSITY OF RHODE ISLAND

2016

ABSTRACT

The exponentially embedded family (EEF) of probability density functions (PDFs) is an important modeling paradigm in a number of signal processing applications including model selection and hypothesis testing. In recent years, the availability of inexpensive sensors that can gather information over wide, distributed areas, has led to many interesting challenges in distributed estimation and data fusion. In these applications there is a need for combining information extracted from data which is only available locally at individual sensors. The EEF is valuable in this context due to its unique properties with respect to the Kullback-Leibler divergence measure. We study the geometry of the EEF in the context of distributed applications where we must combine information at the PDF level. Central to this is the work of Chernoff, in particular, the notion and / or existence of an equidistant PDF between component PDFs or hypotheses. This equidistance is defined with respect to the Kullback Leibler divergence, and has been utilized in recent years as a criterion for combining information (PDFs) in data fusion applications. Much of that work, however, has been limited to combining PDFs in a pairwise fashion. To expand on this we utilize results from Information Geometry in order to establish a formal extension of the “Chernoff point” (equidistant PDF) for the EEF of more than two components. Using well known theory of exponential families, we establish the proper existence criteria of such an extension, and show that much of the geometry involved can be well understood by drawing analogies between the EEF of three component PDFs, and the Euclidean geometry of triangles. We have shown that finding the extended Chernoff point can be guided by considering the orientation of the component PDFs, and that the use of this paradigm can lead to better ways to combine estimates in classical problems such as combining estimates of common means from separate Normal populations.

ACKNOWLEDGMENTS

First of all, I would like to thank my advisor Dr. Steven Kay, for his guidance throughout my studies here at the University of Rhode Island. Over the years I have benefited from his knowledge, experience, and attention to detail which I hope will exhibit itself in my work for many years to come. He is an excellent teacher who has the very rare gift of solving very complex problems, while presenting the results in well constructed, easy to comprehend examples. I am a better student and researcher thanks to him.

I would also like to thank the faculty of the Department of Electrical, Computer, and Biomedical Engineering and the Department of Mathematics. In particular the members of my Graduate Committee for contributing their valuable time and asking very insightful and thought provoking questions. Such interactions will help steer the continuation of my research beyond the dissertation.

I would like to posthumously thank Dr. Lewis Pakula who was taken from us way to soon. Dr. Pakula was an incredible teacher with a magnetic personality. I am very fortunate to have known him and to have learned from such a fine human being.

I am also grateful to my many colleagues and friends whom have helped me throughout the years with their encouragement and technical acumen. In particular, Dr. Tom Wettergren, Dr. Marcus Graham, and Mr. John Baylog of the Naval Undersea Warfare Center (NUWC) whom have always made time for reading my proposals and papers, listening to my ideas, making great comments, and asking insightful questions.

I also owe a debt to my sponsors and managers throughout the years who have supported me both financially and administratively. From the In-house Laboratory Independent Research program at NUWC, Dr. Tony Ruffa and Mr. Neil

Dubois. Special thanks to Neil for being extremely accommodating and supportive in providing the flexibility required to finish the degree requirements. From the management side Mr. Jack O'Sullivan and Mr. Steve O'Grady have been very supportive of my work and have allowed me the freedom to balance my work on numerous Navy projects with the pursuit of my academic research.

Finally I would like to thank my family for their love and support. I am particularly thankful to my wife and best friend Leslie Sherman Costa, who has been so patient throughout this process, despite having to bear a larger burden at home. To my two beautiful daughters Olivia and Abigail for making my life such a wonderful, loving, adventure. To my mother Jacqueline Costa who raised two children on her own and will forever be my hero. To my sister Christine whose independent spirit has fueled her to overcome many obstacles. Finally to my niece Melissa whose perseverance, and uncommon zest for life has taught me to live each day and not let little things get to me.

PREFACE

This thesis is made up of three main manuscripts. In this preface we summarize each manuscript and discuss additional unpublished research that expands beyond each manuscript. These additional results are added as Appendices to the manuscripts.

The main contribution in this dissertation is a novel triangle characterization of the EEF which could be useful in many applications. This contribution extends the work of Chernoff in laying out the details required to find the PDF that is equidistant from component PDFs (hypotheses, models, etc) in a PDF space (statistical manifold). In particular, in distributed signal processing applications such as information fusion, distributed estimation, and distributed detection; algorithms which can account for the orientation of models and / or hypotheses could outperform standard approaches. In this dissertation we show how this characterization can have effects in a previously developed model selection approach. In addition, we show that the EEF can be utilized to introduce a new class of estimators for combining common mean estimates from separate Gaussian populations. Finally, we feel that the use of Euclidean analogies in this Information Geometrical approach should make this burgeoning field of study more accessible to the general statistical signal processing community.

In Chapter 1 we advance the understanding of the EEF Model selection approach developed by Kay, with the main focus on the geometrical aspects of the EEF in a Hypothesis testing framework. We introduce the concept of the information center as an equidistant point in the model space (with respect to Kullback Leibler divergence) and as an alternative origin to the typical null hypothesis. We prove that by embedding each model in turn with the I-center, instead of the “no signal” null hypothesis, we could achieve the optimality of the well known

maximum likelihood or minimum distance rule. In Appendix 1C we add some additional results which help the reader develop a more complete understanding of the geometric interpretation of the EEF. In particular, we numerically generate a visualization of the decision regions in a simple two dimensional example. Further, we add a proof that if the EEF procedure using the I-center as origin, is modified to remove the lower bound constraint, then simply choosing the maximum embedding parameter is equivalent to the ML rule. In Appendix 1D we document some additional work that we did in extending the EEF procedure to an M-ary form.

In Chapter 2 we extend the equidistant property of the Chernoff point for EEF of M component probability density functions. As a result we uncover some interesting higher order geometrical features which relate Chernoff's work to the EEF of more than two components. In particular, we define a novel triangle characterization of the EEF of three components and we state and prove a theorem showing that the Chernoff distance acts as a square distance globally with respect to this triangle characterization. We work out numerous examples for Gaussian components and use these to illustrate these properties. These results could be important to the data fusion community where similar paradigms are prevalent but not fully understood. In Appendices 2C and 2D we include some additional pertinent results. Specifically in Appendix 2C we derive the geodesic equations for univariate Gaussian PDF and show that the pairwise EEF is a solution of these second order partial differential equations. While in Appendix 2D we share some numerical results which illustrate, using a univariate Gaussian example, that the pairwise Chernoff distance acts as a squared distance globally, and that a square root transformation results in a function that meets all criteria for a metric in this PDF space.

In Chapter 3 we apply the EEF paradigm to combining estimates of a com-

mon mean from separate normal populations with unknown and possibly different variance. This is a classical problem and we propose a new class of estimators through the formation of the EEF of the PDFs of the sample mean of each independent population. This family of estimators is narrowed down to two main choices; the Centroid or “center of mass” of the sample mean PDFs, and the Chernoff PDF, which is equi-distant to the PDFs of the estimated sample means. We showed that the Centroid is exactly the classical estimator introduced by Graybill and Deal. In addition, we derive a closed form solution for the Chernoff PDF and show through Monte Carlo simulations that this estimator performs better than the classical estimator for small sample size and variance ratios less than three. In Appendix 3A we look at two examples of applying the EEF paradigm to combining estimates. The first is in the very difficult problem of combining PDF estimates assumed estimated from the same source of data, but with possibly redundant data. In this example we combine two univariate Gaussian PDF estimates and show that the Chernoff appears as a lower bound to the Centroid with respect to Kullback Leibler divergence in the presence of redundant data. The Centroid in the equal sample size problem is the best fixed choice of embedding parameter. In the second example we simplify the PDF example to one of combining variance estimates. That is we adjust the Gaussian PDF combination to one in which we have a common known mean. This gives us a simpler expression for the Chernoff embedding parameter which allows us to prove analytically that the Chernoff parameter converges in probability to $\frac{1}{2}$, i.e. converges to the Centroid. Note that this is a general result that is true in the case of dependent data as well as independent. Further we numerically determine the PDFs of the embedding parameter when the sources are independent and dependent. Interestingly in the independent case the embedding parameter is Gaussian for sample sizes greater than three, so

that the embedding function transforms an $F(m, m)$ distribution to a Gaussian. In the dependent case the PDF is non-Gaussian but with a sharp peak at $\frac{1}{2}$. Thus for large number of samples or higher dependence the Chernoff PDF is very closely approximated by the Centroid. This implies that the Centroid is a better choice in many applications since in general it requires very little computation.

In Chapter 4 we outline some future directions of this research.

TABLE OF CONTENTS

ABSTRACT	ii
ACKNOWLEDGMENTS	iii
PREFACE	v
TABLE OF CONTENTS	ix
LIST OF FIGURES	xii
CHAPTER	
1 A geometrical interpretation of exponentially embedded families of Gaussian probability density functions for model selection	1
Abstract	2
1.1 Introduction	2
1.2 Alternative choice of origin: The Information Center of Component Models	5
1.2.1 Monte Carlo Results	8
1.3 Geometric Interpretation of Results	11
1.4 Conclusion	17
Appendix 1A - Derivation of minimum distance I-center	18
Appendix 1B - Proof of (21)	19
Appendix 1C - Numerical Decision Regions and general proof	20
Appendix 1D - Extension to M-ary EEf	22
List of References	32

	Page
2 An extension of the equidistant property of the Chernoff point in exponentially embedded families of probability density functions	34
Abstract	35
2.1 Introduction	35
2.2 Exponentially Embedded Families and the Extended Chernoff Point	40
2.2.1 Extension of the binary mid-point equivalence property .	42
2.2.2 An interesting property of the Cumulative Generating Function and an extension of Chernoff distance . . .	43
2.3 Geometrical definitions in Euclidean and PDF spaces	44
2.4 EEf of Gaussian PDFs	51
2.4.1 Orthogonal bisectors of Gaussian PDFs	51
2.4.2 EEf for M Gaussians	52
2.4.3 Examples: EEf of two Gaussian components	53
2.5 Characterization of Three Component EEf	55
2.5.1 Numerical examples for finding extended Chernoff point .	69
2.6 Conclusion	72
Appendix 2A - Proof of General EEf of Gaussians	75
Appendix 2B - Projection onto the orthogonal bisector of a pairwise connection	77
Appendix 2C: Properties of the EEf of Univariate Gaussians	78
Appendix 2D: Numerical Results for Chernoff Metric	81
List of References	83
3 On Combining Estimates using the Exponentially Embedded Family Paradigm	87

	Page
3.1 Introduction	88
3.2 Combining Common means	90
3.3 Exponentially Embedded Families	91
3.4 Numerical Results	94
3.5 Conclusion	98
Appendix 3A - Additional Results in using the EEF for Combining Estimates	99
List of References	113
4 Future Work	116
List of References	118

LIST OF FIGURES

Figure		Page
1	Numerical results for $N = 100$: Alternative Signal Models with I-center	9
2	Numerical results for $N = 100$: Probability of choosing the correct model for the EEF, EEF with I-center (mn = min-norm, md = min-distance), and ML approaches.	9
3	The improvement in P_c , the probability of correct model choice, for the I-center of minimum distance (KL-divergence) over the EEF with null origin.	10
4	The improvement in P_c , the probability of correct model choice, for the I-center of minimum norm over the EEF with null origin.	11
5	The difference in P_c , the probability of correct model choice, for the I-center of minimum norm from the optimal ML test.	12
6	The difference in P_c , the probability of correct model choice, for the I-center of minimum distance (KL-divergence) from the optimal ML test.	13
7	Geometrical interpretation of EEF selection procedure for bivariate Gaussian component models with equal covariance $\sigma^2 \mathbf{I}$, (a) null hypothesis as origin, (b) the I-center as origin.	14
8	Decision regions of the Optimal ML test for $\mathbf{s}_1$, $\mathbf{s}_2$, and $\mathbf{s}_3$ (asterisks) in the two-dimensional example.	21
9	Decision regions of the EEF test with null origin for $\mathbf{s}_1$, $\mathbf{s}_2$, and $\mathbf{s}_3$ (asterisks) in the two-dimensional example (see Figure 7(a)).	22
10	Decision regions of the EEF test with I-center origin for $\mathbf{s}_1$, $\mathbf{s}_2$, and $\mathbf{s}_3$ (asterisks) in the two-dimensional example (see Figure 7(b)).	23
11	Two-dimensional decision regions where the model points make up an obtuse triangle: ML test	24
12	Two-dimensional decision regions where the model points make up an obtuse triangle: EEF _{IC}	24

Figure		Page
13	Orthogonal bisection of the geodesic (line) connecting two points in Euclidean space with $d(\cdot, \cdot)$ the Euclidean distance. Note that all points on the orthogonal bisector are circumcenters of circumcircles through the two points. The smallest such circumcircle has a center point which lies on the connecting line.	45
14	Orthogonal bisection of the geodesic (line) connecting two univariate normal PDFs drawn in PDF parameter space with $D(\cdot\ \cdot)$ the KL divergence.	47
15	The depiction of the unique circumcenter, O, of the circumscribed circle of three points in Euclidean space. The circumcenter point exists as the intersection of the orthogonal bisectors of each side of the triangle (i.e. for points in general position). For an acute triangle (left) the circumcenter falls within the triangle. For a right triangle (center), the circumcenter falls on the mid-point of the longest side. For an obtuse triangle (right), the circumcenter falls outside the triangle. In the first two cases the circumcircle coincides with the smallest enclosing circle. For an obtuse triangle they do not coincide as the smallest enclosing circle is centered as in the right triangle, on the mid-point of the longest sided (hypotenuse).	48
16	<i>Geometry of PDFs.</i> An “acute” connection of three univariate Gaussian PDFs. The lines connecting the PDF points (asterisks) are <i>geodesics</i> through the exponential connection in Gaussian parameter space. The orthogonal bisection of each <i>geodesic</i> is shown with the matching color / linetype. The triangles denote the pairwise Chernoff points on the corresponding <i>geodesic</i> connections of each PDF pair. This “triangle” of Gaussians is defined as “acute” since the intersection of the orthogonal bisectors occurs within the region of Gaussian parameter space corresponding to positive embedding weights, $\eta_i > 0$ for all i , i.e. the extended Chernoff point.	58

Figure		Page
17	<i>Geometry of PDFs.</i> A “right” connection of three univariate Gaussian PDFs. The lines connecting the PDF points (asterisks) are <i>geodesics</i> through the exponential connection in Gaussian parameter space. The orthogonal bisection of each <i>geodesic</i> is shown with the matching color / linetype. The triangles denote the pairwise Chernoff points on the <i>geodesic</i> connections of each PDF pair. This “triangle” of Gaussians is defined as “right” due to the intersection of the bisectors (extended Chernoff point denoted by circle) occurring at the pairwise Chernoff point $p_{\eta_{02}^*}$ of the “longest” side, i.e. $D_{02}^* > D_{01}^*, D_{12}^*$	60
18	<i>Geometry of PDFs.</i> An “obtuse” connection of three univariate Gaussian PDFs. The lines connecting the PDF points (asterisks) are <i>geodesics</i> through the exponential connection in Gaussian parameter space. The orthogonal bisection of each <i>geodesic</i> is shown with the matching color, with the triangles denoting the pairwise Chernoff points on the <i>geodesics</i> . This “triangle” of Gaussians is defined as “obtuse” due to the intersection of the bisectors (extended Chernoff point denoted by circle) falling outside the region of Gaussian parameter space corresponding to positive embedding weights. In this case $\boldsymbol{\eta}^*$ has one negative component, $\eta_1 < 0$	61
19	Pythagorean relation of information geometry applied in proof to lemma 1.	63
20	Sketch supporting the necessary part of the proof of theorem 4. . .	66
21	Sketch supporting the sufficient part of the proof of theorem 4. . .	68
22	The Kullback-Liebler divergence relative to the extended Chernoff point / SEB solution point (PDF) for the “acute” connection of three univariate Gaussian PDFs. The black lines connecting the PDF points are <i>geodesics</i> through the exponential connection, and thus make up the boundary of the exponential connection in the normal parameter space. The colors depict the level of the KL divergence of all points (PDFs) relative to the center point (magenta asterisk). The color level is cropped at $D(p_{\boldsymbol{\eta}^*} \ p_i)$ which is drawn as a contour on the image.	70

Figure		Page
23	The Kullback-Liebler divergence relative to the extended Chernoff point / SEB solution point (PDF) for the “right” connection of three univariate Gaussian PDFs. The black lines connecting the PDF points are <i>geodesics</i> through the exponential connection, and thus make up the boundary of the exponential connection in the normal parameter space. The colors depict the level of the KL divergence of all points (PDFs) relative to the center point (magenta asterisk). The color level is cropped at $D(p_{\boldsymbol{\eta}^*} \ p_i)$ which is drawn as a contour on the image.	72
24	The Kullback-Liebler divergence relative to the SEB solution point. With respect to SEB, the “radius” of the ball, $r_s^* = 0.5$	73
25	The Kullback-Liebler divergence relative to the extended Chernoff solution point. The “radius” of the ball is $r^* = 0.541$ with respect to the extended Chernoff point.	74
26	The Chernoff distance of each “side”, i.e. pairwise connection, of the three component EEFs as a function of the mean parameter μ_2 of p_2 . Note that pair p_0, p_2 has the largest Chernoff distance and thus plays the role of hypotenuse (largest squared length).	81
27	The difference of the Chernoff distance of the hypotenuse and the sum of the Chernoff distance of the two “shorter sides”. Note that the triangle properties determined are consistent with the determinations made from the I-center.	82
28	This figure illustrates that the square root of the Chernoff distance satisfies the triangle inequality and thus is a metric in Univariate Gaussian space.	83
29	Performance of Common Mean Estimators as a function of m for $\rho = 1$, $\rho = 2$, and $\rho = 3$	95
30	Performance of Common Mean Estimators $m = 5$	96
31	Performance of Common Mean Estimators $m = 8$	97
32	Kernel estimates for PDF of $\hat{\eta}_*$ for increasing ρ	98
33	Performance of all fixed parameter combined mean estimators vs. Chernoff for fixed $m = 5$ and increasing ρ	99

Figure	Page
34 Double Counting example. Note that the central partition contains data samples common to both sets S_0 and S_1	101
35 Monte Carlo Results of Pairwise Chernoff Fusion under double counting.	102
36 Kullback Leibler divergence of Chernoff fusion PDF relative to Centroid: No overlap	103
37 PDF of Chernoff embedding parameter (no overlap): $m = 2$	107
38 PDF of Chernoff embedding parameter (no overlap): $m = 5$	108
39 First two cumulants of $g(\rho)$	109
40 Third and Fourth cumulants of $g(\rho)$	109
41 Simulation vs Integration for the density of ρ given common samples.	110
42 PDF of $g(\rho)$ with overlap; $m = 8, t = 4$	111

CHAPTER 1

A geometrical interpretation of exponentially embedded families of Gaussian probability density functions for model selection

by

Russell Costa and Steven Kay

Dept. of Electrical, Computer and Biomedical Engineering

University of Rhode Island, Kingston, RI, USA

published in IEEE Transactions on Signal Processing, vol. 61, no. 1, pp. 62-67,

Jan.1, 2013.

Abstract

Model selection via exponentially embedded families (EEF) of probability models has been shown to perform well on many practical problems of interest. A key component in utilizing this approach is the definition of a model origin (i.e. null hypothesis) which is embedded individually within each competing model. In this correspondence we give a geometrical interpretation of the EEF and study the sensitivity of the EEF approach to the choice of model origin in a Gaussian hypothesis testing framework. We introduce the Information Center (I-center) of competing models as an origin in this procedure and compare this to using the standard null hypothesis. Finally we derive optimality conditions for which the EEF using I-center achieves optimal performance in the Gaussian hypothesis testing framework.

1.1 Introduction

Model order estimation is a well studied topic in statistics. Many classical approaches exist for selecting among models, in order to attain the most efficient extraction of information. In recent years the use of exponentially embedded families (EEF) of probability density functions (PDFs) have been shown to yield good results in these types of problems. In [1], the EEF model selection approach was shown to improve on the original form of Minimum Description Length (MDL) [2] in choosing model order when the component models consisted of nested sinusoids. In [3] the EEF approach was introduced as an alternative to the generalized likelihood ratio test (GLRT) in problems with specified alternative signal models of uncertain length. In these limited applications of this approach the choice of an origin (null hypothesis) was assumed as a noise-only PDF. That is, any signal component was assumed to be zero. In this paper we investigate the logic of this assumption of null hypothesis, and study the sensitivity to the chosen origin of the

EEF methodology in a Gaussian hypothesis testing framework.

In the EEF approach, given M models we construct M families of PDFs as

$$p_i(\mathbf{x}; \eta_i) = \exp(\eta_i \ln \frac{p_i(\mathbf{x})}{p_0(\mathbf{x})} - K_0^i(\eta_i) + \ln p_0(\mathbf{x})) \quad i = 1, 2, \dots, M \quad (1)$$

where $K_0^i(\eta_i)$ is the Cumulative Generating Function (or log normalization factor) given as

$$K_0^i(\eta_i) = \ln \int p_i(\mathbf{x})^{\eta_i} p_0(\mathbf{x})^{1-\eta_i} d\mathbf{x} \quad (2)$$

in which $p_0(\mathbf{x})$ is a reference PDF, and η_i is the embedding parameter. This function is associated with the i^{th} model and is valid for all real values of η_i for which (2) is less than infinity. In the EEF procedure the embedding parameter, η_i , is restricted to be in the interval $[0, 1]$ (as in [1]) so that each family of densities are bounded by the PDF pair being embedded. Note that $p_0(\mathbf{x})$ plays the role of origin in the PDF space and the density $p_i(\mathbf{x}; \eta_i)$ is in the exponential family independent of the parametric forms of the component PDFs (provided the PDFs have the same support). This embedding also has the property that if all component PDFs are from the same exponential family, the resultant EEF is within the given family [4].

It is well established [5], that the PDF family in (1) is a *geodesic* (i.e. minimum distance) connection in the PDF space. This *geodesic* is defined with respect to Kullback-Leibler divergence. Letting $T_i(\mathbf{x}) = \ln \frac{p_i(\mathbf{x})}{p_0(\mathbf{x})}$ (the sufficient statistic for η_i), the EEF model selection approach is to choose the model $k \in [1, M]$ if $\max_{\eta_i} [\eta_i T_i(\mathbf{x}) - K_0^i(\eta_i)]$ is maximum for $i = k$. This is equivalent to

$$\arg \max_i \left[\hat{\eta}_i \ln \frac{p_i(\mathbf{x}; \mathbf{s}_i)}{p_0(\mathbf{x})} - K_0(\hat{\eta}_i) \right] \quad (3)$$

where $\hat{\eta}_i$ is the MLE under the i^{th} model. If we assume known Gaussian models such that $p_0(\mathbf{x}) := N(\mathbf{0}, \sigma^2 \mathbf{I})$ and $p_i(\mathbf{x}) := N(\mathbf{s}_i, \sigma^2 \mathbf{I})$ with $\mathbf{x}, \mathbf{s}_i$ as $N \times 1$ vectors and $\mathbf{I}$ an $N \times N$ identity matrix, then $K_0^i(\eta_i) = -\frac{(1-\eta_i)\eta_i}{2\sigma^2} \mathbf{s}_i^T \mathbf{s}_i$ and the MLE of the embedding parameter under the i^{th} model is given as $\hat{\eta}_i = \frac{\mathbf{s}_i^T \mathbf{x}}{\mathbf{s}_i^T \mathbf{s}_i}$. Applying

bound constraints on the MLE, i.e. $\hat{\eta}_i \in [0, 1]$, and by substitution into (3) the EEF selection (EEF₀ for the $\mathbf{0}$ reference / origin) becomes

$$\text{EEF}_0(i) = \begin{cases} \frac{[\mathbf{s}_i^T \mathbf{x}]^2}{2\sigma^2 \mathbf{s}_i^T \mathbf{s}_i} & 0 < \frac{\mathbf{s}_i^T \mathbf{x}}{\mathbf{s}_i^T \mathbf{s}_i} < 1 \\ \frac{\mathbf{s}_i^T \mathbf{x} - \frac{1}{2} \mathbf{s}_i^T \mathbf{s}_i}{\sigma^2} & \frac{\mathbf{s}_i^T \mathbf{x}}{\mathbf{s}_i^T \mathbf{s}_i} \geq 1 \\ 0 & \frac{\mathbf{s}_i^T \mathbf{x}}{\mathbf{s}_i^T \mathbf{s}_i} \leq 0 \end{cases}. \quad (4)$$

The optimal detector for this problem is well known from classical statistics [6] as

$$\arg \max_i \mathbf{x}^T \mathbf{s}_i - \frac{1}{2} \mathbf{s}_i^T \mathbf{s}_i \quad (5)$$

commonly known as the *minimum distance receiver* or *replica correlator*. Note that when the MLE of the embedding parameter is greater than or equal to one, the EEF test statistic is equivalent to the minimum distance receiver.

The EEF₀ allows one to choose among the models by first forming *geodesic* connections between each hypothesis (in turn) and the origin. Then the data is projected onto each *geodesic* through the MLE inference of η_i . This view of projection is supported by a well established duality between maximum likelihood estimation and Kullback-Liebler divergence dating back to Wald [7]. To understand this dual relationship, consider the ML rule as in (5). The equivalent selection with respect to Kullback-Liebler divergence is written as $\min_i \{D(p(\mathbf{x}; \boldsymbol{\mu}, \sigma^2 \mathbf{I}) \| p(\mathbf{x}; \boldsymbol{\mu}_i, \sigma^2 \mathbf{I}))\}_{\boldsymbol{\mu} \rightarrow \mathbf{x}}$ where $D(p_1(\mathbf{x}) \| p_0(\mathbf{x})) = \int_{-\infty}^{\infty} p_1(\mathbf{x}) \ln \frac{p_1(\mathbf{x})}{p_0(\mathbf{x})} d\mathbf{x}$. Note that $\mathbf{x} = \hat{\boldsymbol{\mu}}$, i.e. $\mathbf{x}$ is the MLE of $\boldsymbol{\mu}$. Thus the maximum likelihood selection problem in (5) is equivalent to

$$\min_i \{D(p(\mathbf{x}; \boldsymbol{\mu}, \sigma^2 \mathbf{I}) \| p(\mathbf{x}; \boldsymbol{\mu}_i, \sigma^2 \mathbf{I}))\}_{\boldsymbol{\mu} \rightarrow \hat{\boldsymbol{\mu}}(\mathbf{x})}. \quad (6)$$

The sequence of events in this dual mapping (assuming normal PDFs with known variance) is

1. Find the KL-divergence between the “true” PDF $p(\mathbf{x}; \boldsymbol{\mu}, \sigma^2 \mathbf{I})$ (where arbitrary vector $\boldsymbol{\mu}$ maps out the signal space), and hypothesized model PDFs. For normal PDFs this has a closed form solution as $D(N(\boldsymbol{\mu}, \sigma^2 \mathbf{I}) \| N(\boldsymbol{\mu}_i, \sigma^2 \mathbf{I})) = \frac{\|\boldsymbol{\mu} - \boldsymbol{\mu}_i\|^2}{2\sigma^2}$.
2. Since the KL-divergence is a function of $\boldsymbol{\mu}$, then by the invariance property of the MLE we obtain the MLE of the KL-divergence by replacing $\boldsymbol{\mu}$ with $\hat{\boldsymbol{\mu}}$ such that $\hat{D}(\boldsymbol{\mu}) = D(\hat{\boldsymbol{\mu}}(\mathbf{x}))$.
3. Finally take the minimum over the models $i = 1, \dots, M$ as in (6).

Extending this to the aforementioned geodesic projections we let $\boldsymbol{\mu}_i = \eta_i \mathbf{s}_i$ so that the dual of (4) from (6) is

$$\min_i \min_{\eta_i} \{D(p(\mathbf{x}; \boldsymbol{\mu}, \sigma^2 \mathbf{I}) \| p(\mathbf{x}; \eta_i \mathbf{s}_i, \sigma^2 \mathbf{I}))\}_{\boldsymbol{\mu} \rightarrow \hat{\boldsymbol{\mu}}(\mathbf{x})} \quad (7)$$

which for normal PDFs yields $\hat{\eta}_i = \min_{\eta_i} \frac{\|\mathbf{x} - \eta_i \mathbf{s}_i\|^2}{2\sigma^2}$. Thus choosing the model whose connection is “closest” (with respect to Kullback-Leibler divergence) to the data. This is equivalent to choosing the model for which $p(\mathbf{x}; \hat{\eta}_i)$ is maximum.

1.2 Alternative choice of origin: The Information Center of Component Models

In the interest of studying the dependence and/or sensitivity of the EEF model selection criteria to the choice of origin we introduce the I-center of the given models as an alternative origin. We define the I-center as the PDF $p_{IC}(\mathbf{x})$, equidistant from the models with respect to Kullback-Leibler divergence. That is for M models, the I-center is the PDF p_{IC} , which satisfies $D(p_{IC} \| p_1) = D(p_{IC} \| p_2) = \dots = D(p_{IC} \| p_M)$. To solve for the I-center of PDFs, we adopt a single (arbitrary) model as reference (p_1) and form $M - 1$ equations as

$$D(p_{IC} \| p_i) - D(p_{IC} \| p_1) = 0 \quad i = 2, 3, \dots, M \quad (8)$$

where each equation represents the necessary conditions for p_{IC} to be on the orthogonal bisector (with respect to Kullback-Liebler divergence) of each PDF pairing with p_1 . Thus solving for the I-center (the $M - 1$ simultaneous equations in (8)) entails finding the intersection of all orthogonal bisectors of the pairwise connections. Letting $p_i(\mathbf{x}) := \mathcal{N}(\mathbf{s}_i, \mathbf{C}_i)$, $i = 2, \dots, M$, we define the orthogonal bisector of a given pair of PDFs as the subset of PDFs $p \in \mathcal{P}$ such that

$$\begin{aligned} \Delta^i(\mathcal{P}) &= D(p(\mathbf{x}; \mathbf{s}, \mathbf{C}) \| p_1(\mathbf{x}; \mathbf{s}_1, \mathbf{C}_1)) - D(p(\mathbf{x}; \mathbf{s}, \mathbf{C}) \| p_i(\mathbf{x}; \mathbf{s}_i, \mathbf{C}_i)) \\ &= \ln \frac{|\mathbf{C}_i|}{|\mathbf{C}_1|} + \text{tr}\{\mathbf{C}(\mathbf{C}_i^{-1} - \mathbf{C}_1^{-1})\} + \\ &\quad (\mathbf{s} - \mathbf{s}_i)^T \mathbf{C}_i^{-1} (\mathbf{s} - \mathbf{s}_i) - (\mathbf{s} - \mathbf{s}_1)^T \mathbf{C}_1^{-1} (\mathbf{s} - \mathbf{s}_1) = 0 \end{aligned} \quad (9)$$

where $\mathcal{P}$ is the set of all Gaussian PDFs. Thus the orthogonal bisector of pair p_1 and p_i is the family of PDFs corresponding to the parameters $(\mathbf{s}, \mathbf{C})$ (i.e. $p(\mathbf{x}; \mathbf{s}, \mathbf{C})$) which satisfy (9). This as in Euclidean space, is an infinite set of points (PDFs). If we assume that $\mathbf{C}_i = \sigma^2 \mathbf{I}$, the analytical form of the orthogonal bisectors simplifies to

$$\Delta^i(\mathcal{P}) = \frac{(\mathbf{s}_0 - \mathbf{s}_i)^T (\mathbf{s}_0 - \mathbf{s}_i) - (\mathbf{s}_0 - \mathbf{s}_1)^T (\mathbf{s}_0 - \mathbf{s}_1)}{\sigma^2} = 0 \quad (10)$$

with $\mathbf{s}_0$ as the defining point (or set of points) in $\mathbb{R}^N$ of the I-center. Thus we want to find $\mathbf{s}_0$, the point at which the orthogonal bisectors (lines) in the N-dimensional parameter space intersect. Note that the Kullback Leibler divergence in this case is reduced to squared Euclidean distance [4]. Finally (from (10)) to find the I-center we solve for $\mathbf{s}_0$ in

$$\mathbf{A} \mathbf{s}_0 = \mathbf{b} \quad (11)$$

where

$$\mathbf{A} = \begin{bmatrix} (\mathbf{s}_2 - \mathbf{s}_1)^T \\ \vdots \\ (\mathbf{s}_M - \mathbf{s}_1)^T \end{bmatrix}$$

and

$$\mathbf{b} = \begin{bmatrix} \frac{\mathbf{s}_2^T \mathbf{s}_2 - \mathbf{s}_1^T \mathbf{s}_1}{2} \\ \vdots \\ \frac{\mathbf{s}_M^T \mathbf{s}_M - \mathbf{s}_1^T \mathbf{s}_1}{2} \end{bmatrix}$$

Note that for $N > M - 1$, (11) is an under-determined system of equations with an infinite number of solutions (signals). A unique solution can be found by choosing the solution with minimum ℓ_2 norm. This solution is $\mathbf{s}_0\{\text{min-norm}\} = \mathbf{A}^\dagger \mathbf{b} = \mathbf{A}^T (\mathbf{A} \mathbf{A}^T)^{-1} \mathbf{b}$ where $\mathbf{A}^\dagger$ is the pseudo-inverse of matrix $\mathbf{A}$ [8]. However, an infinite number of solutions can be constructed by adding this minimum norm solution to other vectors in the null space of $\mathbf{A}$, raising an additional question of whether a particular choice of I-center is better than another. For this reason, we also consider the solution of

$$\min_{\mathbf{s}_0 \in \mathcal{S}} (\mathbf{s}_0 - \mathbf{s}_i)^T (\mathbf{s}_0 - \mathbf{s}_i) \quad \text{for any } i \in [1, M] \quad (12)$$

where $\mathcal{S} = \{\mathbf{s}_0 \in \mathbb{R}^N : \|\mathbf{A} \mathbf{s}_0 - \mathbf{b}\|_2 = 0\}$ which is the I-center of minimum distance from the component models, with (arbitrary) $\mathbf{s}_i$ as the reference. By the method of Lagrange Multipliers, it is easy to show that (see appendix 1A)

$$\mathbf{s}_0\{\text{min-dist}\} = [\mathbf{I} - \mathbf{A}^T (\mathbf{A} \mathbf{A}^T)^{-1} \mathbf{A}] \mathbf{s}_i + \mathbf{s}_0\{\text{min-norm}\} \quad i = 1, \dots, M. \quad (13)$$

Thus the PDF corresponding to the I-center in this example is $p_{IC}(\mathbf{x}) := N(\mathbf{s}_0, \sigma^2 \mathbf{I})$, where we will consider both the minimum norm and minimum distance solutions. If we choose this PDF as the origin in the EEF procedure described above, the MLE of the embedding parameters are written as (for $i = 1, \dots, M$) $\hat{\eta}_i = \frac{(\mathbf{s}_i - \mathbf{s}_0)^T (\mathbf{x} - \mathbf{s}_0)}{(\mathbf{s}_i - \mathbf{s}_0)^T (\mathbf{s}_i - \mathbf{s}_0)}$ with $\hat{\eta}_i \in [0, 1]$. Then the EEF with I-Center as origin becomes

$$\text{EEF}_{IC}(i) = \begin{cases} \frac{[(\mathbf{s}_i - \mathbf{s}_0)^T (\mathbf{x} - \mathbf{s}_0)]^2}{2\sigma^2 (\mathbf{s}_i - \mathbf{s}_0)^T (\mathbf{s}_i - \mathbf{s}_0)} & 0 < \frac{(\mathbf{s}_i - \mathbf{s}_0)^T (\mathbf{x} - \mathbf{s}_0)}{(\mathbf{s}_i - \mathbf{s}_0)^T (\mathbf{s}_i - \mathbf{s}_0)} < 1 \\ \frac{(\mathbf{s}_i - \mathbf{s}_0)^T \mathbf{x} - (\mathbf{s}_i^T \mathbf{s}_i - \mathbf{s}_0^T \mathbf{s}_0)/2}{\sigma^2} & \frac{(\mathbf{s}_i - \mathbf{s}_0)^T (\mathbf{x} - \mathbf{s}_0)}{(\mathbf{s}_i - \mathbf{s}_0)^T (\mathbf{s}_i - \mathbf{s}_0)} \geq 1 \\ 0 & \frac{(\mathbf{s}_i - \mathbf{s}_0)^T (\mathbf{x} - \mathbf{s}_0)}{(\mathbf{s}_i - \mathbf{s}_0)^T (\mathbf{s}_i - \mathbf{s}_0)} \leq 0 \end{cases} \quad (14)$$

As in the original EEF derivation, for $\hat{\eta}_i \geq 1$ the EEF_{IC} test is identical to the ML rule. Note that an I-center exists in this reduced Gaussian case as long as the embedded signal models (i.e. log likelihood relative to origin) are linearly independent. In more general probability models the existence would require the sufficient statistics of each model connection to be linearly independent. That is for $i = 2, \dots, M$, $\ln \frac{p_i(\mathbf{x})}{p_1(\mathbf{x})}$ should be linearly independent (see [9] section 1.6 for details).

1.2.1 Monte Carlo Results

To quantify the performance of the EEF under alternative origins we run a series of numerical simulations in which we calculate the probability of choosing the correct model, P_c , given $M = 3$ models defined as $N(\mathbf{s}_i, \sigma^2 \mathbf{I})$ with $\mathbf{s}_i = \sum_{j=1}^i \cos(2\pi f_j n)$, with $f_1 = 0.1$, $f_2 = 0.11$, $f_3 = 0.12$, and $n = 0, 1, \dots, N - 1$. This model was taken from [1]. We calculate P_c via Monte Carlo simulation for varying number of data points, N , spanning from 40 to 120, and noise variance spanning from 1 to 100. For each noise variance 15,000 realizations were performed and used for calculating the criteria in (4), (5), and (14), respectively. The simulation was performed with equal prior probabilities for each model being the true model.

Figure 1 shows for $N = 100$, the three signal models with I-center, while Figure 2, the probability of choosing the correct model as a function of the inverse noise variance for the EEF_0 ($\mathbf{0}$ hypothesis as origin), EEF_{mn} and EEF_{md} (I-center as origin with minimum norm and minimum distance, respectively), and the optimal ML test. We observe that by choosing the I-center as origin (in this example) the EEF for both I-centers, attains the optimal performance of the ML test for moderate to high signal to noise ratio (SNR), while EEF_0 is sub-optimal for all SNR. However, for low SNR the EEF with the I-center does not consistently meet

Figure 1: Numerical results for $N = 100$: Alternative Signal Models with I-center

Figure 2: Numerical results for $N = 100$: Probability of choosing the correct model for the EEF, EEF with I-center (mn = min-norm, md = min-distance), and ML approaches.

Figure 3: The improvement in P_c , the probability of correct model choice, for the I-center of minimum distance (KL-divergence) over the EEF with null origin.

the optimal performance of the ML test. Figures 3 - 6 show the complete simulation results for all data lengths and noise variance. Specifically, Figures 3 and 4 show the gain in performance of the EEF individually for the two distinct I-centers over that of EEF_0 , measured as the change in probability of correct model choice P_c . These plots show that the EEF with I-center consistently performs better than EEF_0 ; with the alternative I-centers performing the same except at the lowest SNRs. At the lowest SNRs the EEF using the minimum distance I-center attains the best performance. This difference in performance at low SNR is most evident in Figures 5 and 6 where we show the difference in performance of the two EEFs with I-center and the optimal ML test. These plots also show that the EEF with minimum distance I-center attains the optimal performance over all SNRs for signal durations of 52 to 100 samples while the minimum norm I-center performs consistently less than optimal at low SNR.

Figure 4: The improvement in P_c , the probability of correct model choice, for the I-center of minimum norm over the EEF with null origin.

1.3 Geometric Interpretation of Results

The above results show that choosing an origin which is an I-center of the component models allows the EEF procedure to match the optimal ML rule over a wide span of conditions. In addition, we observe that choosing some other origin (in this case the noise only hypothesis) degrades the performance in EEF model selection. To visualize the effect of changing origin we consider another three component example, but for $N = 2$. This reduction in dimension allows us to visually interpret the procedure as we are operating in a signal plane. Let $\mathbf{s}_1 = [1 \ 1]^T$, $\mathbf{s}_2 = [2 \ 1]^T$, and $\mathbf{s}_3 = [1.5 \ 3]^T$ with equal variance $\sigma^2 \mathbf{I}$. The I-center in this case is unique (since $N = M - 1$) as the *circumcenter* of the *circumcircle* that intersects points $\mathbf{s}_1$, $\mathbf{s}_2$, and $\mathbf{s}_3$ in $\mathbb{R}^2$. Note that since the points in the plane associated with the component models make up an acute triangle it is well known from plane geometry that the circumcenter is within the triangle (with the model/signal points as vertices), and is found as the unique intersection of the orthogonal bisectors of each side of the triangle. In this case the I-center point

Figure 5: The difference in P_c , the probability of correct model choice, for the I-center of minimum norm from the optimal ML test.

is $\mathbf{s}_0 = [1.5 \ 1.9375]$. In Figure 7 we show the geometry of the sub-manifolds of models for the EEF with null origin (a), and I-center origin (b), respectively. The model points and the origins are depicted as asterisks. The solid lines depict the *connections* (family of PDF parameters) of each model with the origin. Consider a data point $\mathbf{x} = [1.46 \ 1.07]$. Then the classical test would choose the model closest to $\mathbf{x}$ as $\mathbf{s}_1$. In contrast, the EEF selects the model corresponding to the point on the connection (in this case line segment) that is closest to $\mathbf{x}$ (see previous section). In this case the correct choice is $\mathbf{s}_1$. The arrows depict the projection of $\mathbf{x}$ to each model connection in the EEF procedure using both origins. This shows us that in this case EEF_0 (Figure 7(a)) chooses the wrong model $\mathbf{s}_2$, due to the fact that the connection of $\mathbf{s}_2$ and the origin is closer to the data $\mathbf{x}$, than the other connections, despite $\mathbf{x}$ being closer to point $\mathbf{s}_1$ than point $\mathbf{s}_2$. While EEF_{IC} (Figure 7(b)) results in the proper connection (model) choice. In general, this example shows that the decision regions associated with the EEF procedure depend explicitly on the choice of origin and that choosing the I-center as origin avoids asymmetry in the model

Figure 6: The difference in P_c , the probability of correct model choice, for the I-center of minimum distance (KL-divergence) from the optimal ML test.

connections that can occur by an arbitrarily defined origin.

An additional issue for geometrical interpretation is the dependence on signal length of the results. Since changing signal length changes each signal model and potentially their orientation in $\mathbb{R}^N$, the I-center is a function of signal length. Thus the I-center must be calculated separately for each value of N . Motivated by the 2-D geometry we generalize the orientation of models through the location of the I-center. That is we define the orientation of models (connection) as obtuse if the I-center (minimum distance) is outside the subset of $\mathbb{R}^N$ mapped out by the convex combination of the component models, while if it is within this region then the connection is acute [10]. A simple test for determining these conditions (without calculating the I-center) in $\mathbb{R}^N$ was developed in [11] where the application was in finding well-centered triangulations. Since we require the I-center, it is more instructive to rewrite the minimum distance I-center problem in (12) with additional constraints on $\mathbf{s}_0$ such that $\mathbf{s}_0 = \sum_{i=1}^M \alpha_i \mathbf{s}_i$, with $\sum_{i=1}^M \alpha_i = 1$ for the M component

Figure 7: Geometrical interpretation of EEF selection procedure for bivariate Gaussian component models with equal covariance $\sigma^2 \mathbf{I}$, (a) null hypothesis as origin, (b) the I-center as origin.

models. The solution (again found from method of Lagrange multipliers) is [12]

$$\boldsymbol{\alpha} = \mathbf{L}^{-1} \mathbf{d} \quad (15)$$

where $\mathbf{L}$ is an $M \times M$ matrix

$$\mathbf{L} = \begin{bmatrix} 1 & \dots & 1 \\ \mathbf{s}_j^T (\mathbf{s}_i - \mathbf{s}_1) \end{bmatrix} \quad (16)$$

for $i = 2, \dots, M$, $j = 1, \dots, M$, and

$$\mathbf{d} = \begin{bmatrix} 1 \\ \frac{1}{2} (\mathbf{s}_2^T \mathbf{s}_2 - \mathbf{s}_1^T \mathbf{s}_1) \\ \vdots \\ \frac{1}{2} (\mathbf{s}_M^T \mathbf{s}_M - \mathbf{s}_1^T \mathbf{s}_1) \end{bmatrix} \quad (17)$$

Note that $\mathbf{L}$ is of full rank as long as $\mathbf{s}_i - \mathbf{s}_j$ are linearly independent for all i, j . Using (15) yields the minimum distance I-center as well the type of connection, in that if $\alpha_i \geq 0$ for all i , then the connection is acute. Otherwise the connection is obtuse. Thus equations (15)-(17) yield the identical minimum distance I-center as in (12), but also tells us the type of connection by observing the weights.

Revisiting the results of the previous section a couple of things become clear. The cases where the EEF with minimum distance I-center matched optimal performance (for all SNR) corresponded to acute connections. The remaining cases, where performance was sub-optimal for low SNR, corresponded to obtuse connections. Note that although in the obtuse connections the EEF with I-center is sub-optimal, it improves more significantly over the null origin in these cases. This shows that in these connections the choice of origin is more crucial. In contrast, the minimum norm I-center did not attain optimal performance at low SNR for any N thus the minimum distance I-center is the superior choice of origin. The following theorems give conditions for the optimality of the EEF with I-center.

Theorem 1. *Let the data $\mathbf{x}$ be such that the ML rule selects the i^{th} model. That is $\mathbf{x}$ is in the decision region of model $\mathbf{s}_i$, depicted as $\mathbf{x} \in \mathcal{D}(\mathbf{s}_i)$. If $\hat{\eta}_i > 0$, then the EEF_{IC} test selects the same model as the ML test.*

Proof. Let the squared Euclidean distance from the component points $\mathbf{s}_i$ to the I-center $\mathbf{s}_0$ be $(\mathbf{s}_i - \mathbf{s}_0)^T(\mathbf{s}_i - \mathbf{s}_0) = r^2 > 0$. Now from (14)

$$EEF_{IC}(i) = \begin{cases} \frac{[(\mathbf{s}_i - \mathbf{s}_0)^T(\mathbf{x} - \mathbf{s}_0)]^2}{2\sigma^2 r^2} & 0 < (\mathbf{s}_i - \mathbf{s}_0)^T(\mathbf{x} - \mathbf{s}_0) < r^2 \\ \frac{(\mathbf{s}_i - \mathbf{s}_0)^T \mathbf{x} - (\mathbf{s}_i^T \mathbf{s}_i - \mathbf{s}_0^T \mathbf{s}_0)/2}{\sigma^2} & (\mathbf{s}_i - \mathbf{s}_0)^T(\mathbf{x} - \mathbf{s}_0) \geq r^2 \\ 0 & (\mathbf{s}_i - \mathbf{s}_0)^T(\mathbf{x} - \mathbf{s}_0) \leq 0 \end{cases} \quad (18)$$

but

$$(\mathbf{s}_i - \mathbf{s}_0)^T \mathbf{x} - (\mathbf{s}_i^T \mathbf{s}_i - \mathbf{s}_0^T \mathbf{s}_0)/2 = (\mathbf{s}_i - \mathbf{s}_0)^T(\mathbf{x} - \mathbf{s}_0) - \frac{r^2}{2}$$

thus

$$EEF_{IC}(i) = \begin{cases} \frac{[(\mathbf{s}_i - \mathbf{s}_0)^T(\mathbf{x} - \mathbf{s}_0)]^2}{2r^2} < \frac{r^2}{2} & 0 < (\mathbf{s}_i - \mathbf{s}_0)^T(\mathbf{x} - \mathbf{s}_0) < r^2 \\ (\mathbf{s}_i - \mathbf{s}_0)^T(\mathbf{x} - \mathbf{s}_0) - \frac{r^2}{2} \geq \frac{r^2}{2} & (\mathbf{s}_i - \mathbf{s}_0)^T(\mathbf{x} - \mathbf{s}_0) \geq r^2 \\ 0 & (\mathbf{s}_i - \mathbf{s}_0)^T(\mathbf{x} - \mathbf{s}_0) \leq 0 \end{cases} \quad (19)$$

where we have removed σ^2 because it is common over all models. Now with no loss in generality let $\mathbf{s}_0 = \mathbf{0}$, then $\mathbf{s}_i^T \mathbf{s}_i = r^2 = \mathcal{E}$. Note that in this case all signals have the same energy $\mathcal{E}$. Finally the EEF_{IC} becomes

$$\text{EEF}_{IC}(i) = \begin{cases} \frac{(\mathbf{s}_i^T \mathbf{x})^2}{2\mathcal{E}} < \frac{\mathcal{E}}{2} & 0 < \mathbf{s}_i^T \mathbf{x} < \mathcal{E} \\ \mathbf{s}_i^T \mathbf{x} - \frac{\mathcal{E}}{2} \geq \frac{\mathcal{E}}{2} & \mathbf{s}_i^T \mathbf{x} \geq \mathcal{E} \\ 0 & \mathbf{s}_i^T \mathbf{x} \leq 0 \end{cases} \quad (20)$$

For $\mathbf{s}_0 = \mathbf{0}$, $\hat{\eta}_i = \frac{\mathbf{s}_i^T \mathbf{x}}{\mathcal{E}}$ and for $\hat{\eta}_i > 0$ the EEF is strictly monotonic with respect to $\mathbf{s}_i^T \mathbf{x}$ and thus the theorem holds as seen by referring to (4), i.e. ML rule with $\mathbf{s}_i^T \mathbf{s}_i = \mathcal{E}$ for all i . $\square$

In the Monte Carlo simulations it became evident that when the i^{th} model was correct, but $\hat{\eta}_i < 0$ then ties could occur, particularly for obtuse connections. The above theorem shows that in the absence of ties, any I-center can attain optimal performance. This is evident from the simulation results at the moderate to high signal to noise ratios. The following theorem, in addition to the theorem above, describes the conditions for which optimality is achieved independent of signal to noise ratio.

Theorem 2. *Take $\mathbf{s}_0$ to be a minimum distance I-center for (linearly independent) models $\mathbf{s}_1, \dots, \mathbf{s}_M$. If the models make an acute connection, then for any $\mathbf{x}$ in the decision region of model $\mathbf{s}_i$ (i.e. $\mathbf{x} \in \mathcal{D}(\mathbf{s}_i)$), the embedding parameter associated with model $\mathbf{s}_i$, $\hat{\eta}_i$, is greater than zero.*

Proof. Assume $\mathbf{x} \in \mathcal{D}(\mathbf{s}_i)$ and that $\mathbf{s}_0$ is an I-center of the given models. Then by (5) and the equidistant property of $\mathbf{s}_0$ (see appendix B)

$$\mathbf{x} \in \mathcal{D}(\mathbf{s}_i) \Rightarrow \forall_{j \neq i} (\mathbf{x} - \mathbf{s}_0)^T (\mathbf{s}_i - \mathbf{s}_j) > 0 \quad (21)$$

Further we take $\mathbf{s}_0$ as the minimum distance I-center of an acute connection, thus $\mathbf{s}_0 = \sum_{j=1}^M \alpha_j \mathbf{s}_j$ where $\sum_{j=1}^M \alpha_j = 1$ for $\alpha_j > 0$. Thus writing $\mathbf{s}_i = \sum_{j=1}^M \alpha_j \mathbf{s}_j$ we

have

$$\begin{aligned}
(\mathbf{x} - \mathbf{s}_0)^T (\mathbf{s}_i - \mathbf{s}_0) &= (\mathbf{x} - \mathbf{s}_0)^T \left[\sum_{j=1}^M \alpha_j (\mathbf{s}_i - \mathbf{s}_j) \right] \\
&= \sum_{j=1}^M \alpha_j (\mathbf{x} - \mathbf{s}_0)^T (\mathbf{s}_i - \mathbf{s}_j)
\end{aligned} \tag{22}$$

Now since the connection is acute, then $\alpha_j > 0$ for all j and from (21) we have

$$\sum_{j=1}^M \alpha_j (\mathbf{x} - \mathbf{s}_0)^T (\mathbf{s}_i - \mathbf{s}_j) > 0 \Rightarrow (\mathbf{x} - \mathbf{s}_0)^T (\mathbf{s}_i - \sum_{j=1}^M \alpha_j \mathbf{s}_j) = (\mathbf{x} - \mathbf{s}_0)^T (\mathbf{s}_i - \mathbf{s}_0) = r^2 \hat{\eta}_i > 0 \tag{23}$$

where r^2 is the squared Euclidean distance of $\mathbf{s}_i$ from $\mathbf{s}_0$ (greater than zero). Thus $\hat{\eta}_i > 0$. $\square$

Theorems 1 and 2 show that for an acute connection (model orientation), the EEF with the minimum distance I-center is optimal independent of SNR.

1.4 Conclusion

In this paper we present a geometric interpretation of EEF model selection which shows a direct dependence of model selection performance on the choice of model origin. Furthermore, we introduced an alternative to the null hypothesis (as model origin) based on the information center (I-center) of given models. We analyzed the use of the I-center in EEF model selection and compared performance to the typical choice of null hypothesis as origin. The analytical and numerical results show that using the minimum distance I-center as an origin in EEF model selection (with known models) is equivalent to the optimal maximum likelihood selection rule when the models make up an acute connection. In general, regardless of the orientation of component models, the minimum distance I-center was the best choice of origin in that it outperformed the other choices explored.

Appendix 1A - Derivation of minimum distance I-center

To prove (13) from (12) we form the Lagrangian function (for arbitrary i)

$$\mathcal{L}(\mathbf{s}_0, \boldsymbol{\lambda}) = (\mathbf{s}_0 - \mathbf{s}_i)^T(\mathbf{s}_0 - \mathbf{s}_i) + \boldsymbol{\lambda}^T(\mathbf{A}\mathbf{s}_0 - \mathbf{b}) \quad (24)$$

now taking the gradients with respect to $\mathbf{s}_0$ and $\boldsymbol{\lambda}$ we have

$$\nabla_{\mathbf{s}_0} = 2\mathbf{s}_0 - 2\mathbf{s}_i + \mathbf{A}^T \boldsymbol{\lambda} \quad (25)$$

$$\nabla_{\boldsymbol{\lambda}} = \mathbf{A}\mathbf{s}_0 - \mathbf{b}. \quad (26)$$

Setting the gradients to zero and solving for $\boldsymbol{\lambda}$ we have $\boldsymbol{\lambda} = -2(\mathbf{A}\mathbf{A}^T)^{-1}(\mathbf{b} - \mathbf{A}\mathbf{s}_i)$ for the Lagrange multiplier. Substitution into (24) after simplification yields (13).

Appendix 1B - Proof of (21)

We must prove that if $\mathbf{s}_0$ is an I-center for models $\mathbf{s}_i$, $i = 1, 2, \dots, M$, and $\mathbf{x} \in \mathcal{D}(\mathbf{s}_i)$, then for any $j \neq i$

$$(\mathbf{x} - \mathbf{s}_0)^T(\mathbf{s}_i - \mathbf{s}_j) > 0 \quad (27)$$

With no loss in generality let $i = 1$, $j = 2$. Then $\mathbf{x} \in \mathcal{D}(\mathbf{s}_i)$ implies

$$2\mathbf{x}^T \mathbf{s}_1 - \mathbf{s}_1^T \mathbf{s}_1 > 2\mathbf{x}^T \mathbf{s}_2 - \mathbf{s}_2^T \mathbf{s}_2 \quad (28)$$

Note that we can write $\mathbf{s}_1^T \mathbf{s}_1$ in terms of arbitrary vector $\mathbf{s}_0$ as

$$\mathbf{s}_1^T \mathbf{s}_1 = (\mathbf{s}_1 - \mathbf{s}_0)^T(\mathbf{s}_1 - \mathbf{s}_0) + (\mathbf{s}_1 - \mathbf{s}_0)^T \mathbf{s}_0 + \mathbf{s}_0^T \mathbf{s}_1. \quad (29)$$

Similarly,

$$\mathbf{s}_2^T \mathbf{s}_2 = (\mathbf{s}_2 - \mathbf{s}_0)^T(\mathbf{s}_2 - \mathbf{s}_0) + (\mathbf{s}_2 - \mathbf{s}_0)^T \mathbf{s}_0 + \mathbf{s}_0^T \mathbf{s}_2. \quad (30)$$

But if $\mathbf{s}_0$ is an I-center, then $(\mathbf{s}_1 - \mathbf{s}_0)^T(\mathbf{s}_1 - \mathbf{s}_0) = (\mathbf{s}_2 - \mathbf{s}_0)^T(\mathbf{s}_2 - \mathbf{s}_0) = r^2$. Using this (29), and (30) in (28) yields

$$\begin{aligned} 2\mathbf{x}^T \mathbf{s}_1 - r^2 - (\mathbf{s}_1 - \mathbf{s}_0)^T \mathbf{s}_0 - \mathbf{s}_0^T \mathbf{s}_1 &> 2\mathbf{x}^T \mathbf{s}_2 - r^2 - (\mathbf{s}_2 - \mathbf{s}_0)^T \mathbf{s}_0 - \mathbf{s}_0^T \mathbf{s}_2 \\ 2\mathbf{x}^T \mathbf{s}_1 - 2\mathbf{s}_0^T \mathbf{s}_1 &> 2\mathbf{x}^T \mathbf{s}_2 - 2\mathbf{s}_0^T \mathbf{s}_2 \end{aligned}$$

which is equivalent to (27).

Appendix 1C - Numerical Decision Regions and general proof

Expanding on the results in section 1.3, Figures 8, 9, and 10 show the numerically determined decision regions of each test where the value (model index) of the model selected relative to each point in the space is shown. These figures show that EEF_{IC} has identical decision region as the ML test, while EEF_0 suffers from the asymmetry caused by arbitrary choice of origin. However, EEF_{IC} does not always attain the identical decision region as ML as this next example shows. Let $\mathbf{s}_1 = [1 \ 1]$, $\mathbf{s}_2 = [2 \ 1]$, and $\mathbf{s}_3 = [3 \ 3]$. Note that unlike the previous example, these model (non co-linear) points make an obtuse triangle in the plane. Thus from plane geometry the circumcenter is known to exist outside the triangle made by the model points. Figures 11 and 12 show the decision regions for the ML test and EEF_{IC} , respectively, with the decision regions for EEF_{IC} modified slightly by mapping cases in which the EEF is 0 (i.e. when ties occur), for all of the models to zero. In the normal implementation of the EEF this region would be associated with the first listed model which is typically the least complex. This shows that when the model points are oriented as an obtuse triangle there is a portion of the decision space (observation space) that may map to this tie condition.

The occurrence of ties in the obtuse case is a consequence of the bounds on $\hat{\eta}_i$. In particular, the lower bound of zero. If we re-derive the test with I-center as origin, and allow for negative weights, then we can show that the EEF_{IC} is equivalent to the ML test. Indeed, consider the ML test. For $\mathbf{x}$ in the decision region of model $\mathbf{s}_i$ we have

$$\mathbf{x}^T \mathbf{s}_i - \frac{1}{2} \mathbf{s}_i^T \mathbf{s}_i > \mathbf{x}^T \mathbf{s}_j - \frac{1}{2} \mathbf{s}_i^T \mathbf{s}_j$$

for $i \neq j$. With no loss in generality, consider the boundary of $\mathbf{s}_1$ and $\mathbf{s}_2$ and let $\mathbf{x}$ be in the decision region of $\mathbf{s}_1$, then

$$\mathbf{x}^T \mathbf{s}_1 - \frac{1}{2} \mathbf{s}_1^T \mathbf{s}_1 > \mathbf{x}^T \mathbf{s}_2 - \frac{1}{2} \mathbf{s}_2^T \mathbf{s}_2 \quad (31)$$

Figure 8: Decision regions of the Optimal ML test for $\mathbf{s}_1$, $\mathbf{s}_2$, and $\mathbf{s}_3$ (asterisks) in the two-dimensional example.

Further, let $\mathbf{s}_0$ be an I-center for M models, including $\mathbf{s}_1$ and $\mathbf{s}_2$. Then

$$r^2 = (\mathbf{s}_1 - \mathbf{s}_0)^T (\mathbf{s}_1 - \mathbf{s}_0) = (\mathbf{s}_2 - \mathbf{s}_0)^T (\mathbf{s}_2 - \mathbf{s}_0) = \dots = (\mathbf{s}_M - \mathbf{s}_0)^T (\mathbf{s}_M - \mathbf{s}_0) \quad (32)$$

Using (31) in (32), it is easy to show that

$$(\mathbf{x} - \mathbf{s}_0)^T \mathbf{s}_1 > (\mathbf{x} - \mathbf{s}_0)^T \mathbf{s}_2 \quad (33)$$

where this must be true as a consequence of $\mathbf{x}$ in decision region of $\mathbf{s}_1$ and because $\mathbf{s}_0$ is an I-center. Recall that the embedding parameter (MLE) of the i^{th} model, relative to the I-center is

$$\hat{\eta}_i = \frac{(\mathbf{x} - \mathbf{s}_0)^T (\mathbf{s}_i - \mathbf{s}_0)}{(\mathbf{s}_i - \mathbf{s}_0)^T (\mathbf{s}_i - \mathbf{s}_0)} = \frac{(\mathbf{x} - \mathbf{s}_0)^T (\mathbf{s}_i - \mathbf{s}_0)}{r^2} \quad (34)$$

Now subtract $(\mathbf{x} - \mathbf{s}_0)^T \mathbf{s}_0$ from both sides of (32) and divide by r^2 , we have

$$\hat{\eta}_1 > \hat{\eta}_2 \quad (35)$$

Note that these MLEs are unconstrained. Thus when using the I-center as origin, simply choosing the maximum embedding parameter is equivalent to the ML rule.

Figure 9: Decision regions of the EEF test with null origin for $\mathbf{s}_1$, $\mathbf{s}_2$, and $\mathbf{s}_3$ (asterisks) in the two-dimensional example (see Figure 7(a)).

Appendix 1D - Extension to M-ary EEF

Following Manuscript 1 an M-ary version of the EEF was studied where instead of embedding hypotheses in turn (pairwise) with a reference hypothesis, the component hypotheses were embedded directly together. That is we form the PDF

$$p(\mathbf{x}; \boldsymbol{\eta}) = \exp \left(\sum_{i=1}^M \eta_i \ln \frac{p_i(\mathbf{x})}{p_0(\mathbf{x})} - \mathcal{K}_0(\boldsymbol{\eta}) + \ln p_0(\mathbf{x}) \right) \quad (36)$$

where the PDFs $p_i(\mathbf{x})$ are known, where

$$\mathcal{H}_i : \mathbf{x} \sim p_i(\mathbf{x}) \quad (37)$$

Ding and Kay [13] proved that the ML rule for choosing among $\mathcal{H}_i$ is equivalent to choosing the hypothesis for which

$$\arg \min_i D(p(\mathbf{x}; \boldsymbol{\eta}) \| p_i(\mathbf{x})) \quad (38)$$

with $\boldsymbol{\eta} = \hat{\boldsymbol{\eta}}$, the MLE. Note that $\boldsymbol{\eta}$ can take values such that $\mathcal{K}_0(\boldsymbol{\eta})$ is finite, and

$$\hat{\boldsymbol{\eta}} = \arg \max_{\boldsymbol{\eta}} p(\mathbf{x}; \boldsymbol{\eta}) \quad (39)$$

Figure 10: Decision regions of the EEF test with I-center origin for $\mathbf{s}_1$, $\mathbf{s}_2$, and $\mathbf{s}_3$ (asterisks) in the two-dimensional example (see Figure 7(b)).

A simple example can be solved for linearly independent signals in white Gaussian noise. In this case one can solve analytically for the MLE of the embedding parameters.

1D.1 Closed form MLE for EEF of Gaussians with identical covariance

Consider the EEF of M Gaussian components

$$p_{\boldsymbol{\eta}}(\mathbf{x}) = \frac{p_0^{\eta_0}(\mathbf{x})p_1^{\eta_1}(\mathbf{x}) \cdots p_{M-1}^{\eta_{M-1}}(\mathbf{x})}{\int_{\mathbb{R}} p_0^{\eta_0}(\mathbf{x})p_1^{\eta_1}(\mathbf{x}) \cdots p_{M-1}^{\eta_{M-1}}(\mathbf{x})d\mathbf{x}} \quad (40)$$

where component i is $\mathcal{N}(\mathbf{s}_i, \sigma^2 \mathbf{I})$ for $i = 0, 1, \dots, M-1$ and the mean vectors $\mathbf{s}_i$ are linearly independent. Also the dimension of $\mathbf{x}$ (the data), N , is greater than the number of components (i.e. $N > M$). It is easy to show that (40) is also Gaussian with mean $\mathbf{s}(\boldsymbol{\eta})$ and variance $\sigma^2 \mathbf{I}$, where

$$\mathbf{s}(\boldsymbol{\eta}) = \sum_{i=0}^{M-1} \eta_i \mathbf{s}_i \quad (41)$$

and we have enforced the sum constraint $\sum_{i=0}^{M-1} \eta_i = 1$. We want to find the MLE of $\boldsymbol{\eta}$ that meets the sum constraint. To do this in a simple matrix form we

Figure 11: Two-dimensional decision regions where the model points make up an obtuse triangle: ML test

Figure 12: Two-dimensional decision regions where the model points make up an obtuse triangle: EEF_{IC}

utilize some facts about the MLE, and the Kullback Liebler divergence. First, it is well known that maximizing a likelihood function given data is equivalent to minimizing the Kullback Liebler divergence from the empirical distribution (i.e. a simple mapping of the data in the parameter space). Second, for Gaussian components with identical covariance structure, Kullback Liebler divergence is equivalent to a squared Euclidean distance. Thus to find the MLE of $\boldsymbol{\eta}$ in (40) we equivalently solve

$$\arg \min_{\boldsymbol{\eta}} (\mathbf{x} - \mathbf{s}(\boldsymbol{\eta}))^T (\mathbf{x} - \mathbf{s}(\boldsymbol{\eta})) \quad (42)$$

subject to the aforementioned sum constraint. Imposing the sum constraint we can write $\mathbf{s}(\boldsymbol{\eta}) = \mathbf{s}_0 + \sum_{i=1}^{M-1} \eta_i (\mathbf{s}_i - \mathbf{s}_0)$ or in matrix form

$$\mathbf{s}(\boldsymbol{\eta}) = \mathbf{s}_0 + \mathbf{A}\boldsymbol{\eta}_r \quad (43)$$

where $\mathbf{A} = [\mathbf{s}_1 - \mathbf{s}_0 \quad \mathbf{s}_2 - \mathbf{s}_0 \dots \mathbf{s}_{M-1} - \mathbf{s}_0]$ and $\boldsymbol{\eta}_r = [\eta_1 \eta_2 \dots \eta_{M-1}]^T$. After substitution of (43) into (42), taking the derivative with respect to $\boldsymbol{\eta}_r$ and setting equal to zero, we have

$$\boldsymbol{\eta}_r = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T (\mathbf{x} - \mathbf{s}_0) \quad (44)$$

$$\eta_0 = 1 - \boldsymbol{\eta}_r^T \mathbf{1}_{M-1} \quad (45)$$

where $\mathbf{1}_{M-1} = [\mathbf{1} \dots \mathbf{1}]^T$.

1D.1.1 Equivalence of ML rule

We want to show that a two step procedure which consists of (1) Find the MLE of $\boldsymbol{\eta}$ by maximizing the M component EEF (2) Choose the model i closest (with respect to KL-divergence) to the EEF evaluated at the MLE of $\boldsymbol{\eta}$, i.e. $\arg \min_i D(p(\mathbf{x}; \hat{\boldsymbol{\eta}}) \| p_i(\mathbf{x}))$ is equivalent to the well known ML rule. Note in this case the KL-divergence is equivalent to Euclidean distance since the covariance matrices are identical and we enforce the sum constraint.

With no loss in generality take $M = 3$, and $\mathbf{s}_0 = \mathbf{0}$. Then

$$\hat{\boldsymbol{\eta}}_r = [\hat{\eta}_1 \ \hat{\eta}_2]^T = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{x} \quad (46)$$

where $\mathbf{A} = [\mathbf{s}_1 \ \mathbf{s}_2]$. Let $\rho = \mathbf{s}_1^T \mathbf{s}_2$, $\epsilon_1 = \mathbf{s}_1^T \mathbf{s}_1$, $\epsilon_2 = \mathbf{s}_2^T \mathbf{s}_2$ and define

$$ds_i = (\mathbf{s}_i - \mathbf{s}(\hat{\boldsymbol{\eta}}))^T (\mathbf{s}_i - \mathbf{s}(\hat{\boldsymbol{\eta}})) \quad (47)$$

where $\mathbf{s}(\hat{\boldsymbol{\eta}}) = \mathbf{A} \hat{\boldsymbol{\eta}}_r = \mathbf{A}(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{x}$. Thus the EEF based rule will choose model i corresponding to the minimum ds_i . Expanding (47) we have

$$ds_i = \epsilon_i + \mathbf{s}(\hat{\boldsymbol{\eta}})^T \mathbf{s}(\hat{\boldsymbol{\eta}}) - 2\mathbf{s}_i^T \mathbf{s}(\hat{\boldsymbol{\eta}})$$

but since the middle term does not depend on i , we choose model i as

$$\arg \min_i [\epsilon_i - 2\mathbf{s}_i^T \mathbf{s}(\hat{\boldsymbol{\eta}})] = \arg \max_i [2\mathbf{s}_i^T \mathbf{A}(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{x} - \epsilon_i] \quad (48)$$

Now with only three components we can solve the matrix inverse analytically resulting in

$$\mathbf{B} = \mathbf{A}(\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T = \frac{\epsilon_2 \mathbf{s}_1 \mathbf{s}_1^T - \rho \mathbf{s}_2 \mathbf{s}_1^T + \epsilon_1 \mathbf{s}_2 \mathbf{s}_2^T - \rho \mathbf{s}_1 \mathbf{s}_2^T}{\epsilon_1 \epsilon_2 - \rho^2} \quad (49)$$

It is then easy to show that $\mathbf{s}_i^T \mathbf{B} \mathbf{x} = \mathbf{s}_i^T \mathbf{x}$ and thus (48) is equivalent to the ML rule.

As we learned from the pairwise EEF procedure, care must be taken in how the MLE is found. In the M-ary EEF simulations showed that if the MLE of the embedding parameters was calculated using both bound and sum constraints, then the EEF procedure was not equivalent to the ML rule. In particular, it was only equivalent when the connecting models were oriented as an “acute connection”. However, if the MLE was found with only a sum constraint then they were truly equivalent. That is, it must be allowed that some weights may be negative, as the M-ary EEF may still be a valid PDF for all embedding parameters for which the

cumulative generating function is finite. In addition, from [14] we know that for the embedding parameters to be identifiable the sufficient statistics (i.e. the log likelihood ratios $M - 1$ components paired with an arbitrary reference component) must be linearly independent. This is also equivalent to the Fisher information matrix being of full rank [9]. Assuming a unique MLE exists, in many cases there will not be a closed form analytical solution for it. For example, if connecting Gaussian components with different covariance matrices, there is no closed form, and thus it becomes a non linear programming problem to solve for the embedding parameters. To allow the search to include possibly negative weights without diverging one has to consider a constraint on the covariance matrix of the EEF to be positive definite.

1D.2 M-ary EEF of partially unknown PDFs

The above results were for completely known (specified) hypotheses. Consider the problem of choosing among M hypotheses $\mathcal{H}_i$, $i = 1, 2, \dots, M$ where

$$\mathcal{H}_i : \mathbf{x} = A_i \mathbf{s}_i + \mathbf{w} \quad (50)$$

with $\mathbf{w} \equiv \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$, $\mathbf{s}_i$, known linearly independent signals of length n , and A_i an unknown scalar (nuisance) parameter. This class of problem is known as a composite hypothesis test and has no optimal approach associated with it. Under these assumptions a reasonable approach would be to use a Generalized Likelihood Ratio test defined as choosing the i^{th} model as

$$\arg \max_i \frac{p_i(\mathbf{x}; \hat{A}_i)}{p_0(\mathbf{x})} > \gamma \quad (51)$$

where $p_0(\mathbf{x})$ is the PDF of the null hypothesis, i.e. $p_0(\mathbf{x}) \equiv \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$, γ is some pre-defined threshold and $\hat{A}_i$ is the MLE of the amplitude given the data under the i^{th} model. More specifically in this analysis we will disregard the null hypothesis

and consider the selection as

$$\arg \max_i p_i(\mathbf{x}; \hat{A}_i) \quad (52)$$

that is we will choose the hypothesis with maximum likelihood. For the simple Gaussian problem above this is written as choosing $\mathcal{H}_i$ such that

$$\arg \max_i \|\mathbf{x}^T \mathbf{s}_i\|^2 \quad (53)$$

We consider below the M-ary EEF as an alternative approach.

1D.2.1 Derivation of M-ary EEF

Consider M hypotheses $H_1, H_2, \dots, H_M$ where

$$H_i : \mathbf{x} \sim \mathcal{N}(A_i \mathbf{s}_i, \sigma^2 \mathbf{I}) \quad (54)$$

with the data $\mathbf{x}$ given as n iid sampled points. We assume that the noise variance σ^2 is known as well as each of the signal vectors $\mathbf{s}_i$ which are assumed to be linearly independent. The amplitudes (A_i) corresponding to each model are unknown. Given a data realization we want to define an approach to choose the hypothesis which produced it among these given choices where each hypothesis is equal likely (equal priors). We will consider the M-ary EEF written as

$$p(\mathbf{x}; \boldsymbol{\eta}) = \exp \left(\sum_{i=1}^M \eta_i \ell_i(\mathbf{x}) - \mathcal{K}_0(\boldsymbol{\eta}) + \ln p_0(\mathbf{x}) \right) \quad (55)$$

where $\ell_i(\mathbf{x}) = \ln \frac{p_i(\mathbf{x}; \hat{A}_i)}{p_0(\mathbf{x})}$. Now the sufficient statistic for A_i is

$$T_i(\mathbf{x}) = \frac{\mathbf{x}^T \mathbf{s}_i}{\mathbf{s}_i^T \mathbf{s}_i} \quad (56)$$

Using this we can rewrite $\ell_i(\mathbf{x})$ as

$$\ell_i(\mathbf{x}) = \ln \frac{p_i(\mathbf{x}; A_i)}{p_{T_i}(\mathbf{t}_i; A_i) p_0(\mathbf{x})} + \ln p_{T_i}(\mathbf{t}_i; A_i) \quad (57)$$

where p_{T_i} is the PDF of the sufficient statistic. We assume for unknown A_i that the latter term above is constant, i.e. $c_i = p_{T_i}(\mathbf{t}_i; A_i)$ and absorb this into $\bar{\mathcal{K}}_0(\boldsymbol{\eta})$ above. This yields the reduced M-ary EEF as

$$p(\mathbf{x}; \boldsymbol{\eta}) = \exp \left(\sum_{i=1}^M \eta_i \bar{\ell}_i(\mathbf{x}) - \bar{\mathcal{K}}_0(\boldsymbol{\eta}) + \ln p_0(\mathbf{x}) \right) \quad (58)$$

with $\bar{\ell}_i(\mathbf{x}) = \ln \frac{p_i(\mathbf{x}; A_i)}{p_{T_i}(\mathbf{t}_i; A_i) p_0(\mathbf{x})}$. Note that since T_i is a sufficient statistic for A_i , then $\frac{p_i(\mathbf{x}; A_i)}{p_{T_i}(\mathbf{t}_i; A_i)}$ does not depend on A_i . Thus if we let $A_i = 0$ then $\bar{\ell}_i(\mathbf{x})$ reduces as

$$\begin{aligned} \bar{\ell}_i(\mathbf{x}) &= \ln \frac{1}{p_{T_i}(\mathbf{t}_i; 0)} \\ &= c_i + \frac{\mathbf{t}_i^T \mathbf{s}_i^T \mathbf{s}_i \mathbf{t}_i}{2\sigma^2} \\ &= c_i + \frac{\mathbf{x}^T \mathbf{s}_i (\mathbf{s}_i^T \mathbf{s}_i)^{-1} \mathbf{s}_i^T \mathbf{x}}{2\sigma^2} \end{aligned} \quad (59)$$

So that after absorbing the constant into $\bar{\mathcal{K}}_0(\boldsymbol{\eta})$ we have

$$p(\mathbf{x}; \boldsymbol{\eta}) = \exp \left(\sum_{i=1}^M \eta_i \frac{\mathbf{x}^T P_{H_i} \mathbf{x}}{2\sigma^2} - \bar{\mathcal{K}}_0(\boldsymbol{\eta}) + \ln p_0(\mathbf{x}) \right) \quad (60)$$

where $P_{H_i} = \mathbf{s}_i (\mathbf{s}_i^T \mathbf{s}_i)^{-1} \mathbf{s}_i^T$. Now we can write $\bar{\mathcal{K}}_0(\boldsymbol{\eta})$ as

$$\begin{aligned} \bar{\mathcal{K}}_0(\boldsymbol{\eta}) &= \ln \int_{\mathbb{R}} \exp \left(\sum_{i=1}^M \eta_i \frac{\mathbf{x}^T P_{H_i} \mathbf{x}}{2\sigma^2} \right) p_0(\mathbf{x}) d\mathbf{x} \\ &= \ln E_0 \left[\exp \left(\sum_{i=1}^M \eta_i \zeta_i \right) \right] \end{aligned} \quad (61)$$

where $\zeta_i = \frac{\mathbf{x}^T P_{H_i} \mathbf{x}}{2\sigma^2}$. Since $\frac{\mathbf{x}}{\sigma} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ under H_0 , then we can rewrite

$$\sum_{i=1}^M \eta_i \zeta_i = \mathbf{u}^T \left[\frac{\sum_{i=1}^M \eta_i P_{H_i}}{2} \right] \mathbf{u}$$

where $\mathbf{u} = \frac{\mathbf{x}}{\sigma}$. Then it is well known that if $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ then

$$E_0 [\exp(\mathbf{u}^T A \mathbf{u})] = \frac{1}{\det(\mathbf{I} - 2A)^{1/2}} \quad (62)$$

thus the CGF becomes

$$\bar{\mathcal{K}}_0(\boldsymbol{\eta}) = -\frac{1}{2} \ln \det \left(\mathbf{I} - \sum_{i=1}^M \eta_i P_{H_i} \right) \quad (63)$$

and finally M-ary EEF is written as

$$\ln p(\mathbf{x}; \boldsymbol{\eta}) = \sum_{i=1}^M \eta_i \zeta_i + \frac{1}{2} \ln \det \left(\mathbf{I} - \sum_{i=1}^M \eta_i P_{H_i} \right) + \ln p_0(\mathbf{x}) \quad (64)$$

Note the embedding parameters (η_i) must be such that the above is a valid PDF. Since this PDF can be shown to be $\mathcal{N} \left(\mathbf{0}, \sigma^2 \left(\mathbf{I} - \sum_{i=1}^M \eta_i \mathbf{P}_{H_i} \right)^{-1} \right)$, the η_i 's should be such that the matrix $\mathbf{B} = \mathbf{I} - \sum_{i=1}^M \eta_i P_{H_i}$ is positive definite. As a model selection approach we maximize the M-ary EEF with respect to $\boldsymbol{\eta}$ (under the constraint that $\mathbf{B}$ is positive definite) and choose the model with the largest weight. The expression in (64) does not in general yield an analytical solution for the unknown embedding vector $\boldsymbol{\eta}$. However, under some simplified signal assumptions an analytical solution can be found.

1D.2.2 Orthonormal Signal Models

Let the $\mathbf{s}_i$ above be such that for all i , $\mathbf{s}_i^T \mathbf{s}_i = 1$ and $\mathbf{s}_i^T \mathbf{s}_j = 0$, $j \neq i$. Then it can be shown that $\sum_{i=1}^M \eta_i P_{H_i}$ is a rank M matrix with eigenvalues $\eta_1, \dots, \eta_M, 0, \dots, 0$. Thus the eigenvalues of $\mathbf{I} - \sum_{i=1}^M \eta_i P_{H_i}$ are $1 - \eta_1, \dots, 1 - \eta_M, 1, \dots, 1$, and

$$\ln \det \left(\mathbf{I} - \sum_{i=1}^M \eta_i P_{H_i} \right) = \ln \left(\prod_{i=1}^M (1 - \eta_i) \right) = \sum_{i=1}^M \ln (1 - \eta_i) \quad (65)$$

Using (65) in (64) we have

$$\ln p(\mathbf{x}; \boldsymbol{\eta}) = \sum_{i=1}^M \eta_i \zeta_i + \frac{1}{2} \sum_{i=1}^M \ln (1 - \eta_i) + \ln p_0(\mathbf{x}) \quad (66)$$

Taking the derivative with respect to η_i and solving gives

$$\hat{\eta}_i = \frac{\|\mathbf{x}^T \mathbf{s}_i\|^2 - \sigma^2}{\|\mathbf{x}^T \mathbf{s}_i\|^2} \quad (67)$$

thus choosing the model corresponding to the highest embedding parameter (unconstrained in this case) is equivalent to choosing model i such that

$$\arg \max_i \|\mathbf{x}^T \mathbf{s}_i\|^2 \quad (68)$$

which is equivalent to (53).

1D.2.3 Simplified form of M-ary EEF $M = 2$

As shown above, in the case of orthonormal signals, the M-ary EEF is equivalent to the GLRT. In order to study this for more general signal vectors, we derive a simplified form and implement in a simulation for a numerical comparison. Recall that the P_{H_i} are projection matrices so that $\mathbf{B}$ is the sum of M rank one matrices. Thus $\mathbf{B}$ will have $N - M$ eigenvalues equal to 1 and M eigenvalues $1 - \lambda_i$, $i = 1, \dots, M$, where λ_i are the eigenvalues of $\mathbf{A} = \sum_{i=1}^M \eta_i P_{H_i}$ the sum of M rank one matrices. For $M = 2$, $\mathbf{A} = \eta_1 P_{H_1} + \eta_2 P_{H_2}$. Note that by properties of the trace operator that $\text{tr}(\mathbf{A}^k) = \sum_i \lambda_i^k$. Thus for $P_{H_i} = \mathbf{s}_i(\mathbf{s}_i^T \mathbf{s}_i)^{-1} \mathbf{s}_i^T$ the trace of $\mathbf{A}$ is given as

$$\text{tr}(\mathbf{A}) = \eta_1 + \eta_2 = \lambda_1 + \lambda_2 \quad (69)$$

and the trace of $\mathbf{A}^2$ is

$$\text{tr}(\mathbf{A}^2) = \eta_1^2 + \eta_2^2 + 2\eta_1\eta_2 \frac{(\mathbf{s}_1^T \mathbf{s}_2)^2}{\mathbf{s}_1^T \mathbf{s}_1 \mathbf{s}_2^T \mathbf{s}_2} = \lambda_1^2 + \lambda_2^2 \quad (70)$$

Solving these two equations for λ_1, λ_2 we get

$$\lambda_1, \lambda_2 = \frac{\eta_1 + \eta_2 \pm \sqrt{(\eta_1 + \eta_2)^2 - 4\eta_1\eta_2(1 - \gamma)}}{2} \quad (71)$$

where $\gamma = \frac{(\mathbf{s}_1^T \mathbf{s}_2)^2}{\mathbf{s}_1^T \mathbf{s}_1 \mathbf{s}_2^T \mathbf{s}_2}$. Note that for matrix $\mathbf{B} = \mathbf{I} - \mathbf{A}$ to be positive definite, the eigenvalues of $\mathbf{A}$ must all be less than one. In addition, for $\mathbf{s}_1^T \mathbf{s}_2 = 0$, $\gamma = 0$ and $\lambda_1 = \eta_1$, $\lambda_2 = \eta_2$ as expected.

Now with an analytical form of the eigenvalues we can write the determinant of $\mathbf{B}$ as $(1 - \lambda_1)(1 - \lambda_2) = 1 + \lambda_1\lambda_2 - (\lambda_1 + \lambda_2)$, but $\lambda_1 + \lambda_2 = \eta_1 + \eta_2$ and $\lambda_1\lambda_2 = \eta_1\eta_2(1 - \gamma)$. Thus

$$\det \left(\mathbf{I} - \sum_{i=1}^2 \eta_i P_{H_i} \right) = g(\eta_1, \eta_2) = 1 - (\eta_1 + \eta_2) + \eta_1\eta_2(1 - \gamma) \quad (72)$$

Note that γ is clearly in $[0, 1]$ and that valid η_i must be such that $g(\eta_1, \eta_2) > 0$.

We can now re-write our original objective in a simplified form (for $M = 2$) as

$$\ln p(\mathbf{x}; \boldsymbol{\eta}) = \sum_{i=1}^2 \eta_i \zeta_i + \frac{1}{2} \ln g(\eta_1, \eta_2) + \ln p_0(\mathbf{x}) \quad (73)$$

The derivatives of this function are then readily calculated as

$$\frac{\partial \ln p(\mathbf{x}, \boldsymbol{\eta})}{\partial \eta_1} = \zeta_1 + \frac{\eta_2(1 - \gamma) - 1}{2g(\eta_1, \eta_2)} \quad (74)$$

$$\frac{\partial \ln p(\mathbf{x}, \boldsymbol{\eta})}{\partial \eta_2} = \zeta_2 + \frac{\eta_1(1 - \gamma) - 1}{2g(\eta_1, \eta_2)} \quad (75)$$

and the Fisher Information Matrix $\mathbf{I}(\eta_1, \eta_2)$ is

$$\mathbf{I}(\eta_1, \eta_2) = \frac{1}{2g^2(\eta_1, \eta_2)} \begin{bmatrix} (\eta_2(1 - \gamma) - 1)^2 & \gamma \\ \gamma & (\eta_1(1 - \gamma) - 1)^2 \end{bmatrix} \quad (76)$$

We implemented (73) and (53) for two linearly independent sinusoids as in Manuscript 1. Our experimental findings were that due to bound constraints on $\boldsymbol{\eta}$ the M-ary EEF was subject to the same tie phenomena seen in the original EEF. Removal of ties showed that the performance was almost identical to the GLRT. At that point we determined that the M-ary EEF was inferior to the GLRT for these types of problems.

List of References

- [1] S. Kay, “Exponentially embedded families - new approaches to model order estimation,” *IEEE Journal of AES*, vol. 41, no. 1, pp. 333–345, 2005.
- [2] J. Rissanen, “Modeling by shortest data description,” *Automatica*, vol. 14, pp. 465–471, 1978.
- [3] S. Kay, “The multifamily likelihood ratio test for multiple signal model detection,” *IEEE Signal Processing Letters*, vol. 12, no. 5, pp. 369–371, 2005.
- [4] S. Kullback, *Information Theory and Statistics*. Dover, 1968.
- [5] S. Amari and H. Nagaoka, *Methods of Information Geometry*. Oxford University Press, 2000.
- [6] S. M. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory*. Prentice Hall, 1998.

- [7] A. Wald, “Note on the consistency of the maximum likelihood estimate,” *Annals of Mathematical Statistics*, vol. 20, pp. 595–601, 1949.
- [8] A. Bjorck, *Numerical Methods for Least Squares Problems*. SIAM, 1996.
- [9] P. Bickel and K. Doksum, *Mathematical Statistics: Basic Ideas and Selected Topics*. Prentice Hall, 2007.
- [10] V. Rajan, “Optimality of the Delaunay triangulation in R^d ,” *Discrete Computational Geometry*, vol. 12, no. 2, pp. 189–202, 1994.
- [11] E. Vanderzee, A. Hirani, D. Guoy, and E. Ramos, “Well-centered triangulation,” *SIAM Journal on Scientific Computing*, vol. 31, no. 6, pp. 4497–4523, 2010.
- [12] D. Q. Naiman and H. P. Wynn, “On intersecting a point set with Euclidean balls,” *Computational Geometry*, vol. 7, pp. 237–244, 1997.
- [13] Q. Ding and S. Kay, personal communication.
- [14] E. L. Lehmann and G. Casella, *Theory of point estimation*. Springer, 1998.

CHAPTER 2

**An extension of the equidistant property of the Chernoff point in
exponentially embedded families of probability density functions**

by

Russell Costa and Steven Kay

Dept. of Electrical, Computer and Biomedical Engineering

University of Rhode Island, Kingston, RI, USA

submitted to IEEE Transactions on Signal Processing

Abstract

The exponentially embedded family (EEF) of probability density functions (PDFs) is an important modeling paradigm in a number of signal processing applications including data fusion and hypothesis testing. The Chernoff distance (information measure) is also important in establishing asymptotic bounds in Bayes hypothesis tests as well as being used for data fusion and other decision problems. From the work of Chernoff and Kullback it is well known that the two concepts are linked by the Chernoff distribution (or the “Chernoff point”); the PDF within the EEF which is equi-distant to the component PDFs being embedded, where the “distance” measure is given by the Kullback Leibler Divergence (KL-divergence). The equi-distance property has been utilized in recent years as a criterion for combining information (PDFs) in data fusion applications, however, much of that work has been limited to combining PDFs in a pairwise fashion. In this paper we utilize results from information geometry in order to study the extension of the “Chernoff point” for the EEF of more than two components. Using well known theory of exponential families we establish the proper existence criteria of such an extension, and show that much of the geometry involved can be well understood by drawing analogies between the EEF of three component PDFs, and the Euclidean geometry of triangles. To illustrate these relationships we derive simple examples in univariate Gaussian parameter space. Finally, we establish a theorem for determining the proper constraints on embedding parameters for solving the optimization problems associated with determining the extended Chernoff point.

2.1 Introduction

The Chernoff distance [1] is a well established functional measure between two hypotheses (probability density functions) that provides a bound on asymptotic performance in Bayesian hypothesis testing [2, 3, 4]. Given two hypotheses

represented by PDFs, the Chernoff distance is defined as

$$C(p_1, p_0) = - \min_{0 \leq \eta \leq 1} \log \left(\int_{\mathbb{R}^n} p_1^\eta(\mathbf{x}) p_0^{1-\eta}(\mathbf{x}) d\mathbf{x} \right). \quad (77)$$

The result of the minimization η^* , defines the PDF for which it is true that $C(p_1, p_0) = D(p_{\eta^*} \| p_0) = D(p_{\eta^*} \| p_1)$ where $D(\cdot \| \cdot)$ is the Kullback Leibler (KL) divergence written as [5]

$$D(p_1 \| p_0) = \int_{\mathbb{R}^n} p_1(\mathbf{x}) \log \left(\frac{p_1(\mathbf{x})}{p_0(\mathbf{x})} \right) d\mathbf{x} \quad (78)$$

and p_{η^*} is a PDF in the exponentially embedded family (EEF) written as

$$p_\eta(\mathbf{x}) = \frac{p_1^\eta(\mathbf{x}) p_0^{1-\eta}(\mathbf{x})}{\int_{\mathbb{R}^n} p_1^\eta(\mathbf{x}) p_0^{1-\eta}(\mathbf{x}) d\mathbf{x}} \quad (79)$$

Thus the Chernoff distance is given by the KL divergence of a “mid-point” PDF (within a family of PDFs which connect two given component PDFs) from its component PDFs [2] where the equidistant PDF, p_{η^*} is known as the Chernoff distribution or Chernoff point [6]. In particular, the Chernoff distance is a “left-sided” KL-divergence since the parameterized PDF is on the left side. From [7] there exist some interesting connections between the Chernoff point, EEF, KL divergence, and log likelihood ratios. Firstly, the Chernoff point, $p_{\eta^*}(\mathbf{x})$ represents an *information center* (I-center) separating $p_0(\mathbf{x})$ and $p_1(\mathbf{x})$ in that a random sample taken from this density would be noninformative from a Bayesian hypothesis testing (i.e. log likelihood ratio) perspective. That is for a Bayes hypothesis test (with equal priors) asymptotically the probability of choosing the alternative hypothesis $P\{\mathcal{H}_1 | \mathbf{x}\}$ is equal to the probability of choosing the null hypothesis $P\{\mathcal{H}_0 | \mathbf{x}\}$, when $\mathbf{x}$ is distributed as p_{η^*} . Second, Kullback showed that the EEF is the (functional) solution of minimizing (over $p \in \mathcal{P}$ where $\mathcal{P}$ is the space of PDFs) $D(p \| p_0)$ under the constraint $D(p \| p_0) - D(p \| p_1) = c$ for some $c \in \mathbb{R}$. Thus the EEF is a set of PDFs which connect p_0 and p_1 with minimal “distance”, with KL divergence

playing the role of distance. Finally, since $D(p \parallel p_0) - D(p \parallel p_1) = E_p \left[\ln \frac{p_1(\mathbf{x})}{p_0(\mathbf{x})} \right]$, we observe that the EEF connecting $p_0(\mathbf{x})$ to $p_1(\mathbf{x})$ maps out an increasing function (with η) with respect to the log likelihood ratio. It can then be shown that the pdf p_{η^*} is also the solution of

$$\min_{p \in \mathcal{P}} D(p \parallel p_0) \quad \text{such that} \quad E_p \left[\ln \frac{p_1(\mathbf{x})}{p_0(\mathbf{x})} \right] = 0 \quad , \quad (80)$$

or $c = 0$ from above. Thus the Chernoff point (PDF) is not only a mid-point of a *geodesic* but it provides a minimum distance among the set of all PDFs that are equidistant with respect to KL divergence to the component PDFs. See [8, 9] for detailed definitions of *geodesics* in PDF spaces.

In the data fusion / tracking community the EEF is commonly referred to as an exponential mixture family, and has been utilized for conservative fusion of information from component PDFs with unknown correlations [10, 11, 12]. The impetus for such a conservative approach is born of the numerous distributed sensor detection and estimation problems for which the sharing of data among sensors is not plausible, and thus unknown correlations may exist between PDF estimates [13]. The sub-optimal fusion method of covariance intersection [14] is an example of such an approach where the EEF component weights are calculated (via optimization) to minimize the determinant of the fused covariance matrix. Within this framework, the equidistant property of Chernoff has been considered as a criterion for choosing the mixture weights for fusion, with the Chernoff point then representing the solution for the fused information PDF [15, 16]. Since Chernoff's results are defined for pairs of PDFs the primary focus in the fusion community has been on pair-wise fusion as in a sequential Kalman filtering (like) scheme. That said, the sequential pairwise combination of PDFs via their respective Chernoff points can easily be shown to be order dependent. Thus it is difficult to apply to an arbitrary number of PDFs, typical in most practical applications. Application of

these ideas beyond the pairwise paradigm have been undertaken under the name of Generalized Chernoff fusion [17] and convex combinations of PDFs [11], but neither study establishes the existence and theoretical properties of a PDF equidistant to the component PDFs being combined. This shortfall in the use of the Chernoff criteria in distributed data fusion is in part the motivation for this study. In a recent paper Nielsen characterized the Chernoff point in exponential families using Information Geometry [6]. In that paper it was shown that the Chernoff point lies at the intersection of the e-geodesic (EEF) and a bisecting mixture family (m-geodesic), with the m-geodesic corresponding to the family of PDFs which meet the constraint in (80). In this paper we show that the extension of the Chernoff point, to more than two PDFs, exists at the intersection of the m-geodesic bisectors (orthogonal linear families) corresponding to the pairwise EEFs of the component PDFs to be combined (embedded).

Much of the theory to effect the aforementioned extension exists in the work of Cencov [8], where the EEF generalized to an arbitrary number of component PDFs was first introduced. Cencov refers to this family as a geodesic family and shows that as an exponential family it is an affine space governed by the left-sided Kullback Leibler divergence. The tools of information geometry utilized in this paper are mainly due to Cencov and later results of Amari. Amari referred to the EEF as an exponential connection or e-connection from [9]. In both works the machinery of differential geometry is utilized to study PDFs on a finite dimensional manifold of probability measures. Amari focuses primarily on parameterized probability density functions and introduces a broad class of divergences named the α -family of divergences, with corresponding α -connections, of which the KL divergence (left-sided) and EEF are respective members. In a subsequent paper this family of connections was shown to correspond to generalized means of pdfs

[18, 19], where in this context the EEF is a weighted geometric mean of PDFs normalized to sum to one. This generalized mean perspective has also been studied in combining more general information [20] and specifically the geometric mean leading to log opinion pools for combination of experts [21].

A key property of the EEF is that it generates an exponential family. The exponential family of PDFs has many powerful properties for inference such as readily available sufficient statistics and easily obtained bounds on estimation performance [22, 23]. This is particularly useful if the component PDFs are from the same exponential family. In this case the EEF produces a form within this same family. In this paper we focus our efforts on working within the natural parameter space of the exponential family generated by the EEF. Since the geodesic properties of this space are governed with respect to the left-sided KL divergence, all our definitions are with respect to that directional divergence. We also make use of Pythagorean theorems which apply to a broad class of directional divergences (α -divergence, Bregman, etc) [24]. Furthermore we exploit the reproductive property with Gaussian components for a more pedagogical perspective on Information Geometry and how it applies to the extension of the equidistant property of the Chernoff point. We refer to the PDF attaining this equidistant property for greater than two EEF components as an “extended Chernoff point”. While this result may have broad yet unknown applications, the main contribution of this paper is the extension of equidistance property motivated primarily by its use in fusion of information represented by PDFs, such as the distributed sensing and log opinion pools applications mentioned above. In section 2.2 we establish the existence criteria for an extension of the Chernoff point for greater than two component EEFs. In section 2.3 we define information centers and smallest enclosing balls of PDFs analogous to Euclidean geometrical objects such as circumcenters and smallest

enclosing circles in the plane, respectively. We restrict all definitions to that of the natural parameters of the exponential family generated by the EEF, rather than utilizing dual parameters as in [25]. In our opinion, by avoiding this extra notation our main results are more accessible to the general statistical signal processing community. In section 2.4, we analytically and numerically solve a number of examples using univariate Gaussian components, where the above reproductive property of the EEF is in full display. In section 2.5 we utilize these examples to expose some interesting properties of the extended Chernoff point and state and prove a theorem for establishing triangle like characterization of the the EEF of three components.

2.2 Exponentially Embedded Families and the Extended Chernoff Point

Definition 1. *EEF of M component PDFs*

We define the EEF extended to an arbitrary number of component PDFs, M , with component PDFs $p_i(\mathbf{x})$, $i = 0, \dots, M - 1$ and embedding weights η_i as

$$p(\mathbf{x}; \boldsymbol{\eta}) = \exp \left(\sum_{i=0}^{M-1} \eta_i \ln p_i(\mathbf{x}) - \mathcal{K}_0(\boldsymbol{\eta}) \right) \quad (81)$$

with the cumulant generating function (CGF) $\mathcal{K}_0(\boldsymbol{\eta}) = \ln \int p_0^{\eta_0}(\mathbf{x}) p_1^{\eta_1}(\mathbf{x}) \cdots p_{M-1}^{\eta_{M-1}}(\mathbf{x}) d\mathbf{x}$, and all $p_i(\mathbf{x})$ defined on common support $\mathcal{X}$. This is a valid PDF family for all $\boldsymbol{\eta} \in \Upsilon$ such that $\mathcal{K}_0(\boldsymbol{\eta})$ is finite, i.e. $\Upsilon = \{\boldsymbol{\eta} : |\mathcal{K}_0(\boldsymbol{\eta})| < \infty\}$ [23].

The M -component EEF (81), is readily recognized as a canonical (or natural) parametrization of an exponential family [23], and will have *rank* $M - 1$ if and only if the sufficient statistics $T_1(\mathbf{x}), \dots, T_{M-1}(\mathbf{x})$ are linearly independent functions of $\mathbf{x}$ with positive probability [26], where $T_i(\mathbf{x}) = \ln \frac{p_i(\mathbf{x})}{p_0(\mathbf{x})}$.

Definition 2. *The natural parameter space of the M -component EEF and its subsets*

Let $\mathcal{Y}$ be the natural parameter space for the EEF of M component PDFs such that the CGF $\mathcal{K}_0(\boldsymbol{\eta})$ is finite for all $\boldsymbol{\eta} \in \mathcal{Y}$. We define subsets of this parameter space $\mathcal{Y}_a$ and $\mathcal{Y}_p$ as

$$\mathcal{Y}_a = \{\boldsymbol{\eta} \in \mathcal{Y} : \sum_{i=0}^{M-1} \eta_i = 1\} \quad (82)$$

$$\mathcal{Y}_p = \{\boldsymbol{\eta} \in \mathcal{Y} : \sum_{i=0}^{M-1} \eta_i = 1, \forall_i \eta_i \geq 0\} \quad (83)$$

Clearly $\mathcal{Y}_p \subset \mathcal{Y}_a \subset \mathcal{Y}$.

The full *rank* condition of the EEF insures that the embedding parameters $\boldsymbol{\eta}$ are identifiable, i.e. each parameter value maps out a unique PDF in the family and that the cumulant generating function $\mathcal{K}_0(\boldsymbol{\eta})$ is convex in $\mathcal{Y}$. See [8, 23, 26] for further details on the properties of exponential families. In this paper we will assume that the component PDFs always meet this full rank condition. In practice, an exponential family can always be reduced to a *minimal* representation, i.e. a $k < M - 1$ full rank family, by removing the redundant components [27].

Definition 3. *Extended Chernoff point*

We write the extended Chernoff point (PDF) as $p_{\boldsymbol{\eta}^}$ where*

$$\boldsymbol{\eta}^* = \arg \min_{\boldsymbol{\eta} \in \mathcal{Y}_a} \mathcal{K}_0(\boldsymbol{\eta}) \quad (84)$$

with $\mathcal{K}_0(\boldsymbol{\eta})$ as the CGF of the M -component EEF and $\boldsymbol{\eta}$ is an $M \times 1$ vector with $M - 1$ free parameters (corresponding to component PDFs).

We will show that the embedding parameter, $\boldsymbol{\eta}^*$ at which the minimum of the CGF occurs, defines a PDF within the EEF such that $D(p_{\boldsymbol{\eta}^*} \| p_i) = D(p_{\boldsymbol{\eta}^*} \| p_j)$ for all $i, j \in (0, 1, \dots, M - 1)$. This is essentially an extension of the well known binary case in (80). Note that the restriction to $\mathcal{Y}_a \subset \mathcal{Y}$ refines the problem to the proper

space which contains the desired generalization of the Chernoff PDF and results in an interesting relationship between the CGF and the KL-divergence of the EEF to its component PDFs.

2.2.1 Extension of the binary mid-point equivalence property

Define $p_\eta(\mathbf{x})$ as the EEF of PDFs $p_0(\mathbf{x})$, $p_1(\mathbf{x})$ as in (79). Let $\mathcal{K}_0(\eta)$ be the CGF for $p_\eta(\mathbf{x})$, then it is well established (an exercise in Cover and Thomas [2]) that the two distinct conditions

$$\eta^* = \arg_{\eta \in [0,1]} (D(p_\eta \| p_0) = D(p_\eta \| p_1)) \quad (85)$$

$$\eta_* = \arg \min_{\eta \in [0,1]} \mathcal{K}_0(\eta) \quad (86)$$

result in $\eta^* = \eta_*$ where both are equivalent to the condition $E_{p_\eta} \left[\ln \frac{p_1(\mathbf{x})}{p_0(\mathbf{x})} \right] = 0$, for $\eta = \eta^* = \eta_*$. As a new result we can extend this mid-point equivalence to M component PDFs. To do this we must prove that both approaches yield the vector $\boldsymbol{\eta}$ such that $E_{p\boldsymbol{\eta}} \left[\ln \frac{p_i(\mathbf{x})}{p_0(\mathbf{x})} \right] = 0$ for all $i = 1, \dots, M-1$. We can write the KL-divergence of the M -component EEF with respect to an arbitrary component density $p_i(\mathbf{x})$ using (81) as

$$D(p_{\boldsymbol{\eta}} \| p_i) = E_{p\boldsymbol{\eta}} \left[\sum_{j=0}^{M-1} \eta_j \ln p_j(\mathbf{x}) - \mathcal{K}_0(\boldsymbol{\eta}) - \ln p_i(\mathbf{x}) \right] \quad (87)$$

If we take $i = 0$ as a reference, we can form the $M-1$ equations

$$\Delta_i = D(p_{\boldsymbol{\eta}} \| p_i) - D(p_{\boldsymbol{\eta}} \| p_0) = E_{p\boldsymbol{\eta}} \left[\ln \frac{p_i(\mathbf{x})}{p_0(\mathbf{x})} \right] \quad (88)$$

Now we must show that minimizing the CGF is equivalent to the same criteria on the log likelihood ratio. Note that since the EEF is an exponential family for all $\boldsymbol{\eta} \in \mathcal{T}$, its CGF is a convex function of $\boldsymbol{\eta}$. Thus we can find a unique minimum solution by taking the derivative with respect to $\boldsymbol{\eta}$ and setting this equal to zero.

We can write for all i

$$\begin{aligned}\frac{\partial \mathcal{K}_0(\boldsymbol{\eta})}{\partial \eta_i} &= \frac{\partial / \partial \eta_i \int \exp \left(\sum_{j=0}^{M-1} \eta_j \ln \frac{p_j(\mathbf{x})}{p_0(\mathbf{x})} + \ln p_0(\mathbf{x}) \right) d\mathbf{x}}{\exp(\mathcal{K}_0(\boldsymbol{\eta}))} \\ &= E_{p\boldsymbol{\eta}} \left[\ln \frac{p_i(\mathbf{x})}{p_0(\mathbf{x})} \right] = 0\end{aligned}\tag{89}$$

which is equivalent to (88), when $\Delta_i = 0$. Recall that a unique solution exists under the conditions that the sufficient statistics of the EEF are linearly independent, resulting in a *full* exponential family. Thus we observe that under proper conditions the equidistant property of the Chernoff point (equidistant PDF) can be extended to the EEF of M component densities. Note that when $M = 2$ the embedding parameter (scalar) η is necessarily in the interval $[0, 1]$, due to the monotonic property of the KL divergence. However, as we will see later in the paper, for $M > 2$ the extended Chernoff point may correspond to some negative embedding parameters.

2.2.2 An interesting property of the Cumulative Generating Function and an extension of Chernoff distance

As pointed out earlier, the Chernoff distance is just the KL divergence of the Chernoff point to the component PDFs. Further for the M -component case we can write the KL-divergence of the EEF to a component PDF $p_j(\mathbf{x})$ as

$$D(p\boldsymbol{\eta} \| p_j) = E_{p\boldsymbol{\eta}} \left[\sum_{i \neq j} \eta_i \ln \frac{p_i(\mathbf{x})}{p_j(\mathbf{x})} \right] - \mathcal{K}_0(\boldsymbol{\eta})\tag{90}$$

Consider the function

$$R(\boldsymbol{\eta}) = \sum_{j=0}^{M-1} \eta_j D(p\boldsymbol{\eta} \| p_j)\tag{91}$$

the weighted sum of the KL-divergence of the EEF to all of its component densities.

Then by substitution of (90) into (91) we have

$$R(\boldsymbol{\eta}) = -\mathcal{K}_0(\boldsymbol{\eta}) + \sum_{j=0}^{M-1} \eta_j E_{p\boldsymbol{\eta}} \left[\sum_{i \neq j} \eta_i \ln \frac{p_i}{p_j} \right] = -\mathcal{K}_0(\boldsymbol{\eta})$$

since the sum in the second term is equal to zero (by the sum constraint on the embedding weights η_i). Thus the CGF of the EEF can be written as the negative of the weighted sum of KL-divergences relative to its component densities. An immediate consequence of this is that *the extended Chernoff point is the PDF in the EEF which maximizes the weighted sum of the divergences*. From this result we see that the actual measure (value) of Chernoff distance (77) is a weighted sum of KL divergences of the EEF to its components. Such an extension of Chernoff distance has been mentioned in several papers as a potential bound for classification error in multi-class problems [28, 29] but with little or no details on existence criteria or in calculating it in practice. Since the Chernoff distance measure is in general a sum of KL divergence measures, it inherits several useful properties, such as being a positive measure and being equal to zero if and only if the component pdfs are equal almost everywhere. In contrast to the KL divergence, the Chernoff distance is a symmetric measure in that $C(p_0, p_1) = C(p_1, p_0)$, but it does not satisfy the triangle inequality thus it is not a metric distance. Despite that limitation, we will show through examples that the Chernoff distance does behave like a squared distance in the space of PDF mapped out by the EEF.

2.3 Geometrical definitions in Euclidean and PDF spaces

In this section we interpret the geometry of PDF spaces (governed by a directional divergence not a metric) through analogous concepts in Euclidean geometry. Consider two points $\mathbf{z}_1 = (x_1, y_1)$, $\mathbf{z}_2 = (x_2, y_2)$ in $\mathbb{R}^2$. Clearly an equidistant point exists on the line connecting these two points as the mid-point of this line. However, an infinite number of equidistant points exists in $\mathbb{R}^2$ which lie on a line orthogonal to the connecting line, passing through the mid-point (see figure 13). This line is defined as the *orthogonal bisector* of the connecting line. In Euclidean space a line is a minimum distance connection of any two points. In differential geometry which

applies to more general, non-Euclidean spaces, this minimum distance connection is defined as a *geodesic*. Similarly, in spaces of PDFs the notion of minimum distance connection (*geodesic*) and orthogonal bisector can be defined relative to KL divergence [24]. Figure 14 shows an example with two univariate Gaussian PDFs defined in Gaussian parameter space (μ, σ^2) , i.e. mean and variance. The *geodesic* connecting the two PDFs is defined by the EEF, and the *orthogonal bisector* is the set of PDFs which are equi-distant with respect to KL divergence from the component densities (as for the Chernoff point).

Figure 13: Orthogonal bisection of the geodesic (line) connecting two points in Euclidean space with $d(\cdot, \cdot)$ the Euclidean distance. Note that all points on the orthogonal bisector are circumcenters of circumcircles through the two points. The smallest such circumcircle has a center point which lies on the connecting line.

Definition 4. *Orthogonal bisectors of PDFs* Consider two PDFs $p(\mathbf{x}; \boldsymbol{\theta}_0), p(\mathbf{x}; \boldsymbol{\theta}_1)$ as distinct points on a smooth statistical manifold $\mathcal{M} = \{p(\mathbf{x}, \boldsymbol{\theta}) | \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ [9] where $\boldsymbol{\theta}$ is an $n_\theta \times 1$ parameter vector. We define $\mathcal{P}_\perp \subset \mathcal{M}$ as the set of all PDFs equidistant in KL divergence with respect to this pair, and denote it as the orthog-

onal bisector of the pair. That is,

$$\begin{aligned}\mathcal{P}_\perp &= \{p(\mathbf{x}; \boldsymbol{\theta}) \in \mathcal{M} : \\ &D(p(\mathbf{x}; \boldsymbol{\theta}) \| p(\mathbf{x}; \boldsymbol{\theta}_0)) = D(p(\mathbf{x}; \boldsymbol{\theta}) \| p(\mathbf{x}; \boldsymbol{\theta}_1))\}\end{aligned}$$

and is equivalent to

$$\mathcal{P}_\perp = \{p(\mathbf{x}; \boldsymbol{\theta}) \in \mathcal{M} : E_{p(\mathbf{x}; \boldsymbol{\theta})} \left[\ln \frac{p(\mathbf{x}; \boldsymbol{\theta}_1)}{p(\mathbf{x}; \boldsymbol{\theta}_0)} \right] = 0\} \quad (92)$$

Note that due to the directional nature of KL divergence, the equi-distant conditions necessary to the orthogonality definition only apply in a single direction (ordering of PDFs in the functional argument of the KL divergence). The particular order of interest in this work is the “left-sided” KL divergence which has been shown to be the relevant divergence corresponding to the EEF [18, 30]. It is also noteworthy to recognize that the EEF being a *geodesic* in this space, has nothing to do with the line being straight in parameter space, but that the EEF is a solution of a particular partial differential equation in these univariate Gaussian coordinates [24]. Fortuitously, for this parametrization it just happens to be a line as seen in Figure 14 (see Appendix 2C). This is no longer true for (μ, σ) Gaussian coordinates (mean, standard deviation) for which it is known to be a hyperbola [31].

To better understand the geometry of the EEF consisting of more than two component PDFs, in particular the extended Chernoff point, $p\boldsymbol{\eta}^*$, we consider some additional simple facts of two dimensional Euclidean spaces. In Euclidean geometry the *circumscribed circle* (CC) of a polygon with center point $c^* \in \mathbb{R}^2$ (circumcenter), is a circle which passes through all the vertices of the polygon. A polygon which has a CC is defined as *cyclic*. Most regular (planar) shapes such as triangles, squares, rectangles, etc. are *cyclic*. Considering the triangle, the CC exists uniquely (for three points in a plane) when the points are in general position,

Figure 14: Orthogonal bisection of the geodesic (line) connecting two univariate normal PDFs drawn in PDF parameter space with $D(\cdot||\cdot)$ the KL divergence.

for example if three points in a plane make a valid triangle. However, in $\mathbb{R}^n$, where the CC becomes more generally a *circumscribed ball*, (CB), the CB will exist for the same conditions, but will not necessarily be unique. For example if $n > M - 1$ where M is the number of points, then there are an infinite number of CBs in the space. When all points are not in general position, then a related geometric object called the *smallest enclosing circle* SEC (in $\mathbb{R}^2$), or *smallest enclosing ball* (SEB) for arbitrary dimension, can be a useful approximation and can be shown to always exist. In many cases the SEC coincides with the CC. In general the SEC is defined as a circle which minimally encloses a set of points, where the circle may pass through some arbitrary subset of the overall set.

Focusing on planar geometry, we can consider these concepts in three point problems, where at least one point is not collinear (i.e. all three points are not on a line). In these three point problems we see that the properties of the triangle defined by the points reveal the relationship between the CC and the SEC. Figure

Figure 15: The depiction of the unique circumcenter, O , of the circumscribed circle of three points in Euclidean space. The circumcenter point exists as the intersection of the orthogonal bisectors of each side of the triangle (i.e. for points in general position). For an acute triangle (left) the circumcenter falls within the triangle. For a right triangle (center), the circumcenter falls on the mid-point of the longest side. For an obtuse triangle (right), the circumcenter falls outside the triangle. In the first two cases the circumcircle coincides with the smallest enclosing circle. For an obtuse triangle they do not coincide as the smallest enclosing circle is centered as in the right triangle, on the mid-point of the longest sided (hypotenuse).

15 shows constructively how the center of the CC is found as the intersection of the orthogonal bisectors of each side of a triangle. These pictures show that for *acute* triangles the center point of the CC exists within the triangle, for *right* triangles it exists on the mid-point of the *hypotenuse* (longest side), and for an *obtuse* triangle the CC exists outside the triangle, on the orthogonal bisector of the hypotenuse. In addition, for acute and right triangles the center of the CC and SEC coincide, but for obtuse triangles the center of the SEC is at the mid-point of the hypotenuse, and thus does not coincide with the CC. These conditions on the CC are necessary and sufficient for characterizing the type of a triangle in Euclidean space, and are defined solely by the orthogonal bisections. This leads us to question for three PDFs what defines acute, right, and obtuse with relation to similar geometric objects in information space. Mathematical definitions for these geometric notions follow, for both Euclidean points and their PDF counterparts. We define geometrical concepts of CC and SEC (CB and SEB in n -dimensional parameter space) in parameter spaces of PDFs where KL divergence is the “distance” and the EEF is

the pairwise *geodesic* connection.

Definition 5. *Circumcenter of circumscribed ball in $\mathbb{R}^n$*

Consider n_p points $\mathbf{p}_i \in \mathbb{R}^n$, $i = 0, 1, \dots, n_p - 1$ in general position within n -dimensional Euclidean space. A point $\mathbf{c}^* \in \mathbb{R}^n$ is the center of a circumscribed ball if

$$d(\mathbf{c}^*, \mathbf{p}_0) = d(\mathbf{c}^*, \mathbf{p}_1) = \dots = d(\mathbf{c}^*, \mathbf{p}_{n_p-1}) = r^* \quad (93)$$

where r^* is the radius of the circumscribed ball and $d(\cdot, \cdot)$ is the Euclidean distance. Note for $\mathbf{p}_i \in \mathbb{R}^2$, $i = 0, 1, 2$, this becomes the circumscribed circle (CC).

Definition 6. *Smallest enclosing ball in $\mathbb{R}^n$*

Consider n_p points $\mathbf{p}_i \in \mathbb{R}^n$, $i = 1, \dots, n_p$ in n -dimensional Euclidean space. A point $\mathbf{s}^* \in \mathbb{R}^n$ is the center point of an enclosing ball with radius r_s^* , (denotes as the pair $(\mathbf{s}^*, r_s^*)$) if

$$\forall_{i=1}^{n_p} d(\mathbf{s}, \mathbf{p}_i) \leq r$$

where the radius r is given as $r = \max_i d(\mathbf{s}, \mathbf{p}_i)$. The smallest enclosing ball $(\mathbf{s}^*, r^*)$ is then

$$\mathbf{s}^* = \arg \min_{\mathbf{s} \in \mathbb{R}^n} \max_{i=1, \dots, n_p} d(\mathbf{s}, \mathbf{p}_i) \quad (94)$$

$$r_s^* = \max_{i=1, \dots, n_p} d(\mathbf{s}^*, \mathbf{p}_i) \quad (95)$$

Note for $\mathbf{p}_i \in \mathbb{R}^2$ this becomes the smallest enclosing circle (SEC).

In spaces of PDFs where the proper distance measures are not generally Euclidean these geometric concepts can be important in data fusion, multiple hypothesis testing, and machine learning frameworks. Thus the above definitions can be re-written in PDF space with respect to a directional divergence, namely the aforementioned left-sided KL divergence. For simplicity we assume as in [9] that we are always working with parameterized PDFs and we restrict ourselves to the sub-manifold of PDF space defined by the EEF.

Definition 7. *Circumcenter of PDFs within the EEF*

Consider n_p PDFs $p(\mathbf{x}; \boldsymbol{\theta}_i)$ as points on a smooth statistical manifold $\mathcal{M} = \{p(\mathbf{x}, \boldsymbol{\theta}) | \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$. The EEF of these component PDFs $p(\mathbf{x}; \boldsymbol{\eta})$ as in definition 1 defines a sub-manifold $\mathcal{M}_{\mathcal{Y}_a} \subset \mathcal{M}$. A PDF (point on the manifold) restricted to this connection $p(\mathbf{x}; \boldsymbol{\theta}(\boldsymbol{\eta}^*))$ is the unique circumcenter of PDFs if

$$\begin{aligned} D(p(\mathbf{x}; \boldsymbol{\theta}(\boldsymbol{\eta}^*)) || p(\mathbf{x}; \boldsymbol{\theta}_0)) &= D(p(\mathbf{x}; \boldsymbol{\theta}(\boldsymbol{\eta}^*)) || p(\mathbf{x}; \boldsymbol{\theta}_1)) \\ &= \dots \\ &= D(p(\mathbf{x}; \boldsymbol{\theta}(\boldsymbol{\eta}^*)) || p(\mathbf{x}; \boldsymbol{\theta}_{n_p-1})) \end{aligned} \tag{96}$$

where $\boldsymbol{\theta}(\boldsymbol{\eta})$ is a restricted subset of $\boldsymbol{\Theta}$ constrained by $\boldsymbol{\eta} \in \mathcal{Y}_a$.

A circumcenter of PDFs, like its Euclidean counterpart, does not always exist or can have infinite solutions. In the Euclidean case a unique circumcenter can be found by setting additional restrictions on the solution, such as “minimum norm” or “minimum distance”. In [32], the authors defined an I-center (Information center) of PDFs which is essentially a circumcenter and found that restriction to the EEF was essentially the I-center of minimum distance. Note that the circumcenter of PDFs within the EEF is exactly the extended Chernoff point.

Definition 8. *Smallest enclosing ball of PDFs*

Consider n_p PDFs $p(\mathbf{x}; \boldsymbol{\theta}_i)$ as points on a smooth statistical manifold $\mathcal{M} = \{p(\mathbf{x}, \boldsymbol{\theta}) | \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ [9]. The EEF of these component PDFs $p(\mathbf{x}; \boldsymbol{\eta})$ as in definition 1 defines a sub-manifold $\mathcal{M}_{\mathcal{Y}_a} \subset \mathcal{M}$. A PDF (point on the manifold) restricted to this connection $p(\mathbf{x}; \boldsymbol{\eta}^*)$ is at the center point of the smallest enclosing ball (SEB) if

$$\boldsymbol{\eta}^* = \arg \min_{\boldsymbol{\theta}(\boldsymbol{\eta}) \in \boldsymbol{\Theta}} \max_{i=0, \dots, n_p-1} D(p(\mathbf{x}; \boldsymbol{\theta}(\boldsymbol{\eta})) || p(\mathbf{x}; \boldsymbol{\theta}_i)) \tag{97}$$

$$r^* = \max_{i=0, \dots, n_p-1} D(p(\mathbf{x}; \boldsymbol{\theta}(\boldsymbol{\eta}^*)) || p(\mathbf{x}; \boldsymbol{\theta}_i)) \tag{98}$$

where $\boldsymbol{\theta}(\boldsymbol{\eta})$ is a restricted subset of $\boldsymbol{\Theta}$ constrained by $\boldsymbol{\eta} \in \mathcal{Y}_p \subset \mathcal{Y}_a$.

The smallest enclosing ball of PDFs has been considered more generally by Nielsen and Nock [33] with respect to a larger class of divergences (i.e. Bregman divergence class). In [34] it is shown that the solution of the minimum ball problem is the generalized mean associated with the proper divergence. Specifically with respect to the left-sided KL divergence, the SEB of PDFs lies within the geometric mean of PDFs, which is essentially the EEF with the weights restricted to be positive and to sum to one, i.e. $\boldsymbol{\eta} \in \mathcal{T}_p$. In section 2.5 it will be evident that if $\boldsymbol{\eta}^*$ is in $\mathcal{T}_p$, the extended Chernoff point and SEB coincide (i.e. are the same PDF) as a simple corollary of theorem 4. However, when the extended Chernoff point is in $\mathcal{T}_a \setminus \mathcal{T}_p$ (i.e. has at least one negative embedding weight), then these are two distinct points in PDF space. Understanding the conditions for which this separation occurs can be important a number of applications, in particular data fusion.

2.4 EEF of Gaussian PDFs

In this section we derive the closed form solutions for orthogonal bisectors of Gaussians as well as for the EEF of M Gaussians, as a tool to illustrate the above geometric properties through simple analytical and numerical examples.

2.4.1 Orthogonal bisectors of Gaussian PDFs

As described above, orthogonal bisectors exist for pairs of PDFs with respect to KL divergence just as they do for pairs of points in Euclidean space with respect to Euclidean distance. To find the orthogonal bisector of PDFs $p(\mathbf{x}; \boldsymbol{\theta}_0)$ and $p(\mathbf{x}; \boldsymbol{\theta}_1)$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\mathbf{x} \in \mathbb{R}^n$, we must find the family of PDFs $\mathcal{P}$ which (from definition 4) meet the condition $\forall_{p(\mathbf{x}; \boldsymbol{\theta}) \in \mathcal{P}} D(p(\mathbf{x}; \boldsymbol{\theta}) \| p(\mathbf{x}; \boldsymbol{\theta}_0)) = D(p(\mathbf{x}; \boldsymbol{\theta}) \| p(\mathbf{x}; \boldsymbol{\theta}_1))$. Since the EEF is a multi-dimensional geodesic connection of PDFs with the additional closure property for exponential families, it is instructive to construct exam-

ples using component PDFs that are in the exponential family. Gaussian PDFs are one such example. Let $p(\mathbf{x}; \boldsymbol{\theta}_0) := \mathcal{N}(\boldsymbol{\mu}_0, \mathbf{C}_0)$, $p(\mathbf{x}; \boldsymbol{\theta}_1) := \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{C}_1)$, then the orthogonal bisector of PDFs can be shown to be the set of PDFs $p(\mathbf{x}; \boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$ such that

$$\begin{aligned} \Delta(\mathcal{P}) &= \ln \frac{|\mathbf{C}_1|}{|\mathbf{C}_0|} + \text{tr}\{\mathbf{C}(\mathbf{C}_1^{-1} - \mathbf{C}_0^{-1})\} \\ &\quad + (\boldsymbol{\mu} - \boldsymbol{\mu}_1)^T \mathbf{C}_1^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_1) \\ &\quad - (\boldsymbol{\mu} - \boldsymbol{\mu}_0)^T \mathbf{C}_0^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0) = 0 \end{aligned} \quad (99)$$

where $\Delta(\mathcal{P}) = D(p(\mathbf{x}; \boldsymbol{\theta}) \| p(\mathbf{x}; \boldsymbol{\theta}_1)) - D(p(\mathbf{x}; \boldsymbol{\theta}) \| p(\mathbf{x}; \boldsymbol{\theta}_0))$. Thus the orthogonal bisector of $p(\mathbf{x}; \boldsymbol{\theta}_0)$ and $p(\mathbf{x}; \boldsymbol{\theta}_1)$ is the family of PDFs corresponding to the parameters, $(\boldsymbol{\mu}, \mathbf{C})$ (i.e. $p(\mathbf{x}; \boldsymbol{\mu}, \mathbf{C})$), which satisfy (99). With p_0 and p_1 taken as univariate Gaussian (99) becomes

$$\begin{aligned} \Delta(\mathcal{P}) &= \log \frac{\sigma_1^2}{\sigma_0^2} + \left[\frac{\sigma_0^2 - \sigma_1^2}{\sigma_1^2 \sigma_0^2} \right] \sigma^2 \\ &\quad + \frac{(\mu - \mu_1)^2}{\sigma_1^2} - \frac{(\mu - \mu_0)^2}{\sigma_0^2} = 0 \end{aligned} \quad (100)$$

shown in Figure 14 for $\mu_0 = 1, \sigma_0^2 = 3, \mu_1 = 2.6, \sigma_1^2 = 1$.

2.4.2 EEF for M Gaussians

The functional form for the EEF of M Gaussian PDFs has been reported in numerous papers but to be complete we state the theorem and prove it in the Appendix.

Theorem 3. *Consider the EEF $p(\mathbf{x}; \boldsymbol{\eta})$ of M Gaussian component density functions $p_i(\mathbf{x}) := \mathcal{N}(\boldsymbol{\mu}_i, \mathbf{C}_i)$, for $i = 0, 1, \dots, M-1$, where $\boldsymbol{\mu}_i$ are $N \times 1$ mean vectors, $\mathbf{C}_i$ are $N \times N$ covariance matrices (positive definite symmetric), and $\boldsymbol{\eta} \in \mathcal{Y}_a$. Then $p(\mathbf{x}; \boldsymbol{\eta})$ is also a Gaussian PDF with mean $\boldsymbol{\mu}_{\boldsymbol{\eta}}$ and covariance $\mathbf{C}_{\boldsymbol{\eta}}$ given as*

$$\boldsymbol{\mu}_{\boldsymbol{\eta}} = \left(\sum_{i=0}^{M-1} \eta_i \mathbf{C}_i^{-1} \right)^{-1} \left(\sum_{i=0}^{M-1} \eta_i \mathbf{C}_i^{-1} \boldsymbol{\mu}_i \right) \quad (101)$$

$$\mathbf{C}_{\boldsymbol{\eta}} = \left(\sum_{i=0}^{M-1} \eta_i \mathbf{C}_i^{-1} \right)^{-1} \quad (102)$$

where $\boldsymbol{\eta} \in \mathcal{Y}_a$ is equivalent to the set of all $\boldsymbol{\eta}$ such that $\mathbf{C}_{\boldsymbol{\eta}}$ is positive definite due to the finite condition on the CGF.

proof: (see Appendix 2A).

In addition to the above analytical form of the EEF of M Gaussian components, one can also derive the following closed form of the CGF.

$$\mathcal{K}_0(\boldsymbol{\eta}; M) = \ln \mathcal{G}(\boldsymbol{\eta}; M) - \frac{1}{2} \sum_{j=1}^{M-1} \mathcal{D}_j(\boldsymbol{\eta}) + \frac{1}{2} \ln |\mathbf{C}_{\boldsymbol{\eta}}|. \quad (103)$$

where for $j = 1, 2, \dots, M-1$ and letting $\mathbf{A}_i = \eta_i (\mathbf{C}_i)^{-1}$, $\mathbf{c}_0 = \boldsymbol{\mu}_0$, we have

$$\mathcal{D}_j(\boldsymbol{\eta}) = (\boldsymbol{\mu}_j - \mathbf{c}_{j-1})^T \mathbf{M}_{ij} (\boldsymbol{\mu}_j - \mathbf{c}_{j-1})$$

with $\mathbf{M}_{ij} = \mathbf{A}_j \left(\sum_{i=0}^j \mathbf{A}_i \right)^{-1} \left(\sum_{i=0}^{j-1} \mathbf{A}_i \right)$, $\mathbf{c}_j = \left(\sum_{i=0}^j \mathbf{A}_i \right)^{-1} \left[\sum_{i=0}^j \mathbf{A}_i \boldsymbol{\mu}_i \right]$, and $\mathcal{G}(\boldsymbol{\eta}; M) = \prod_{i=0}^{M-1} \left(\frac{1}{|\mathbf{C}_i|^{1/2}} \right)^{\eta_i}$. As the CGF of an exponential family this function is known to be convex [22] for all $\boldsymbol{\eta} \in \mathcal{Y}_a \subset \mathcal{Y}$.

2.4.3 Examples: EEF of two Gaussian components

Given the general form of the EEF of Gaussians (101), (102), and (103) we can consider several examples for studying the relationship between the Chernoff point and the SEB of PDFs with respect to KL divergence. Since Gaussian distributions are within the exponential family, from [35, 6] we know that the Chernoff point can be found in closed form for a 1-parameter family. From the perspective of the EEF this means that this condition should hold when the component PDFs only differ in one parameter. The following example illustrates this point.

Example 4.1: *EEF of two Gaussian components with the same mean*

Take $p_0(\mathbf{x}) := \mathcal{N}(\mathbf{0}, \sigma_0^2 \mathbf{I})$, $p_1(\mathbf{x}) := \mathcal{N}(\mathbf{0}, \sigma_1^2 \mathbf{I})$, with $\mathbf{x} \in \mathbb{R}^n$. Note that to find the Chernoff point (I-center) we must perform the minimization as in (77), but

this is equivalent to the minimization of the CGF for two components, from (103) given as

$$\mathcal{K}_0(\eta; 2) = \log \mathcal{G}(\eta; 2) - \frac{1}{2} \mathcal{D}_1(\eta) + \frac{n}{2} \log(2\pi) + \frac{1}{2} |\mathbf{C}_\eta| \quad (104)$$

or we must solve for η corresponding to $\mathcal{K}_0'(\eta) = 0$, since $\mathcal{K}_0(\eta)$ is convex. By substitution of $p_0(\mathbf{x}), p_1(\mathbf{x})$ into (104) and differentiating with respect to η we obtain

$$\mathcal{K}_0'(\eta; 2) = n \log \frac{\sigma_0}{\sigma_1} - \frac{n}{2} \frac{(\sigma_0^2 - \sigma_1^2)}{\eta(\sigma_0^2 - \sigma_1^2) + \sigma_1^2} = 0 \quad (105)$$

Solving this for η we get

$$\eta^* = \frac{\sigma_1^2}{\sigma_1^2 - \sigma_0^2} - \frac{1}{\log \frac{\sigma_1^2}{\sigma_0^2}} \quad (106)$$

To obtain the PDF corresponding to the Chernoff point $p(\mathbf{x}; \eta^*) = \mathcal{N}(\mathbf{0}, \mathbf{C}_{\eta^*})$ we substitute back into the form of $\mathbf{C}_\eta$ yielding

$$\mathbf{C}_{\eta^*} = \frac{\sigma_1^2 \sigma_0^2 \log \frac{\sigma_1^2}{\sigma_0^2}}{\sigma_1^2 - \sigma_0^2} \mathbf{I} \quad (107)$$

As more parameters differ, the problem quickly becomes analytically intractable. As a result in [6] a bisection algorithm is introduced to approximate the Chernoff point along the geodesic. However, since the EEF is a member of the exponential family over the restricted interval $0 \leq \eta \leq 1$, the function $\mathcal{K}_0(\eta)$ is convex over this interval [22], thus a unique solution always exists on the geodesic, and can be found exactly using non-linear programming [36]. That being said in low dimensional cases we may find solutions quickly through evaluation of multiple solutions and choosing the solution which lies on the geodesic.

Example 4.2: *Two Gaussians with different means and variances*

Take $p_0(\mathbf{x}) := \mathcal{N}(\mu_0 \mathbf{1}, \sigma_0^2 \mathbf{I})$, $p_1(\mathbf{x}) := \mathcal{N}(\mu_1 \mathbf{1}, \sigma_1^2 \mathbf{I})$, with $\mathbf{x} \in \mathbb{R}^n$. In this example we have two component PDF parameters that are different. After substitution for

the PDFs and differentiating with respect to η we have

$$\begin{aligned} \mathcal{K}_0'(\eta; 2) = & n \log \frac{\sigma_0}{\sigma_1} - \frac{n}{2} \frac{(\sigma_0^2 - \sigma_1^2)}{\eta(\sigma_0^2 - \sigma_1^2) + \sigma_1^2} + \\ & \frac{n(\mu_1 - \mu_0)^2}{2} \frac{(\sigma_0^2 - \sigma_1^2)\eta^2 + \sigma_1^2(2\eta - 1)}{[(\sigma_0^2 - \sigma_1^2)\eta + \sigma_1^2]^2} \end{aligned} \quad (108)$$

which can be written as a quadratic in η as $a\eta^2 + b\eta + c = 0$ where

$$\begin{aligned} a &= (\sigma_0^2 - \sigma_1^2) \left[(\sigma_0^2 - \sigma_1^2) \log \frac{\sigma_0^2}{\sigma_1^2} + (\mu_1 - \mu_0)^2 \right] \\ b &= 2\sigma_1^2 \left[(\mu_1 - \mu_0)^2 + (\sigma_0^2 - \sigma_1^2) \log \frac{\sigma_0^2}{\sigma_1^2} \right] - (\sigma_0^2 - \sigma_1^2)^2 \end{aligned}$$

and

$$c = \sigma_1^2 \left[\sigma_1^2 \log \frac{\sigma_0^2}{\sigma_1^2} - (\sigma_0^2 - \sigma_1^2) - (\mu_1 - \mu_0)^2 \right]$$

which can be solved by the quadratic equation and picking the solution which is on the geodesic. Note that if we let $\sigma_1^2 = \sigma_0^2$, then $\eta^* = \frac{1}{2}$ as the KL-divergence reduces to squared Euclidean distance.

2.5 Characterization of Three Component EEF

As observed in section 2.3 and 2.4 the extended Chernoff point is analogous to a CC, and the center of the SEB of PDFs is analogous to the SEC in plane geometry. With that in mind we consider the three component EEF as a “triangle” of PDFs. Our motivation is that if in fact these analogies hold true, then the extended Chernoff point will have a similar relationship to the SEB of PDFs that the CC shares with the SEC. In particular, the type of the triangle is bijective with the location of the CC relative to the triangle, and subsequently determines whether the CC coincides with the SEC for a given triangle. Determining this separation in the PDF space is important to calculating and utilizing these geometrical criteria in data fusion. In the following definitions we impose the common triangle characterizations of acute, right, and obtuse into the three component EEF based

solely on the form of the embedding components associated with the extended Chernoff point. Ultimately we will utilize these characterizations for separating the problems of finding the extended Chernoff point and the SEB, respectively.

In definition 1 we defined the M -component EEF under its most general conditions, i.e. relative to embedding parameters in which the CGF is finite. We will define the three component EEF corresponding to (81) for $M = 3$ and $\gamma \in \gamma_a$. Then by imposing the sum constraint (and ignoring the dummy variable) we write the three component EEF as

$$p_{\boldsymbol{\eta}} = \exp \left(\eta_1 \ln \frac{p_1}{p_0} + \eta_2 \ln \frac{p_2}{p_0} + \ln p_0 - \mathcal{K}_0(\eta_1, \eta_2) \right) \quad (109)$$

Further we note that this three component EEF can be thought of as a “triangle” of PDFs with “sides” comprised of the *pairwise connections* of its three components.

Definition 9. *Pairwise connections of three-component EEF: Triangle of PDFs*
Consider the three component EEF $p_{\boldsymbol{\eta}}$ (equation (109)). We define the pairwise connections p_{01} , p_{02} , and p_{12} as sub-manifolds of $p_{\boldsymbol{\eta}}$ written as

$$\begin{aligned} p_{01} &= \frac{1}{c_{01}} (p_1^{\eta_{01}} p_0^{1-\eta_{01}}) \\ p_{02} &= \frac{1}{c_{02}} (p_2^{\eta_{02}} p_0^{1-\eta_{02}}) \\ p_{12} &= \frac{1}{c_{12}} (p_2^{\eta_{12}} p_1^{1-\eta_{12}}) \end{aligned} \quad (110)$$

where $\eta_{ij} \in [0, 1]$. Further we denote $p_{\eta_{01}^*}$, $p_{\eta_{02}^*}$, and $p_{\eta_{12}^*}$ as the pairwise Chernoff points of each pairwise connection, respectively. That is, they represent the PDFs (within each pairwise connection) corresponding to the conditions

$$\begin{aligned} D_{01}^* &= D(p_{\eta_{01}^*} \| p_0) = D(p_{\eta_{01}^*} \| p_1) \\ D_{02}^* &= D(p_{\eta_{02}^*} \| p_0) = D(p_{\eta_{02}^*} \| p_2) \\ D_{12}^* &= D(p_{\eta_{12}^*} \| p_1) = D(p_{\eta_{12}^*} \| p_2) \end{aligned} \quad (111)$$

where we define D_{01}^* , D_{02}^* , and D_{12}^* as the pairwise Chernoff distance measures (“lengths”) of each connection (“side”).

Note that each pairwise connection corresponds to the three component EEF $p_{\boldsymbol{\eta}}$ with the embedding weight of the opposite component set to zero, e.g. $p_{01} = p_{\boldsymbol{\eta}}$, with $\boldsymbol{\eta} = [1 - \eta_{01}, \eta_{01}, 0]^T$. In addition, for each *pairwise connection*, we can also refine our previous definition of an orthogonal bisector (definition 4) to be restricted to the three component EEF, $p_{\boldsymbol{\eta}}$.

Definition 10. *Orthogonal bisector as a sub-manifold of the three component EEF*
We define

$$p_{\boldsymbol{\eta}(t)} = \{p \in p_{\boldsymbol{\eta}} : E_p \left[\ln \frac{p_1}{p_0} \right] = 0\} \quad (112)$$

as the sub-manifold of the three component EEF which orthogonally bisects the pairwise connection p_{01} , where $t \in \mathbb{R}$ and $\boldsymbol{\eta}(t) = [\eta_1(t) \ \eta_2(t)]^T$.

Note that $p_{\boldsymbol{\eta}(t)}$ is a curve in the PDF space parameterized by t .

Definition 11. *Orthogonal projection onto $p_{\boldsymbol{\eta}(t)}$*

We say that $p_{2\perp} \in p_{\boldsymbol{\eta}(t)}$ is the orthogonal projection of PDF p_2 onto $p_{\boldsymbol{\eta}(t)}$ when [9, 37]

$$p_{2\perp} = \arg \min_{p \in p_{\boldsymbol{\eta}(t)}} D(p || p_2) \quad (113)$$

We show in Appendix 2B, Fact 1, that for the three component EEF $p_{2\perp}$ is given as

$$p_{2\perp} = \exp \left(\eta_1 \ln \frac{p_1}{p_0} - \mathcal{K}_0(\eta_1, 1) + \ln p_2 \right) \quad (114)$$

which means that the closest PDF within the orthogonal bisector $p_{\boldsymbol{\eta}(t)}$ of a given pairwise connection p_{01} , to the opposite PDF component e.g. $p_2 \in p_{\boldsymbol{\eta}}$ (in the three component EEF), corresponds to a value of one for the opposite component PDF (i.e. $\eta_2 = 1$). Given these definitions of orthogonal bisection and projection in

the PDF space, we can now define characterizations of the three component EEF analogous to that of triangles in Euclidean space.

Definition 12. *Acute connection of PDFs* The three component EEF $p_{\boldsymbol{\eta}}$ is an “acute” connection if the intersection of the orthogonal bisectors (of each pairwise connection) occurs within the space of PDFs defined by the subset of the $p_{\boldsymbol{\eta}}$ for which $\eta_i \geq 0$ for all i , i.e. for $\boldsymbol{\eta} \in \Upsilon_p$.

Figure 16: *Geometry of PDFs.* An “acute” connection of three univariate Gaussian PDFs. The lines connecting the PDF points (asterisks) are *geodesics* through the exponential connection in Gaussian parameter space. The orthogonal bisection of each *geodesic* is shown with the matching color / linetype. The triangles denote the pairwise Chernoff points on the corresponding *geodesic* connections of each PDF pair. This “triangle” of Gaussians is defined as “acute” since the intersection of the orthogonal bisectors occurs within the region of Gaussian parameter space corresponding to positive embedding weights, $\eta_i > 0$ for all i , i.e. the extended Chernoff point.

This condition is a defined extension to PDFs of the Euclidean property of a

circumcenter within a convex hull of points that define an acute triangle, and is illustrated in figure 16 for univariate Gaussian components. This illustration utilizes our previously derived analytical closed forms for the pairwise EEF (101, 102) and the orthogonal bisectors (100). The lines connecting the points in the PDF space are the univariate Gaussian parameter values of the pairwise connections. Note that since the orthogonal bisectors intersect within the region of Gaussian parameter space mapped by the positive embedding weights of the EEF, this connection (from definition 12) is deemed acute. One other feature of note in this figure is the connection of p_0 and p_2 . In this case the component PDFs differ only in the mean. It is well known [7], that KL divergence is essentially a squared Euclidean distance. Thus the orthogonal bisector is simply a line as in Euclidean space.

Definition 13. *Right connection of PDFs* The three component EEF $p_{\boldsymbol{\eta}}$ is a “right” connection (as in figure 17) if the intersection of the orthogonal bisectors (of each pairwise connection) occurs on the mid-point of the “longest” pairwise connection (side) of $p_{\boldsymbol{\eta}}$. That is, if we take p_{02} as the “longest” side ($D_{02}^* > D_{01}^*, D_{12}^*$), then the orthogonal bisectors of the pairwise connections intersect on the boundary denoted as η_{02}^* (the pairwise Chernoff point of the connection of p_0 and p_2) with $\eta_0, \eta_2 > 0$, and $\eta_1 = 0$.

Definition 14. *Obtuse connection of PDFs* The three component EEF $p_{\boldsymbol{\eta}}$ is an “obtuse” connection (as in figure 18) if the intersection of the orthogonal bisectors (of each pairwise connection) occurs at a PDF point $p_{\boldsymbol{\eta}^*}$ such that $\boldsymbol{\eta}^* \notin \Upsilon_p$. That is, the extended Chernoff point (i.e. unique I-center of PDFs corresponding to sum constraint) has one negative embedding weight.

Definition 9 suggests that the pairwise Chernoff distance acts as a “squared length” of each side of a triangle of PDFs. In figure 17 we observe that by increasing

Figure 17: *Geometry of PDFs*. A “right” connection of three univariate Gaussian PDFs. The lines connecting the PDF points (asterisks) are *geodesics* through the exponential connection in Gaussian parameter space. The orthogonal bisection of each *geodesic* is shown with the matching color / linetype. The triangles denote the pairwise Chernoff points on the *geodesic* connections of each PDF pair. This “triangle” of Gaussians is defined as “right” due to the intersection of the bisectors (extended Chernoff point denoted by circle) occurring at the pairwise Chernoff point $p_{\eta_{02}^*}$ of the “longest” side, i.e. $D_{02}^* > D_{01}^*, D_{12}^*$.

the mean of p_2 (relative to the acute case) we have increased the squared length of the connection of p_0 and p_2 (and thus D_{02}^*) achieving the condition of a right connection. Similarly, for an additional increase in length we achieve the obtuse condition shown in figure 18. Note that in all three cases $D_{02}^* > D_{01}^*, D_{12}^*$ therefore making side p_{02} akin to an “hypotenuse” of the PDF triangles in these examples. Although these analogous triangle characterizations seem somewhat obscure, they are key to understanding when the SEB and extended Chernoff point coincide (acute and right), and when they do not (obtuse). In addition, we will show that this separation can be determined without solving for the extended Chernoff point directly. This fact may be quite useful in practice when the extended Chernoff point is extremely difficult to solve for (find) compared to the SEB. The following

Figure 18: *Geometry of PDFs*. An “obtuse” connection of three univariate Gaussian PDFs. The lines connecting the PDF points (asterisks) are *geodesics* through the exponential connection in Gaussian parameter space. The orthogonal bisection of each *geodesic* is shown with the matching color, with the triangles denoting the pairwise Chernoff points on the *geodesics*. This “triangle” of Gaussians is defined as “obtuse” due to the intersection of the bisectors (extended Chernoff point denoted by circle) falling outside the region of Gaussian parameter space corresponding to positive embedding weights. In this case $\boldsymbol{\eta}^*$ has one negative component, $\eta_1 < 0$.

lemmas are needed to establish the needed theorem.

Lemma 1. *Extended Chernoff point relative to pairwise Chernoff point*

If $p_{\boldsymbol{\eta}^*}$ is the extended Chernoff point of the three component EEF, $p_{\boldsymbol{\eta}}$, $\boldsymbol{\eta} \in \Upsilon_a$, with components p_0, p_1, p_2 , then the KL-divergence of the extended Chernoff point to its PDF components is greater than or equal to the KL divergence of each pairwise Chernoff point to its respective component PDFs. For example for pairwise connection p_{01} with corresponding pairwise Chernoff point $p_{\eta_{01}^*}$ we have

$$D(p_{\boldsymbol{\eta}^*} \| p_2) = D(p_{\boldsymbol{\eta}^*} \| p_0) \geq D(p_{\eta_{01}^*} \| p_0) = D(p_{\eta_{01}^*} \| p_1) = D_{01}^* \quad (115)$$

Similarly for pairwise connections p_{02} and p_{12} .

Proof. With no loss in generality, let D_{01}^* be the Chernoff distance (“squared length”) of the longest side. That is $D_{01}^* > D_{12}^*, D_{02}^*$. Since $p_{\boldsymbol{\eta}^*}$ is the extended

Chernoff point, of the EEF of p_0, p_1, p_2 , then $p_{\boldsymbol{\eta}^*}$ lies on the *orthogonal bisector* for all of the pairwise EEFs. In particular, $p_{\boldsymbol{\eta}^*} \in p_{\boldsymbol{\eta}(t)}$, the orthogonal bisector of p_{01} (see definition 10), and from Fact 2 this intersection occurs at $p_{\eta_{01}^*} = [1 - \eta_{01}^* \ \eta_{01}^* \ 0]$. Consequently it can be shown that the following Pythagorean relation holds: [9, 38]

$$\forall_{p \in p_{\boldsymbol{\eta}(t)}} D(p \| p_0) = D(p \| p_{\eta_{01}^*}) + D(p_{\eta_{01}^*} \| p_0) \quad (116)$$

Specifically for the extended Chernoff point we have (see figure 19)

$$D(p_{\boldsymbol{\eta}^*} \| p_0) = D(p_{\boldsymbol{\eta}^*} \| p_{\eta_{01}^*}) + D(p_{\eta_{01}^*} \| p_0) \quad (117)$$

Since $D(p_{\boldsymbol{\eta}^*} \| p_{\eta_{01}^*}) \geq 0$ unless $p_{\boldsymbol{\eta}^*} = p_{\eta_{01}^*}$ almost everywhere, then (116) implies

$$D(p_{\boldsymbol{\eta}^*} \| p_0) \geq D(p_{\eta_{01}^*} \| p_0) \quad (118)$$

and as $p_{\boldsymbol{\eta}^*}$ is the extended Chernoff point then we have

$$D(p_{\boldsymbol{\eta}^*} \| p_2) = D(p_{\boldsymbol{\eta}^*} \| p_0) \geq D(p_{\eta_{01}^*} \| p_0) = D(p_{\eta_{01}^*} \| p_1) \quad (119)$$

□

Note that if lemma 1 holds with equality then $p_{\boldsymbol{\eta}}$ is a “right” connection.

Lemma 2. *Monotonocity property of EEF*

Let $p_{\boldsymbol{\eta}}$ be the three component EEF of PDFs p_0, p_1, p_2 and take $p_{\boldsymbol{\eta}(t)}$ as the orthogonal bisector (PDF family) of the pairwise EEF p_{01} , defined in equation (112). The KL-divergence of this family to the opposite PDF component p_2 , $D(p_{\boldsymbol{\eta}(t)} \| p_2)$ is a decreasing function of η_2 , where η_2 is the embedding parameter associated with PDF component p_2 .

Proof. Let $\boldsymbol{\eta}(t) = [\eta_1(t) \ \eta_2(t) = t]^T$ where $t \in [0, 1]$. Then from facts 1 and 2, $p_{\boldsymbol{\eta}(1)} = p_{2\perp}$ and $p_{\boldsymbol{\eta}(0)} = p_{0\perp}$, respectively. We can write the KL-divergence $D(p_{\boldsymbol{\eta}(t)} \| p_2)$ as

$$D(p_{\boldsymbol{\eta}(t)} \| p_2) = E_{p_{\boldsymbol{\eta}(t)}} \left[\eta_1(t) \ln \frac{p_1}{p_0} + (\eta_2(t) - 1) \ln \frac{p_2}{p_0} - \mathcal{K}_0(\boldsymbol{\eta}(t)) \right]$$

Figure 19: Pythagorean relation of information geometry applied in proof to lemma 1.

but by definition 10, $E_{p\boldsymbol{\eta}(t)} \left[\ln \frac{p_1}{p_0} \right] = 0$ thus

$$\begin{aligned} D(p_{\boldsymbol{\eta}(t)} \| p_2) &= E_{p\boldsymbol{\eta}(t)} \left[(\eta_2(t) - 1) \ln \frac{p_2}{p_0} - \mathcal{K}_0(\boldsymbol{\eta}(t)) \right] \\ &= (t - 1) E_{p\boldsymbol{\eta}(t)} \left[\ln \frac{p_2}{p_0} \right] - \mathcal{K}_0(\eta_1(t), t) \end{aligned} \quad (120)$$

Now we must calculate the total derivative of $D(p_{\boldsymbol{\eta}(t)} \| p_2)$ as

$$\frac{dD(\eta_1(t), \eta_2(t))}{dt} = \frac{\partial D}{\partial \eta_1} \frac{\partial \eta_1}{\partial t} + \frac{\partial D}{\partial \eta_2} \frac{\partial \eta_2}{\partial t} \quad (121)$$

where

$$\frac{\partial D}{\partial \eta_1} = (t - 1) \frac{\partial}{\partial \eta_1} E_{p\boldsymbol{\eta}(t)} \left[\ln \frac{p_2}{p_0} \right] - \frac{\partial \mathcal{K}_0}{\partial \eta_1} \quad (122)$$

$$\frac{\partial D}{\partial \eta_2} = (t - 1) \frac{\partial}{\partial \eta_2} E_{p\boldsymbol{\eta}(t)} \left[\ln \frac{p_2}{p_0} \right] + E_{p\boldsymbol{\eta}(t)} \left[\ln \frac{p_2}{p_0} \right] - \frac{\partial \mathcal{K}_0}{\partial \eta_2} \quad (123)$$

but since $\frac{\partial \eta_2}{\partial t} = 1$, $\frac{\partial \mathcal{K}_0}{\partial \eta_2} = E_{p\boldsymbol{\eta}(t)} \left[\ln \frac{p_2}{p_0} \right]$, and $\frac{\partial \mathcal{K}_0}{\partial \eta_1} = 0$, then by substitution of (122),

(123) into (121) we get

$$\frac{dD}{dt} = (t - 1) \frac{\partial^2 \mathcal{K}_0}{\partial \eta_1 \partial \eta_2} \cdot \frac{\partial \eta_1}{\partial t} + (t - 1) \frac{\partial^2 \mathcal{K}_0}{\partial \eta_2^2} \quad (124)$$

It remains to solve for $\frac{\partial \eta_1}{\partial t} = \frac{\partial \eta_1}{\partial \eta_2}$. We do this by use of the implicit function theorem. Since we have a function $g(\eta_1, \eta_2) = E_{p\boldsymbol{\eta}_{(t)}} \left[\ln \frac{p_1}{p_0} \right] = 0$ which depends implicitly on η_1 . Let $T_1 = \ln \frac{p_1}{p_0}$ and $T_2 = \ln \frac{p_2}{p_0}$, then

$$\begin{aligned} \frac{dg}{d\eta_2} &= \frac{d}{d\eta_2} \left[\int \exp(\eta_1 T_1 + \eta_2 T_2 + \ln p_0 - \mathcal{K}_0(\eta_1, \eta_2)) T_1 dx \right] = 0 \\ &= E_{p\boldsymbol{\eta}_{(t)}} \left[T_1 \left(\frac{\partial \eta_1}{\partial \eta_2} T_1 + T_2 - \frac{\partial \mathcal{K}_0(\eta_1, \eta_2)}{\partial \eta_2} \right) \right] = 0 \end{aligned} \quad (125)$$

where

$$\frac{\partial \mathcal{K}_0}{\partial \eta_2} = E_{p\boldsymbol{\eta}_{(t)}} \left[\frac{\partial \eta_1}{\partial \eta_2} T_1 + T_2 \right] \quad (126)$$

Substituting (126) into (125) and after some algebra we have

$$\begin{aligned} &\frac{\partial \eta_1}{\partial \eta_2} E_{p\boldsymbol{\eta}_{(t)}} \left[T_1 \left(T_1 - E_{p\boldsymbol{\eta}_{(t)}} [T_1] \right) \right] \\ &+ E_{p\boldsymbol{\eta}_{(t)}} \left[T_1 \left(T_2 - E_{p\boldsymbol{\eta}_{(t)}} [T_2] \right) \right] = 0 \end{aligned} \quad (127)$$

but the first expectation is just the variance of T_1 , while the second is the covariance of T_1 and T_2 . Thus we have

$$\frac{\partial \eta_1}{\partial \eta_2} = \frac{-cov(T_1, T_2)}{var(T_1)} = \frac{\partial \eta_1}{\partial t} \quad (128)$$

Finally substitution of (128) into (124) yields

$$\frac{dD}{dt} = (t-1) var(T_2) \left[1 - \frac{cov^2(T_1, T_2)}{var(T_1)var(T_2)} \right] \quad (129)$$

Now by the Cauchy-Schwartz inequality the bracketed term is greater than zero, as is the variance of T_2 , thus $\frac{dD}{dt} < 0$ for all $t < 1$, and $D(p\boldsymbol{\eta}_{(t)} \| p_2)$ is a decreasing function of η_2 with increasing η_2 . $\square$

The following lemma also holds with the proof omitted due to its similarity to the above proof.

Lemma 3. *The KL-divergence of the orthogonal bisector $p_{\boldsymbol{\eta}(t)}$ of p_{01} to component PDF p_0 (or p_1) is an increasing function of t . That is*

$$\frac{d}{dt} D(p_{\boldsymbol{\eta}(t)} \| p_0) = t \text{var}(T_2) \left(1 - \frac{\text{cov}^2(T_1, T_2)}{\text{var}(T_1) \text{var}(T_2)} \right) > 0 \quad (130)$$

for $\eta_2(t) = t > 0$.

As a consequence of Lemmas 2 and 3 we observe that there is a direction associated with the PDF family $p_{\boldsymbol{\eta}(t)}$ and the component p_2 . That is, as we increase η_2 (up to 1) we move closer to p_2 , and as we decrease η_2 we move further from p_2 . The pairwise connection p_{01} acts as a boundary, where on one side $p_{\boldsymbol{\eta}(t)}$ corresponds to positive η_2 , and on the other to negative η_2 , since $\eta_2 = 0$ at p_{01}^* (fact 2). This notion of direction is useful in the proof of the following theorem which provides a method to characterize a given connection by only finding its pairwise Chernoff points. That is, we do not need to determine the extended Chernoff point to characterize the connection, and as a result we can choose the fusion strategy after determining the connection, which can avoid needless calculations.

Theorem 4. *Characterization of “triangle” of PDFs Let $p_{\boldsymbol{\eta}}$, $\boldsymbol{\eta} \in \mathcal{Y}_a$ be the three component EEF of p_0, p_1, p_2 . Take p_i as an arbitrary component PDF and let $p_{\eta_{jk}^*}$ be the Chernoff point of the pairwise EEF p_{jk} , where $j, k \neq i$. Then $p_{\boldsymbol{\eta}}$ is an acute connection if and only if for all $i \in \{0, 1, 2\}$*

$$D(p_{\boldsymbol{\eta}_{jk}^*} \| p_i) > D(p_{\boldsymbol{\eta}_{jk}^*} \| p_j) = D(p_{\boldsymbol{\eta}_{jk}^*} \| p_k) = D_{jk}^* \quad (131)$$

For example, the KL-divergence of the pairwise Chernoff point of $p_{\eta_{01}^*}$ to p_2 (the opposite PDF component) is greater than the KL-divergence of $p_{\eta_{01}^*}$ to its components p_0, p_1 . This is also true for pairwise Chernoff points $p_{\eta_{02}^*}$, and $p_{\eta_{12}^*}$. (Note that in Euclidean geometry of a triangle (131) is equivalent to (for any side) the median of the triangle being greater than the halflength).

Figure 20: Sketch supporting the necessary part of the proof of theorem 4.

Proof. ($\Rightarrow$) Assume $p_{\boldsymbol{\eta}}$ is *acute*. Then $\boldsymbol{\eta}^* \in \mathcal{Y}_p \subset \mathcal{Y}_a$. That is, $\boldsymbol{\eta}^*$ consists of strictly positive components. By the properties of the extended Chernoff point

$$D(p_{\boldsymbol{\eta}^*} \| p_0) = D(p_{\boldsymbol{\eta}^*} \| p_1) = D(p_{\boldsymbol{\eta}^*} \| p_2)$$

and $p_{\boldsymbol{\eta}^*}$ lies at the intersection of the perpendicular bisectors of each side of the triangle of PDFs. With no loss in generality, take p_{01} as the pairwise EEF of p_0, p_1 to represent an arbitrary side of the triangle and as in (112) we take $p_{\boldsymbol{\eta}(t)}$ as the orthogonal bisector of p_{01} . From Fact 2 p_{01} and $p_{\boldsymbol{\eta}(t)}$ intersect at $p_{\eta_{01}^*}$, the pairwise Chernoff point and from Fact 1 component PDF p_2 projects orthogonally onto $p_{\boldsymbol{\eta}(t)}$ at point $p_{2\perp}$. Thus $p_{2\perp}$, $p_{\eta_{01}^*}$, and $p_{\boldsymbol{\eta}^*}$ are all in $p_{\boldsymbol{\eta}(t)}$ and the embedding parameters associated with each of these PDFs are given as

$$p_{\eta_{01}^*} \Rightarrow (\eta_{01}^*, 0)$$

$$p_{\boldsymbol{\eta}^*} \Rightarrow (\eta_1^*, \eta_2^* < 1)$$

$$p_{2\perp} \Rightarrow (\eta_{1\perp}, 1)$$

Thus η_2 is increasing (along $p\boldsymbol{\eta}_{(t)}$) from $p_{\eta_{01}^*}$, through $p\boldsymbol{\eta}^*$, up to $p_{2\perp}$ (see figure 20). Since from Lemma 2 $D(p\boldsymbol{\eta}_{(t)}\|p_2)$ is a decreasing function, then

$$D(p_{\eta_{01}^*}\|p_{2\perp}) > D(p\boldsymbol{\eta}^*\|p_{2\perp}) \quad (132)$$

Now by the aforementioned Pythagorean relation (116) we have, for all $p \in p\boldsymbol{\eta}_{(t)}$

$$D(p\|p_2) = D(p\|p_{2\perp}) + D(p_{2\perp}\|p_2) \quad (133)$$

Thus

$$D(p\boldsymbol{\eta}^*\|p_2) = D(p\boldsymbol{\eta}^*\|p_{2\perp}) + D(p_{2\perp}\|p_2) \quad (134)$$

$$D(p_{\eta_{01}^*}\|p_2) = D(p_{\eta_{01}^*}\|p_{2\perp}) + D(p_{2\perp}\|p_2) \quad (135)$$

and combining (132), (134), and (135) implies

$$D(p_{\eta_{01}^*}\|p_2) > D(p\boldsymbol{\eta}^*\|p_2) \quad (136)$$

Finally by lemma 1 we have

$$D(p\boldsymbol{\eta}^*\|p_2) > D(p_{\eta_{01}^*}\|p_1) = D(p_{\eta_{01}^*}\|p_0) \quad (137)$$

Thus $D(p_{\eta_{01}^*}\|p_2) > D(p_{\eta_{01}^*}\|p_1) = D(p_{\eta_{01}^*}\|p_0)$ and since we took “side” p_{01} as arbitrary, this holds for all sides and the forward implication is proved.

($\Leftarrow$) We again take p_{01} as an arbitrary side and assume that

$$D(p_{\eta_{01}^*}\|p_2) > D(p_{\eta_{01}^*}\|p_1) = D(p_{\eta_{01}^*}\|p_0) \quad (138)$$

We must prove that $p\boldsymbol{\eta}$ is an acute connection, that is $\boldsymbol{\eta}^* \in \gamma_p$. Since $p_{\eta_{01}^*}$ and $p\boldsymbol{\eta}^*$ are in $p\boldsymbol{\eta}_{(t)}$, then by the Pythagorean relation of information geometry (as in equation (117)),

$$D(p\boldsymbol{\eta}_{(t)}\|p_0) = D(p\boldsymbol{\eta}_{(t)}\|p_{\eta_{01}^*}) + D(p_{\eta_{01}^*}\|p_0) \quad (139)$$

Figure 21: Sketch supporting the sufficient part of the proof of theorem 4.

but from (138) this implies

$$D(p_{\boldsymbol{\eta}(t)} \| p_0) < D(p_{\boldsymbol{\eta}(t)} \| p_{\eta_{01}^*}) + D(p_{\eta_{01}^*} \| p_2) \quad (140)$$

and by the equidistant property of the extended Chernoff point we have

$$D(p_{\boldsymbol{\eta}(t)} \| p_2) < D(p_{\boldsymbol{\eta}(t)} \| p_{\eta_{01}^*}) + D(p_{\eta_{01}^*} \| p_2) \quad (141)$$

at $p_{\boldsymbol{\eta}(t)} = p_{\boldsymbol{\eta}^*}$. Note that $p_{\boldsymbol{\eta}(t)}$ is a function of $\eta_2(t)$ as shown in figure 21 and consider the function

$$G(\boldsymbol{\eta}(t)) = D(p_{\boldsymbol{\eta}(t)} \| p_2) - D(p_{\boldsymbol{\eta}(t)} \| p_{\eta_{01}^*}) - D(p_{\eta_{01}^*} \| p_2) \quad (142)$$

where $G(p_{\eta_{01}^*}) = 0$ and by (141), $G(\boldsymbol{\eta}^*) < 0$. Now if $G(\boldsymbol{\eta}(t))$ is a decreasing function of $\eta_2(t) = t$, then $\eta_2^* > \eta_2(0) = 0$ as we move along $p_{\boldsymbol{\eta}(t)}$ from $p_{\eta_{01}^*}$ to $p_{2\perp}$. Taking the derivative of (142) gives

$$\frac{dG(\boldsymbol{\eta}(t))}{dt} = \frac{d}{dt} D(p_{\boldsymbol{\eta}(t)} \| p_2) - \frac{d}{dt} D(p_{\boldsymbol{\eta}(t)} \| p_{\eta_{01}^*}) \quad (143)$$

but by lemma 2 the first term is less than zero for all $\eta_2(t) = t < 1$. Similarly, lemma 3 implies that the second term is positive for all $t > 0$. Thus G is decreasing and $\eta_2^* > 0$. Since we took side p_{01} as arbitrary, this holds for all sides and we conclude that $p_{\boldsymbol{\eta}}$ is acute. $\square$

This theorem has the following important corollary.

Corollary 1. *If the three component EEF, $p_{\boldsymbol{\eta}}$, is an acute or right connection then the extended Chernoff point and the center of the SEB coincide. If $p_{\boldsymbol{\eta}}$ is an obtuse connection then the radius (the KL divergence of the PDF point to its components) of the extended Chernoff point is strictly larger than that of the SEB, and the center of the SEB occurs at the pairwise Chernoff point of the longest side.*

2.5.1 Numerical examples for finding extended Chernoff point

In figures 16 - 18 we observed the geometry of the extended Chernoff point in the three component EEF of univariate Gaussians, where we identified the extended Chernoff point as the intersection of the orthogonal bisectors of the pairwise connections. In this case a gridding procedure could easily identify the desired result. However, in practice this would be inefficient. A better approach alluded to earlier is to solve the numerical optimization problem using non-linear programming techniques. In these univariate Gaussian examples this is particularly efficient as we have a closed form expression for the objective to be minimized, i.e. the CGF of the M -component PDF (103). For the constraints we have $\boldsymbol{\eta} \in \mathcal{Y}_a$ which for the Gaussian problem becomes

$$\boldsymbol{\eta}^* = \arg \min_{\boldsymbol{\eta}} \mathcal{K}_0(\boldsymbol{\eta}) \quad (144)$$

such that $\sum_{i=0}^{M-1} \eta_i = 1$, and the variance parameter of the EEF (as a function of the embedding weights), $\sigma^2(\boldsymbol{\eta})$, must be strictly greater than zero. This second constraint on the variance of the EEF is equivalent to the CGF being finite for

the univariate PDF. For solving the SEB we add the constraint for the embedding weights to be positive. The addition of this constraint assures that the variance of the EEF will be positive since it can not take values outside of the range in variance of the components (see equation (102)).

Figure 22: The Kullback-Liebler divergence relative to the extended Chernoff point / SEB solution point (PDF) for the “acute” connection of three univariate Gaussian PDFs. The black lines connecting the PDF points are *geodesics* through the exponential connection, and thus make up the boundary of the exponential connection in the normal parameter space. The colors depict the level of the KL divergence of all points (PDFs) relative to the center point (magenta asterisk). The color level is cropped at $D(p_{\boldsymbol{\eta}^*} \| p_i)$ which is drawn as a contour on the image.

To illustrate this approach we produce three examples in which we numerically determine the extended Chernoff point. As in the previous illustrations we take our component PDFs to be Gaussian univariate functions with $p_0(x) = \mathcal{N}(1, 1)$ and $p_1(x) = \mathcal{N}(1, 3)$, while $p_2(x)$ has a variance of 1, but we vary the mean to produce different types of connections. To produce an acute connection we choose $\mu_2 = 2.2$. The result of our optimization is then $\boldsymbol{\eta}^* = [0.277, 0.335, 0.388]$.

The extended Chernoff point corresponds to Gaussian univariate coordinates $\mu_* = 1.6, \sigma_*^2 = 1.29$. It is then easy to verify that $D(p_{\boldsymbol{\eta}^*} \| p_i) = 0.1974$, for all $i = 0, 1, 2$. Similarly to produce a right connection we increase the mean of $p_2(x)$, to $\mu_2 = 2.61$. The result of our optimization is then $\boldsymbol{\eta}^* = [0.5, 0, 0.5]$, which is the mid-point (or pairwise Chernoff point) of the connection of p_0 and p_2 . The extended Chernoff point then corresponds to Gaussian univariate coordinates $\mu_* = 1.81, \sigma_*^2 = 1$, and $D(p_{\boldsymbol{\eta}^*} \| p_i) = D(p_{\eta_{01}^*} \| p_0) = D(p_{\eta_{01}^*} \| p_1) = 0.324$. Figures 22 and 23 illustrate the “radius” of the KL ball (the set of PDFs within a prescribed KL-divergence) relative to the extended Chernoff point in these acute and right connections, respectively. The optimization was performed using the MATLAB optimization toolbox and the color image depicts the KL divergence $D(p_{\boldsymbol{\eta}^*} \| p_{\mu, \sigma^2})$ over a grid of the univariate Gaussian parameters. In these cases the resultant optimizations of extended Chernoff point and SEB coincide as expected. The colors represent the KL divergence of the relevant point to all other points in the Gaussian univariate parameter space. The color-scale is clipped at that of the smallest divergence among the three component densities which are equal in these cases as expected.

In our third example we set $\mu_2 = 3$, which results in an obtuse connection. In this case the alternative optimizations for the extended Chernoff point and SEB do not result in the same PDF (point in PDF parameter space). The optimization with sum and variance constraints resulted in $\boldsymbol{\eta}^* = [1.043, -0.815, 0.772]$, while the sum and positive coefficients constraints resulted in $\boldsymbol{\eta}^* = [0.5, 0, 0.5]$. Thus the extended Chernoff point is $\mu_* = 2, \sigma_*^2 = 0.648$, while the center point of the SEB is $\mu_* = 2, \sigma_*^2 = 1$. For the extended Chernoff case the KL-ball radius is $D(p_{\boldsymbol{\eta}^*} \| p_i) = 0.541$, while for the SEB point we have $D(p_{\boldsymbol{\eta}^*} \| p_0) = 0.5$. The alternative results are shown in Figures 24 and 25 where it is clear that in the obtuse

Figure 23: The Kullback-Liebler divergence relative to the extended Chernoff point / SEB solution point (PDF) for the “right” connection of three univariate Gaussian PDFs. The black lines connecting the PDF points are *geodesics* through the exponential connection, and thus make up the boundary of the exponential connection in the normal parameter space. The colors depict the level of the KL divergence of all points (PDFs) relative to the center point (magenta asterisk). The color level is cropped at $D(p_{\eta^*} \| p_i)$ which is drawn as a contour on the image.

case the divergence at the component densities are not equal relative to the center of the SEB, i.e. $D(p_{\eta^*} \| p_1) < D(p_{\eta^*} \| p_0)$. The KL-ball of the extended Chernoff solution point does show this equality of KL divergence at the components, as expected.

2.6 Conclusion

In this paper we have studied the higher-level geometrical properties of the EEF of three PDFs. In particular, we take a novel perspective of characterizing this exponential connection of component PDFs as a triangle of PDFs. Since this space of PDFs is well known to be an exponential family that is governed by a directional divergence, we use the Chernoff distance as a symmetric (order inde-

Figure 24: The Kullback-Liebler divergence relative to the SEB solution point. With respect to SEB, the “radius” of the ball, $r_s^* = 0.5$.

pendent) measure (i.e. $C(p_1, p_2) = C(p_2, p_1)$) that (from our results) acts as a squared length between points in PDF space. We have shown that the equidistant property of the Chernoff point can be extended to an arbitrary number of PDFs within the M -component EEF, utilizing the theory of exponential families, and have established the existence criteria for such an extension. As a result we have better defined the proper constraints of the embedding parameters for both general and Gaussian PDF components. In particular, that there is utility in having some negative embedding weights in the M -component construction of the geodesic families due to Cencov and Amari. This is a novel extension of this powerful modeling paradigm. By studying three component EEFs we found that the extended Chernoff point behaves like a circumcenter in Euclidean geometry, with its dependence on connection type akin to characterizations of triangles. We found that the extended Chernoff point is equivalent to the smallest enclosing ball

Figure 25: The Kullback-Liebler divergence relative to the extended Chernoff solution point. The “radius” of the ball is $r^* = 0.541$ with respect to the extended Chernoff point.

of PDFs when the connection is acute, but the two geometrical objects correspond to separate points in the PDF space (of the EEF) if the connection is obtuse. This is important to defining proper constraints for finding the extended Chernoff point by non-linear programming optimization when no analytical solution exists. This result can be significant to utilizing the equidistant criteria for combining information in bandwidth limited sensing problems, where sharing of data amongst sensors is very limited or not feasible.

Appendix 2A - Proof of General EEF of Gaussians

Lemma 4. *Combination of Quadratic Forms [39] Take $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ as $n \times 1$ mean vectors and $\mathbf{C}_1, \mathbf{C}_2$ as $n \times n$ symmetric covariance matrices. Then*

$$\begin{aligned} (\mathbf{x} - \boldsymbol{\mu}_1)^T \mathbf{C}_1^{-1} (\mathbf{x} - \boldsymbol{\mu}_1) + (\mathbf{x} - \boldsymbol{\mu}_2)^T \mathbf{C}_2^{-1} (\mathbf{x} - \boldsymbol{\mu}_2) \\ = (\mathbf{x} - \boldsymbol{\mu}_a)^T \mathbf{C}_a^{-1} (\mathbf{x} - \boldsymbol{\mu}_a) + \mathcal{D} \end{aligned} \quad (145)$$

where

$$\boldsymbol{\mu}_a = \mathbf{C}_a (\mathbf{C}_1^{-1} \boldsymbol{\mu}_1 + \mathbf{C}_2^{-1} \boldsymbol{\mu}_2) \quad (146)$$

$$\mathbf{C}_a^{-1} = \mathbf{C}_1^{-1} + \mathbf{C}_2^{-1} \quad (147)$$

$$\mathcal{D} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \mathbf{C}_1^{-1} (\mathbf{C}_1^{-1} + \mathbf{C}_2^{-1})^{-1} \mathbf{C}_2^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \quad (148)$$

Proof. Recall we can write the EEF as

$$p(\mathbf{x}; \boldsymbol{\eta}) = \frac{p_{M-1}(\mathbf{x})^{\eta_{M-1}} \cdots p_1(\mathbf{x})^{\eta_1} p_0(\mathbf{x})^{\eta_0}}{\int_{-\infty}^{\infty} p_{M-1}(\mathbf{x})^{\eta_{M-1}} \cdots p_1(\mathbf{x})^{\eta_1} p_0(\mathbf{x})^{\eta_0} d\mathbf{x}}$$

Then the product of the M Gaussians is written as

$$\mathcal{G}(\boldsymbol{\eta}; M) \cdot \exp \left(-\frac{1}{2} \left[\sum_{i=0}^{M-1} (\mathbf{x} - \boldsymbol{\mu}_i)^T [\eta_i \mathbf{C}_i^{-1}] (\mathbf{x} - \boldsymbol{\mu}_i) \right] \right)$$

where

$$\mathcal{G}(\boldsymbol{\eta}; M) = \prod_{i=0}^{M-1} \left(\frac{1}{(2\pi)^{n/2} |\mathbf{C}_i|^{1/2}} \right)^{\eta_i} \quad (149)$$

Since $\mathcal{G}$ does not depend on $\mathbf{x}$, this term cancels and we are left with

$$p(\mathbf{x}; \boldsymbol{\eta}) = \frac{\exp \left(-\frac{1}{2} \left[\sum_{i=0}^{M-1} (\mathbf{x} - \boldsymbol{\mu}_i)^T [\eta_i \mathbf{C}_i^{-1}] (\mathbf{x} - \boldsymbol{\mu}_i) \right] \right)}{\int_{-\infty}^{\infty} \exp \left(-\frac{1}{2} \left[\sum_{i=0}^{M-1} (\mathbf{x} - \boldsymbol{\mu}_i)^T [\eta_i \mathbf{C}_i^{-1}] (\mathbf{x} - \boldsymbol{\mu}_i) \right] \right) d\mathbf{x}}$$

Now let $A_i = \eta_i \mathbf{C}_i^{-1}$ then from the lemma we can combine the first two quadratic terms (i.e. p_0, p_1) as

$$(\mathbf{x} - \mathbf{c}_1)^T (A_0 + A_1) (\mathbf{x} - \mathbf{c}_1) + \mathcal{D}_1$$

where $\mathcal{D}_1$ does not depend on $\mathbf{x}$ (and thus cancels out), while $\mathbf{c}_1 = (A_0 + A_1)^{-1} (A_0 \boldsymbol{\mu}_0 + A_1 \boldsymbol{\mu}_1)$. Applying the lemma again to combine

$$(\mathbf{x} - \mathbf{c}_1)^T (A_0 + A_1) (\mathbf{x} - \mathbf{c}_1) + (\mathbf{x} - \boldsymbol{\mu}_2)^T (A_2) (\mathbf{x} - \boldsymbol{\mu}_2)$$

we have

$$(\mathbf{x} - \mathbf{c}_2)^T (A_0 + A_1 + A_2) (\mathbf{x} - \mathbf{c}_2) + \mathcal{D}_2$$

where $\mathbf{c}_2 = (A_0 + A_1 + A_2)^{-1} ((A_0 + A_1) \mathbf{c}_1 + A_2 \boldsymbol{\mu}_2)$ but substitution in for $\mathbf{c}_1$ results in

$$\mathbf{c}_2 = (A_0 + A_1 + A_2)^{-1} (A_0 \boldsymbol{\mu}_0 + A_1 \boldsymbol{\mu}_1 + A_2 \boldsymbol{\mu}_2).$$

Repeating this process $M - 1$ times gives the required Gaussian quadratic form. The integration in the denominator then gives the leading term for proper normalization. □

Appendix 2B - Projection onto the orthogonal bisector of a pairwise connection

Fact 1. *Let $p_{2\perp}$ be the orthogonal projection of PDF p_2 onto the family $p_{\boldsymbol{\eta}(t)}$ (see definition 10). Then*

$$p_{2\perp} = \exp \left(\eta_1 \ln \frac{p_1}{p_0} - \mathcal{K}_0(\eta_1, 1) + \ln p_2 \right) \quad (150)$$

Proof. From Kullback [5] the solution of $\min_p D(p||p_2)$ such that $E_p[T] = \theta$ is given as

$$p_{2\perp} = \frac{\exp(\lambda T + \ln p_2)}{\int \exp(\lambda T + \ln p_2) dx} \quad (151)$$

where λ is chosen such that $E_p[T] = \theta$. Let $T = \ln \frac{p_1}{p_0}$ and $\theta = 0$. Then $E_{p_{2\perp}} \left[\ln \frac{p_1}{p_0} \right] = 0$ and thus $p_{2\perp} \in p_{\boldsymbol{\eta}(t)}$. Now

$$p_{2\perp} = \frac{\exp \left(\lambda \ln \frac{p_1}{p_0} + \ln p_2 \right)}{\int \exp \left(\lambda \ln \frac{p_1}{p_0} + \ln p_2 \right) dx}$$

but the CGF of the three component PDF is given as

$$\mathcal{K}_0(\eta_1, \eta_2) = \ln \int \exp \left(\eta_1 \ln \frac{p_1}{p_0} + \eta_2 \ln \frac{p_2}{p_0} + \ln p_0 \right) dx$$

so that $\mathcal{K}_0(\lambda, 1) = \ln \int \exp \left(\lambda \ln \frac{p_1}{p_0} + \ln p_2 \right) dx$ and thus

$$p_{2\perp} = \exp \left(\lambda \ln \frac{p_1}{p_0} - \mathcal{K}_0(\lambda, 1) + \ln p_2 \right)$$

Replacing λ with η_1 yields (150) □

Fact 2. *The projection of p_0 (a component of pairwise connection p_{01}) onto orthogonal bisector $p_{\boldsymbol{\eta}(t)}$ occurs at the pairwise Chernoff point $p_{\eta_{01}^*}$. That is*

$$p_{0\perp} = \exp(\eta_{01}^* \ln \frac{p_1}{p_0} - \mathcal{K}_0(\eta_{01}^*, 0) + \ln p_0) \quad (152)$$

Note that $p_{0\perp} \in p_{\boldsymbol{\eta}(t)}$ and corresponds to $\eta_2 = 0$.

Proof. Similar to proof for fact 1. □

Appendix 2C: Properties of the EEF of Univariate Gaussians

Consider the EEF $p(x; \eta)$ of two univariate Gaussian PDFs $p_0(x; \mu_0, \sigma_0^2)$ and $p_1(x; \mu_1, \sigma_1^2)$. Recall that $p(x, \eta)$ is also Gaussian with mean and variance given as (functions of η)

$$\sigma^2(\eta) = \frac{\sigma_0^2 \sigma_1^2}{\sigma_1^2 - \eta(\sigma_1^2 - \sigma_0^2)} \quad (153)$$

$$\mu(\eta) = \sigma^2(\eta) \left(\frac{(1-\eta)\mu_0}{\sigma_0^2} + \frac{\eta\mu_1}{\sigma_1^2} \right) = \frac{(1-\eta)\sigma_1^2\mu_0 + \eta\sigma_0^2\mu_1}{\sigma_1^2 - \eta(\sigma_1^2 - \sigma_0^2)} \quad (154)$$

It is easy to show that this family (curve) maps as a straight line in the μ, σ^2 coordinates. Let $F(\eta) = (1-\eta)\sigma_1^2 + \eta\sigma_0^2$. Then this family of densities consists of a subset of all univariate densities mapped out by parameter η . We can write the slope of this family in the univariate Gaussian space as

$$s(\eta_1, \eta_2) = \frac{\sigma^2(\eta_2) - \sigma^2(\eta_1)}{\mu(\eta_2) - \mu(\eta_1)} = \frac{\Delta\sigma^2}{\Delta\mu} \quad (155)$$

for any $\eta_2 > \eta_1$, $\eta_1, \eta_2 \in (0, 1)$. After some more algebra we have

$$\Delta\sigma^2 = \frac{\sigma_0^2 \sigma_1^2 (\eta_2 - \eta_1) (\sigma_1^2 - \sigma_0^2)}{F(\eta_1) F(\eta_2)} \quad (156)$$

$$\Delta\mu = \frac{\sigma_0^2 \sigma_1^2 (\eta_2 - \eta_1) (\mu_1 - \mu_0)}{F(\eta_1) F(\eta_2)} \quad (157)$$

and thus the slope of the EEF of two univariate Gaussian PDFs is

$$s(\eta_1, \eta_2) = \frac{\sigma_1^2 - \sigma_0^2}{\mu_1 - \mu_0} \quad (158)$$

which is independent of η (i.e. constant over the connection) and thus maps a straight line in the Gaussian parameter space. We have shown that the EEF maps a straight line in Gaussian parameter space but is it a geodesic? For the curve mapped out by the EEF to be a geodesic in these coordinates it must be true that [9]

$$\ddot{\theta}_k + \sum_{i,j} \Gamma_{ij}^k \dot{\theta}_i \dot{\theta}_j = 0 \quad (159)$$

with $\theta_1 = \mu$, $\theta_2 = \sigma^2$ and all derivatives taken with respect to η . The Christoffel symbols are calculated as

$$\Gamma_{ij}^k = g^{kl} E_{p(x; \boldsymbol{\theta})} \left[\frac{\partial^2 \ln p(x; \boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \cdot \frac{\partial \ln p(x; \boldsymbol{\theta})}{\partial \theta_l} \right] \quad (160)$$

where

$$\left[\frac{\partial^2 \ln p(x; \boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right] = \begin{bmatrix} -\frac{1}{\theta_2} & -\frac{x-\theta_1}{\theta_2^2} \\ -\frac{x-\theta_1}{\theta_2^2} & \frac{1}{2\theta_2^2} - \frac{(x-\theta_1)^2}{\theta_2^3} \end{bmatrix} \quad (161)$$

and

$$g^{kl} = \mathcal{I}^{-1}(\boldsymbol{\theta}) = \begin{bmatrix} \theta_2 & 0 \\ 0 & 2\theta_2^2 \end{bmatrix} \quad (162)$$

where g^{kl} is a metric equal to the inverse Fisher information and the first derivatives (scores) are

$$\frac{\partial \ln p(x; \boldsymbol{\theta})}{\partial \theta_1} = \frac{x - \theta_1}{\theta_2} \quad (163)$$

$$\frac{\partial \ln p(x; \boldsymbol{\theta})}{\partial \theta_2} = -\frac{1}{\theta_2} + \frac{1}{2\theta_2} (x - \theta_1)^2 \quad (164)$$

Given these derivatives we can calculate the Christoffel symbols as [40]

$$\Gamma_{ij}^1 = \begin{bmatrix} 0 & -\frac{1}{\theta_2} \\ -\frac{1}{\theta_2} & 0 \end{bmatrix} \quad (165)$$

$$\Gamma_{ij}^2 = \begin{bmatrix} 0 & 0 \\ 0 & -\frac{2}{\theta_2} \end{bmatrix} \quad (166)$$

Substituting (165) and (166) into (159) we get

$$\ddot{\theta}_1 - \frac{2}{\theta_2} \dot{\theta}_1 \dot{\theta}_2 = 0 \quad (167)$$

$$\ddot{\theta}_2 - \frac{2}{\theta_2} \dot{\theta}_2^2 = 0 \quad (168)$$

Thus for the EEF to be a geodesic in the given univariate coordinates these conditions must hold. To see this we calculate the coordinate derivatives as

$$\dot{\theta}_1 = \dot{\mu}(\eta) = \frac{\sigma_0^2 \sigma_1^2 (\mu_1 - \mu_0)}{[\sigma_1^2 - \eta(\sigma_1^2 - \sigma_0^2)]^2} \quad (169)$$

$$\dot{\theta}_2 = \dot{\sigma}^2(\eta) = \frac{\sigma_0^2 \sigma_1^2 (\sigma_1^2 - \sigma_0^2)}{[\sigma_1^2 - \eta(\sigma_1^2 - \sigma_0^2)]^2} \quad (170)$$

and second derivatives as

$$\ddot{\theta}_1 = \ddot{\mu}(\eta) = \frac{2\sigma_1^2\sigma_0^2(\sigma_1^2 - \sigma_0^2)(\mu_1 - \mu_0)}{[\sigma_1^2 - \eta(\sigma_1^2 - \sigma_0^2)]^3} \quad (171)$$

$$\ddot{\theta}_2 = \ddot{\sigma}(\eta) = \frac{2\sigma_1^2\sigma_0^2(\sigma_1^2 - \sigma_0^2)^2}{[\sigma_1^2 - \eta(\sigma_1^2 - \sigma_0^2)]^3} \quad (172)$$

Substitution of (169 - 172) into (167-168) shows that the EEF satisfies the conditions for a geodesic in the given coordinates.

Appendix 2D: Numerical Results for Chernoff Metric

In the above we have defined the Chernoff distance as a measure of “squared length” and we have illustrated using univariate Gaussians, some triangle properties of PDFs which mirror that of Euclidean geometry. It is clear that the Chernoff distance is a symmetric measure, i.e. $C(p_0, p_1) = C(p_1, p_0)$, and by the properties of KL-divergence, $C(p_0, p_1) \geq 0$ with equality if and only if $p_0 = p_1$ almost everywhere. In addition, if the Chernoff distance does indeed act as a squared distance globally in the space, this implies that a simple transformation (square root) can result in a true metric. We explore this with a numerical example. Consider the triangle of PDFs with components $p_0(x) = \mathcal{N}(1, 1)$, $p_1(x) = \mathcal{N}(1, 3)$ and $p_2(x) = \mathcal{N}(\mu, 1)$, where we define μ to vary smoothly from 2.2 to 2.8. Figure 26 shows the Chernoff distance between each pair of PDFs as a function of μ , where D_{02}^* is the Chernoff distance of p_0 and p_2 analogous to the “squared length” of the “side” mapped by the EEF of p_0 and p_2 (previously defined as p_{02}).

Figure 26: The Chernoff distance of each “side”, i.e. pairwise connection, of the three component EEFs as a function of the mean parameter μ_2 of p_2 . Note that pair p_0, p_2 has the largest Chernoff distance and thus plays the role of hypotenuse (largest squared length).

Figure 27: The difference of the Chernoff distance of the hypotenuse and the sum of the Chernoff distance of the two “shorter sides”. Note that the triangle properties determined are consistent with the determinations made from the I-center.

Similarly D_{01}^* and D_{12}^* correspond to p_{01} and p_{12} , respectively. Note that D_{01}^* is constant as expected and that D_{02}^* is greatest for all μ . Thus “side” p_{02} plays the role of hypotenuse for this triangle characterization for all μ . Now as in plane geometry, we can characterize the triangle type by the relationship of the lengths and / or squared lengths of the sides. Indeed, if D_{02}^* is the squared distance of the hypotenuse, then the connection of PDFs is acute if and only if $D_{02}^* - (D_{01}^* + D_{12}^*) < 0$. Similarly if the difference is greater than zero, the connection is obtuse, and if equal it is right. Figure 27 shows the result of our numerical experiment where we found the extended Chernoff point for every μ and thus validated the characterization shown. We did observe some instability of the optimization near the crossing from acute to obtuse but was able to remedy by forcing more iterations, i.e. making more stringent stopping criteria. This result is consistent with our previous result of the triangle of PDFs formed by the EEF obeying an analogous planar geometry triangle property, namely the median-halflength relation of Theorem 4.

So the next question is can we produce a metric based on Chernoff distance

Figure 28: This figure illustrates that the square root of the Chernoff distance satisfies the triangle inequality and thus is a metric in Univariate Gaussian space.

by simply taking the square root? Define L_{02} as the square root of D_{02}^* (similarly for L_{01} and L_{12}). Then $L_{02} > L_{01}, L_{12}$, and the triangle inequality holds if $L_{02} < L_{01} + L_{12}$. In Figure 28 it is shown that this measure satisfies the triangle inequality and thus $\sqrt{D^*}$ meets all conditions of a metric, in a univariate Gaussian PDF space. To prove this generally is a source of future work and could hold significance in a number of applications.

List of References

- [1] H. Chernoff, "A measure of asymptotic efficiency of tests of a hypothesis based on the sum of observations," *Annals of Mathematical Statistics*, vol. 23, no. 4, pp. 493–507, 1952.
- [2] T. M. Cover and J. A. Thomas, *Elements of information theory*, 2nd ed. John Wiley and Sons, 2006.
- [3] M. Westover, "Asymptotic geometry of multiple hypothesis testing," *Information Theory, IEEE Transactions on*, vol. 54, no. 7, pp. 3327–3329, July 2008.
- [4] C. Leang and D. Johnson, "On the asymptotics of M-hypothesis Bayesian detection," *IEEE Transactions on Information Theory*, vol. 43, no. 1, pp. 280–282, January 1997.

- [5] S. Kullback, *Information Theory and Statistics*. Dover, 1968.
- [6] F. Nielsen, “An Information-Geometric characterization of Chernoff information,” *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 269–272, 2013.
- [7] S. Kay, “Exponentially embedded families - new approaches to model order estimation,” *IEEE Journal of AES*, vol. 41, no. 1, pp. 333–345, 2005.
- [8] N. Cencov, *Statistical decision rules and optimal inference*. American Mathematical Society, 1982.
- [9] S. Amari and H. Nagaoka, *Methods of Information Geometry*. Oxford University Press, 2000.
- [10] Y. Wang and X. Li, “Distributed estimation fusion under unknown cross-correlation: An analytic center approach,” in *Information Fusion (FUSION), 2010 13th Conference on*, July 2010, pp. 1–8.
- [11] Y. Wang and X. R. Li, “Distributed estimation fusion with unavailable cross-correlation,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 48, no. 1, pp. 259–278, January 2012.
- [12] T. Bailey, S. Julier, and G. Agamennoni, “On conservative fusion of information with unknown non-Gaussian dependence,” in *15th International Conference on Information Fusion*. IEEE, 2012, pp. 1876–1883.
- [13] B. Khaleghi, A. Khamis, F. Karray, and S. Razavi, “Multisensor data fusion: A review of the state-of-the-art,” *Information Fusion*, vol. 14, no. 1, pp. 28–44, 2013.
- [14] S. J. Julier, “An empirical study into the use of Chernoff information for robust, distributed fusion of Gaussian mixture models,” in *9th International Conference on Information Fusion*. IEEE, 2006, pp. 1–8.
- [15] M. Hurley, “An Information-Theoretic justification for Covariance Intersection and its generalization,” Lincoln Laboratory (Massachusetts Institute of Technology), Lexington, MA, Tech. Rep. TR 1074, August 2001.
- [16] N. R. Ahmed and M. Campbell, “Fast consistent Chernoff Fusion of Gaussian mixtures for Ad Hoc sensor networks,” *IEEE Transactions on Signal Processing*, vol. 60, no. 12, pp. 6739–6745, 2012.
- [17] W. Farrell and C. Ganesh, “Generalized Chernoff fusion approximation for practical distributed data fusion,” in *Information Fusion, 2009. FUSION '09. 12th International Conference on*, 2009, pp. 555–562.
- [18] S. Amari, “Integration of stochastic models by minimizing alpha-divergence,” *Neural Computation*, vol. 19, pp. 2780–2796, 2007.

- [19] G. Hardy, J. Littlewood, and G. Polya, *Inequalities*. Cambridge University Press, 1967.
- [20] J. Aczél, “On weighted synthesis of judgments,” *Aequationes Mathematicae*, vol. 27, no. 1, pp. 288–307, 1984.
- [21] T. Heskes, “Selecting weighting factors in logarithmic opinion pools,” *Advances in neural information processing systems*, pp. 266–272, 1998.
- [22] O. Barndorff-nielsen, *Information and exponential families in statistical theory*. John Wiley and Sons, 1978.
- [23] E. L. Lehmann and G. Casella, *Theory of point estimation*. Springer, 1998.
- [24] S. Amari and A. Cichocki, “Information geometry of divergence functions,” *Bulletin of the Polish Academy of Sciences: Technical Sciences*, vol. 58, no. 1, pp. 183–195, 2010.
- [25] S.-I. Amari, “Information geometry on hierarchy of probability distributions,” *Information Theory, IEEE Transactions on*, vol. 47, no. 5, pp. 1701–1711, Jul 2001.
- [26] P. J. Bickel and K. A. Doksum, *Mathematical Statistics: Basic ideas and selected topics Vol. 1*. Pearson Prentice Hall, 2007.
- [27] L. D. Brown, *Fundamentals of statistical exponential families with applications in statistical decision theory*, S. S. Gupta, Ed. Institute of Mathematical Statistics, 1986.
- [28] J. Ginebra, “On the measure of the information in a statistical experiment,” *Bayesian Analysis*, vol. 2, no. 1, pp. 167–211, 2007.
- [29] I. Fazekas and F. Liese, “Some properties of the Hellinger transform and its application in classification problems,” *Computers & Mathematics with Applications*, vol. 31, no. 8, pp. 107–116, 1996.
- [30] F. Nielsen and R. Nock, “Sided and symmetrized bregman centroids,” *IEEE Trans. Inf. Theor.*, vol. 55, pp. 2882–2904, June 2009. [Online]. Available: <http://portal.acm.org/citation.cfm?id=1669524.1669559>
- [31] S. Costa, S. Santos, and J. Strapasson, “Fisher information distance: a geometrical reading?” *arXiv preprint arXiv:1210.2354v2*, 2012.
- [32] R. Costa and S. Kay, “A geometrical interpretation of exponentially embedded families of Gaussian probability density functions for model selection,” *Signal Processing, IEEE Transactions on*, vol. 61, no. 1, pp. 62–67, 2013.

- [33] F. Nielsen and R. Nock, “Approximating smallest enclosing balls,” in *International Conference on Computational Science and Its Applications (ICCSA)*, 2009.
- [34] F. Nielsen and R. Nock, “On the smallest enclosing information disk,” *Information Processing Letters*, vol. 105, pp. 93–97, 2008.
- [35] F. Nielsen. “Chernoff information of exponential families.” February 2011. [Online]. Available: <http://arxiv.org/pdf/1102.2684>
- [36] R. Fletcher, *Practical Methods of Optimization*. John Wiley and Sons, 1987.
- [37] I. Csiszar and F. Matus, “Information projections revisited,” *Information Theory, IEEE Transactions on*, vol. 49, no. 6, pp. 1474–1490, june 2003.
- [38] I. Csiszár and P. C. Shields, *Information theory and statistics: A tutorial*. Now Publishers Inc, 2004.
- [39] K. Petersen and M. Pedersen. “The matrix cookbook.” November 2008. [Online]. Available: <http://matrixcookbook.com>
- [40] M. K. Murray and J. W. Rice, *Differential geometry and statistics*. CRC Press, 1993, vol. 48.

CHAPTER 3

On Combining Estimates using the Exponentially Embedded Family Paradigm

by

Russell Costa and Steven Kay

Dept. of Electrical, Computer and Biomedical Engineering

University of Rhode Island, Kingston, RI, USA

submitted to IEEE Signal Processing Letters

Abstract

The exponentially embedded family (EEF) of probability density functions (PDFs) is an important modeling paradigm in a number of signal processing applications including model selection and hypothesis testing. In recent years the availability of inexpensive sensors that can gather information over wide, distributed areas, has led to many interesting challenges in distributed estimation and data fusion. In many such applications there is a need for combining estimates extracted from data only available locally at individual sensors. In this letter we apply the EEF to the classical problem of combining mean estimates from independent populations. This problem provides a tractable example for the use of this paradigm that may provide insight into the broader application of this approach.

3.1 Introduction

In numerous scientific fields there is a need for combining estimates obtained from independent as well as dependent experiments. One well established application is in the combination of experts [1, 2], where the expert's opinion is represented by a probability distribution on some parameters of interest. In this vein, various forms of opinion pools, e.g. both linear and log-based, have been proposed and studied [3, 4]. There is also related research in combination of forecasts [5], or in a Bayesian framework, elicitation of prior information [6]. In recent years, the explosion in distributed systems and large scale sensor networks has made the need for combining information, reduced / extracted from data, a significant challenge that has garnered a lot of attention [7]. In these types of problems combining state estimates [8], classifiers [9], and detection decisions [10], have all been considered. In this paper we consider a particular approach to these types of problems which utilize the EEF of probability density functions. The EEF is essentially a log-opinion pool and has been applied to numerous problems including model

selection [11] and data fusion [12] where it is described as an exponential mixture model [13]. The main motivation for this paradigm in the fusion literature is as a conservative approach which is useful when the correlation among sensor data is unknown [14].

While the EEF has been shown useful for data fusion, there is always a question when utilizing this approach of how to choose the particular weight among the infinite possibilities. In Covariance Intersection (CI) approaches applied to combining state estimates, the weight is chosen to minimize the trace or determinant of the fused covariance matrix. In [12] the CI approach is shown to be equivalent to choosing the PDF with minimum entropy (assuming Gaussian components). In the same work the Chernoff criterion is suggested, which chooses the PDF within the EEF that is equidistant from the two PDFs being combined. More recently, the tracking community has found the centroid (equal weights) to perform well [15]. However, despite the large number of papers written very few have compared various criteria on a tractable problem.

In this work we apply the EEF to the classical problem of estimating the common mean from two independent normal sources [16] with unknown variance. We form the EEF of the estimated sample mean PDFs and present it as a potential family of solutions to the common mean estimation problem. Within this family of estimates we focus our study on the relationship between choosing the centroid versus the Chernoff equi-distance property as a fusion criteria. Our results show that in the common mean problem, the centroid corresponds to the classical Graybill-Deal estimator and that the Chernoff approach outperforms this estimator for small sample size and variance ratios less than three.

3.2 Combining Common means

Consider two separate data sets $\mathbf{x}_0, \mathbf{x}_1$ of size n_0, n_1 respectively. We assume that the two data sets consist of independent identically distributed (i.i.d.) samples from Gaussian processes with mean, μ and variance σ_0^2 and σ_1^2 , respectively. For these independent samples we can estimate the sample means, μ_i as

$$\hat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij} \quad (173)$$

where x_{ij} is the j^{th} sample from the data vector $\mathbf{x}_i$, and clearly the sample means are independent. It is well known that [17] the distribution of the sample mean for each Gaussian i.i.d. sample is given as $\hat{\mu}_i \sim \mathcal{N}\left(\mu, \frac{\sigma_i^2}{n_i}\right)$.

If the variances of each sample are known, then we can estimate the common mean by the best linear unbiased estimator (BLUE), written as [17]

$$\hat{\mu} = \frac{n_0 \sigma_1^2 \hat{\mu}_0 + n_1 \sigma_0^2 \hat{\mu}_1}{n_1 \sigma_0^2 + n_0 \sigma_1^2} \quad (174)$$

When the variances are unknown this form was adopted by Graybill and Deal with the sample variances (unbiased) taking the place of the known variances.

$$\hat{\mu}_{GD} = \frac{n_0 \hat{\sigma}_1^2 \hat{\mu}_0 + n_1 \hat{\sigma}_0^2 \hat{\mu}_1}{n_1 \hat{\sigma}_0^2 + n_0 \hat{\sigma}_1^2}$$

This is known as the Graybill-Deal estimator [18] and although there have been various improvements [16, 19], this simple closed form solution leads to its continued use. When $n_0 = n_1 = m$ then

$$\hat{\mu}_{GD} = \frac{\hat{\sigma}_1^2 \hat{\mu}_0 + \hat{\sigma}_0^2 \hat{\mu}_1}{\hat{\sigma}_0^2 + \hat{\sigma}_1^2} = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_1^2 + \hat{\sigma}_0^2} \hat{\mu}_0 + \frac{\hat{\sigma}_0^2}{\hat{\sigma}_1^2 + \hat{\sigma}_0^2} \hat{\mu}_1 \quad (175)$$

Graybill and Deal showed that this is an unbiased estimator, and for $m > 9$ this estimator is uniformly better than either individual estimate $\hat{\mu}_i$. That is, it is unbiased and has a smaller MSE for all variance ratios $\rho = \frac{\sigma_1^2}{\sigma_0^2}$.

3.3 Exponentially Embedded Families

The EEF for a pairwise connection of two component PDFs, $p_0(x)$, $p_1(x)$ is written as

$$p(x; \eta) = \frac{p_0(x)^{1-\eta} p_1(x)^\eta}{\int p_0(x)^{1-\eta} p_1(x)^\eta dx} \quad (176)$$

where η is a scalar parameter. We can rewrite (176) as

$$p(x; \eta) = \exp \left(\eta \log \frac{p_1(x)}{p_0(x)} + \log p_0(x) - \mathcal{K}_0(\eta) \right) \quad (177)$$

Then $p(x; \eta)$ is an exponential family of PDFs over all values of η for which $\mathcal{K}_0(\eta)$ is finite, where $\mathcal{K}_0(\eta) = \log \int p_0^{1-\eta}(x) p_1^\eta(x) dx$ is the cumulant generating function. It can be shown that the EEF with two Gaussian components, reproduces a Gaussian family, which yields $\mathcal{N}(\mu_\eta, \sigma_\eta^2)$, with mean

$$\mu_\eta = \frac{(1-\eta)\sigma_1^2\mu_0 + \eta\sigma_0^2\mu_1}{(1-\eta)\sigma_1^2 + \eta\sigma_0^2} \quad (178)$$

and variance $\sigma_\eta^2 = \frac{\sigma_0^2\sigma_1^2}{(1-\eta)\sigma_1^2 + \eta\sigma_0^2}$. Note that each value of $\eta \in [0, 1]$ corresponds to a particular PDF on the geodesic connecting the two component PDFs. In addition, the variance parameterized by η can be used as a conservative estimate of the variance when there are unknown dependencies in the data sets contributing to the sample mean estimates [20]. However, this work will focus on mean estimation only.

This family of PDFs has a special information geometric property in that it is the weighted geometric mean of PDFs which is optimal with respect to the weighted sum of “left-sided” Kullback-Leibler divergence (KL-divergence). That is $p(x; \eta)$ is the result of [15, 21]

$$\arg \min_{p(x)} [(1-\eta)D(p(x)||p_0(x)) + \eta D(p(x)||p_1(x))] \quad (179)$$

where $D(p(x)||p_0(x)) = \int p(x) \log \frac{p(x)}{p_0(x)} dx$. Thus any fixed weight choice optimizes a specific weighting of relative information.

As an alternative to the GD estimator, we consider the EEF of the estimated PDFs of the sample means. This is just the weighted geometric mean of the PDFs of the independent sample means. Specifically, we construct the EEF of the two density estimates $p(x; \hat{\mu}_0, \hat{\sigma}_0^2)$ and $p(x; \hat{\mu}_1, \hat{\sigma}_1^2)$ to form $p(x; \mu_\eta, \sigma_\eta^2)$ where we have accounted for the sample size (m) in the variance estimates of the sample mean and we restrict $\eta \in [0, 1]$. Note that equation (178) provides a family of potential estimators for the common mean, where we replace the known means and variances with their estimates to provide $\hat{\mu}_\eta$. Within this family we will consider two choices for the embedding parameter η . The first is to choose $\eta = \frac{1}{2}$. This is commonly known as the Centroid and corresponds to the PDF which is minimal with respect to the average KL-divergence, that is

$$\eta_c = \arg \min_{\eta} \left[\frac{1}{2} D(p(x; \eta) \| p_0) + \frac{1}{2} D(p(x; \eta) \| p_1) \right] \quad (180)$$

which is trivial given (179). The common mean estimate then becomes from (178) with estimated means and variances

$$\hat{\mu}_c = \frac{\hat{\sigma}_1^2 \hat{\mu}_0 + \hat{\sigma}_0^2 \hat{\mu}_1}{\hat{\sigma}_0^2 + \hat{\sigma}_1^2} \quad (181)$$

which is identical to (175), the GD estimator of common mean. The other choice we will consider is the embedding parameter η corresponding to the Chernoff criteria [22]. This criteria consists of the embedding weight for which the fused PDF is equidistant from the two components with respect to left-sided KL-divergence. That is the Chernoff embedding parameter (or Chernoff point, [23]) is defined as η_* such that $D(p_{\eta_*} \| \hat{p}_1) = D(p_{\eta_*} \| \hat{p}_0)$, or equivalently

$$\eta_* = \arg \min_{\eta} \mathcal{K}_0(\eta) \quad (182)$$

The PDF corresponding to η_* , $p(x; \eta_*)$, can be found as a closed form solution for pairwise Chernoff of Gaussian PDFs (unknown mean and variance). Removing

the hat notation for estimates, take $p_0(x)$ as $\mathcal{N}(\mu_0, \sigma_0^2)$, and $p_1(x)$ as $\mathcal{N}(\mu_1, \sigma_1^2)$, with $x \in \mathbb{R}$. After substitution of the PDFs into $\mathcal{K}_0(\eta)$ we have

$$\mathcal{K}_0(\eta) = \log \sigma_1 + \eta \log \frac{\sigma_0}{\sigma_1} - \frac{1}{2} \frac{\eta(1-\eta)(\mu_1 - \mu_0)^2}{\sigma_1^2 + \eta(\sigma_0^2 - \sigma_1^2)} - \frac{1}{2} \log [\sigma_1^2 + \eta(\sigma_0^2 - \sigma_1^2)] \quad (183)$$

Then differentiating with respect to η and solving the quadratic we have

$$\eta_* = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \quad (184)$$

where

$$a = (\sigma_0^2 - \sigma_1^2) \left[(\sigma_0^2 - \sigma_1^2) \log \frac{\sigma_0^2}{\sigma_1^2} + (\mu_1 - \mu_0)^2 \right] \quad (185)$$

$$b = 2\sigma_1^2 \left[(\mu_1 - \mu_0)^2 + (\sigma_0^2 - \sigma_1^2) \log \frac{\sigma_0^2}{\sigma_1^2} \right] - (\sigma_0^2 - \sigma_1^2)^2 \quad (186)$$

$$c = \sigma_1^2 \left[\sigma_1^2 \log \frac{\sigma_0^2}{\sigma_1^2} - (\sigma_0^2 - \sigma_1^2) - (\mu_1 - \mu_0)^2 \right] \quad (187)$$

which is the unique result of picking the solution which is on the geodesic (i.e. $\eta_* \in [0, 1]$). Let $\sigma_d = \sigma_0^2 - \sigma_1^2$, $\rho = \frac{\sigma_1^2}{\sigma_0^2}$, and $\mu_d = \mu_1 - \mu_0$. In addition, let $\tau = \mu_d^2 - \sigma_d \log \rho$. Then after some algebra

$$\hat{\eta}_* = \frac{\sigma_d^2 - 2\sigma_1^2\tau + \sqrt{4\sigma_0^2\sigma_1^2\mu_d^2\tau + \sigma_d^4}}{2\sigma_d\tau} \quad (188)$$

Substitution of this into (178) gives

$$\hat{\mu}_{\eta_*} = \hat{A}_0 \hat{\mu}_0 + \hat{A}_1 \hat{\mu}_1 \quad (189)$$

where

$$\hat{A}_0 = \frac{\sigma_1^2 \left(2\sigma_0^2\tau - \sigma_d^2 - \sqrt{4\sigma_0^2\sigma_1^2\mu_d^2\tau + \sigma_d^4} \right)}{\sigma_d \left(\sigma_d^2 + \sqrt{4\sigma_0^2\sigma_1^2\mu_d^2\tau + \sigma_d^4} \right)} \quad (190)$$

$$\hat{A}_1 = \frac{\sigma_0^2 \left(\sigma_d^2 - 2\sigma_1^2\tau + \sqrt{4\sigma_0^2\sigma_1^2\mu_d^2\tau + \sigma_d^4} \right)}{\sigma_d \left(\sigma_d^2 + \sqrt{4\sigma_0^2\sigma_1^2\mu_d^2\tau + \sigma_d^4} \right)} \quad (191)$$

and each variable on the right hand side is estimated from data $\mathbf{x}_0$ and $\mathbf{x}_1$. Note that it is easy to show that $\hat{A}_0 + \hat{A}_1 = 1$, meaning that the Chernoff estimator is

(like the centroid) a convex combination of the individual sample mean estimates. In practice, the embedding parameter $\hat{\eta}_*$ will take on different values in $[0, 1]$, depending on the estimated PDFs that are to be combined. In contrast, the centroid is a fixed choice of embedding parameter.

3.4 Numerical Results

We perform Monte Carlo simulations to quantify the performance of the Chernoff common mean estimator relative to the Centroid / Graybill-Deal estimator. In our simulations we take the true mean as $\mu = 2$ and the true variance of sample $\mathbf{x}_0$ as $\sigma_0^2 = 1$. Over our simulations we take σ_1^2 as variable from 1 to 10, so that the ratio ρ also varies over this range. In addition we vary m in the range 2 to 40. In each Monte Carlo iteration we draw m i.i.d. samples from p_0 and p_1 , respectively. We then calculate the sample mean, and sample variance of each vector and use them to form $\hat{\mu}_c$, (181) and $\hat{\mu}_{\eta_*}$, (189). For each estimate of the common mean, we calculate the mean squared error (MSE) performance and average over all Monte Carlo iterations.

Figure 29 shows the difference in MSE performance of the two estimators for fixed $\rho = 1$, $\rho = 2$, and $\rho = 3$ respectively. Specifically we show the MSE of Centroid minus the MSE of Chernoff, calculated over 100,000 Monte Carlo iterations. Note that in all cases the Chernoff estimator outperforms the Centroid estimator for smaller sample sizes. Specifically for $\rho = 1$, the Chernoff estimator outperforms Centroid for all sample sizes calculated (up to and including $m = 40$), for $\rho = 2$ up to $m = 26$, and for $\rho = 3$ up to $m = 8$. Thus we see that the gains in performance diminish as the ratio in variance between the two samples increases. To better illustrate the measure of performance as a function of the variance ratio ρ , we performed simulations with fixed $m = 5$ and $m = 8$. Figures 30 and 31 show the MSE of the respective estimators as a function of ρ . We observe that for $m = 5$

Figure 29: Performance of Common Mean Estimators as a function of m for $\rho = 1$, $\rho = 2$, and $\rho = 3$.

the Chernoff estimator performs better out to $\rho \approx 4$, while for $m = 8$, it performs better to $\rho \approx 3$. Thus there is only appreciable gain in using the Chernoff approach (for small sample sizes) when the unknown variances are not too far apart. Note that the figures also show the MSE of the individual sample mean $\hat{\mu}_0$ calculated from the sample with smaller true variance, σ_0^2 . It is well known that the GD estimator will not outperform the better of the individual sample mean estimators for all ρ , when $m < 9$ [18].

Viewing the previous MSE results it is clear that dynamically choosing the combined estimate to correspond to the mean of the equidistant PDF, within the EEF connecting the estimated sample mean PDFs does have some advantage, but why do the benefits diminish as ρ increases? To answer this we expand our analysis / simulation beyond the relationship of the Chernoff and Centroid estimators. First we look at how the PDF of the Chernoff embedding parameter, $\hat{\eta}_*$, is affected by the

Figure 30: Performance of Common Mean Estimators $m = 5$

change in variance ratio, ρ . We note that in the unknown but equal variance case, the results of our simulations show that the expectation of the Chernoff embedding parameter is $\frac{1}{2}$, with a PDF symmetric about $\frac{1}{2}$, becoming more concentrated as m increases. In Figure 32 we plot the PDF of $\hat{\eta}_*$ for $\rho = 1$, $\rho = 2$, and $\rho = 3$, respectively, for fixed $m = 5$. The PDFs are estimated by applying the kernel density estimator (function KSDENSITY) in MATLAB to $\hat{\eta}_*$ calculated from each Monte Carlo iteration. Clearly for $\rho = 1$ the PDF is non-Gaussian and symmetric about $\frac{1}{2}$. However as ρ increases it becomes asymmetric and shifts to the right (higher weighting for p_1). Thus we observe that as the variance ratio ρ increases, the expected value of the Chernoff embedding parameter increases. This means that the weight corresponding to equi-distance is higher for the sample mean estimate $\hat{\mu}_1$ from density $p_1(x)$, the higher variance sample. To view how this affects MSE performance of the common mean, consider as an alternative to Chernoff and Centroid, all other fixed choices of η , $\hat{\mu}_\eta$ from (178) where we utilize

Figure 31: Performance of Common Mean Estimators $m = 8$

the estimated means and variance. In our simulation we calculate the MSE of $\hat{\mu}_\eta$ for all fixed embedding parameter choices between 0 and 1. In Figure 33, we plot this MSE performance as a function of η for the three different variance ratios, ρ . We also plot the MSE of the Chernoff approach in each case (constant lines). We observe that as ρ is increased, $\hat{\eta}_*$ increases (from Figure 32), but the optimal fixed choice (minimum of each solid curve marked by *) for η decreases. As a result the best fixed weight choice adds weight to the lower variance estimate $\hat{\mu}_0$ derived from the samples of p_0 (the smaller variance sample). This result is more intuitive. That said, we note that despite the increased weight on the higher variance sample, the Chernoff common mean estimate outperforms all fixed weighting choices for $\hat{\mu}_\eta$ up until about $\rho = 3$, when the existence of better performing fixed weighting choices emerge. This implies that a criteria that is a function of the data has merit but in this case also has limited utility.

Figure 32: Kernel estimates for PDF of $\hat{\eta}_*$ for increasing ρ

3.5 Conclusion

In this work we have shown that the classical Graybill-Deal estimator of the common mean from two normal populations with different, unknown variances is identical to choosing the centroid or the geometric mean of the PDFs of the independent sample means. As an alternative we have introduced the Chernoff common mean estimator and shown through numerical simulations that this estimator performs better than the classical estimator under certain conditions, primarily for low number of samples and relatively close variances. This work contributes to understanding the implications of choosing these particular criteria for combining estimates in the independent sources case. This insight should help direct further study in the case of dependent sources.

Figure 33: Performance of all fixed parameter combined mean estimators vs. Chernoff for fixed $m = 5$ and increasing ρ .

Appendix 3A - Additional Results in using the EEF for Combining Estimates

In Manuscript 3 we have applied the EEF paradigm to the problem of combining independent, unbiased estimates of mean from separate normal populations. These populations are assumed to have a common mean, but have (possibly) different variance. This work was motivated by questions which arise in the broader application of this paradigm in problems such as combining expert opinions [1], and combining state estimates in tracking applications [8]. In both of these applications the estimates being combined are in the form of PDFs, where the component PDFs to be combined are estimated from separate populations derived from the same underlying source. Thus the estimates may have some form of interdependency (redundant data, common prior, etc.). This is a difficult problem for which the EEF provides a sub-optimal solution, however the EEF is utilized because it

provides a conservative combined estimate that does not double count information [13]. In contrast, naive Bayes approaches, which assume PDFs derived from purely independent sources are well known to underestimate the level of uncertainty of the combined estimates [24]. In this appendix we construct some simple examples to gain more insight on the EEF as a method to combine PDF estimates, and to obtain some further numerical and analytical results on the convergent relationship between the Chernoff and Centroid approaches to weighting the information.

3A.1 Combining PDF estimates

As we observed in applying the EEF to the common mean problem, when the PDFs to be combined are Gaussian we reproduce a Gaussian family parametrized by η . In the common mean problem we focused on the choice of η and its effect on the MSE performance of the combined estimates of the common mean of two Gaussian PDFs. In the more general application we seek the best fusion of two estimated PDFs. In particular, we will compare the Centroid and the Chernoff equidistant PDF as choices for combining two local PDF estimates into a combined PDF estimate. Since this is a PDF estimation problem we will adopt an information divergence measure, namely KL-divergence, to compare relative performance of combined PDF estimates.

Define $p(x)$, $x \in \mathbb{R}$ as a probability density function. Consider N independent and identically distributed (i.i.d.) samples $x_1, x_2, \dots, x_N \in \mathcal{X}$ from $p(x)$. Now consider the formation of two sub-collections of the N i.i.d. samples, denoted as S_0 and S_1 where both are subsets of $\mathcal{X}$ and we assume that $S_0 \cup S_1 = \mathcal{X}$ and the cardinality of both sets are equal, i.e. $|S_0| = |S_1| = m$. The intersection of these sets may or may not be empty. That is, it is possible that sets S_1 and S_2 contain some common data. This model introduces a simple dependency between the samples in sets S_0 , and S_1 known as double counting. That is any function

Figure 34: Double Counting example. Note that the central partition contains data samples common to both sets S_0 and S_1 .

($f : \mathbb{R}^m \mapsto \mathbb{R}$) of S_0 and S_1 results in two random variables which are dependent through the effect of the redundant (common) data samples. See Figure 34 for an illustration. The problem of interest is to estimate the PDF $\hat{p}(x)$ without direct access to the data. In lieu of data we have the PDF estimates $\hat{p}_0(x)$, $\hat{p}_1(x)$, estimated from the data in sets S_0 and S_1 , respectively. Note that since we do not have access to the data, we cannot determine and / or account for any common data samples used in estimating $\hat{p}_0$ and $\hat{p}_1$. We form the EEF of the local PDF estimates in order to combine $\hat{p}_0$ and $\hat{p}_1$ to produce an estimate $\hat{p}(x)$ which does not double count the same information. The combined PDF estimate then becomes

$$\hat{p}(x; \eta) = \exp \left(\eta \log \frac{\hat{p}_1(x)}{\hat{p}_0(x)} + \log \hat{p}_0(x) - \mathcal{K}_0(\eta) \right) \quad (192)$$

As a measure of performance we will calculate

$$D(\hat{p}(x; \eta) \| p(x)) = \int \hat{p}(x; \eta) \log \frac{\hat{p}(x; \eta)}{p(x)} dx \quad (193)$$

the KL-divergence of the combined estimate relative to the true PDF.

3A.1.1 A simple numerical experiment

Assume the underlying PDF is Gaussian. Then the problem becomes to estimate the individual mean and variance of the local PDFs and then combine in some way. In this case the previously derived mean estimators for the Centroid (181), and Chernoff (189) can be utilized with corresponding variance components included to define the PDF. Assuming a Gaussian form transforms the combining PDFs problem into a combining of parameter estimates problem, but in the interest of gaining insight into the Centroid and Chernoff approaches to combining PDFs we will ignore practical considerations and treat this as a PDF estimation problem. The most common application is to combine Gaussian mixture models [25] but can only be solved approximately.

Figure 35: Monte Carlo Results of Pairwise Chernoff Fusion under double counting.

Given the above assumptions, we perform the following Monte Carlo experi-

Figure 36: Kullback Leibler divergence of Chernoff fusion PDF relative to Centroid: No overlap

ment consisting of 10,000 independent realizations of $N = 60$ i.i.d samples from a uni-variate Gaussian $\mathcal{N}(2, 4)$ where $m = |S_0| = |S_1| = 40$. Thus on average the two sub-collections have twenty samples that are identical, i.e. double counted. Figure 35 shows the result, where the vertical axis is the Monte Carlo average of KL-divergence of the local PDFs estimated via maximum likelihood ($\text{mle}_1, \text{mle}_2$), as well as the MLE using all the data points, mle_t , (which is asymptotically optimal). The dashed curve is the result associated with fixed choices of the embedding parameter (η) and the solid blue line below it shows the average error when using $\hat{\eta}_*$, corresponding to the Chernoff criteria. This simulation result suggests that the Chernoff fused PDF performs best among all other embedding parameter choices. Further experiments have shown that as m gets large, the Chernoff and Centroid PDFs coincide. That is, the variance of $\hat{\eta}_*$ (centered at 0.5) decreases asymptotically with m for a constant level of overlap / double counting. For example, see

figure 36 for the data length dependence for no overlap. This relationship between the Chernoff Fusion PDF and the Centroid is the focus of the rest of this appendix. In particular, we want to understand the convergence properties with and without overlap, and quantify the performance gains of the Chernoff fusion relative to the Centroid for small samples.

3A.2 A simplified example: Combining variance estimates

In both the combined mean estimate and the combined PDF problem, the statistics of the Chernoff embedding parameter are analytically intractable. To simplify the analysis, suppose that we take the pairwise univariate Gaussian problem above and assume that we have a known mean. Thus $\hat{\mu}_1 = \hat{\mu}_0 = \mu$ and the sample variance is then

$$\hat{\sigma}_i^2 = \frac{1}{m} (\mathbf{x}_i - \mu)^T (\mathbf{x}_i - \mu) \quad (194)$$

Now when the two estimates are combined by the EEF, then $p_\eta \equiv \mathcal{N}(\mu, \sigma^2(\eta))$ where the fused variance as a function of η is

$$\hat{\sigma}^2(\eta) = \frac{\hat{\sigma}_1^2 \hat{\sigma}_0^2}{\eta(\hat{\sigma}_0^2 - \hat{\sigma}_1^2) + \hat{\sigma}_1^2} \quad (195)$$

In general when the sample variances are derived from i.i.d. populations with no common data (i.e. no dependencies) this is a nonsensical approach, since the average of the variance estimates will perform better. That said, when there are potential dependencies this approach can improve over the individual estimates in a conservative fashion [13]. Recalling that the mid-point density is equivalently defined by

$$\eta_* = \arg \min_{\eta} \mathcal{K}_0(\eta) \quad (196)$$

we can solve in closed form as

$$\begin{aligned}\hat{\eta}_* &= \frac{1}{\ln\left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_1^2}\right)} + \frac{\hat{\sigma}_1^2}{\hat{\sigma}_1^2 - \hat{\sigma}_0^2} \\ &= \frac{\hat{\sigma}_1^2 \ln \frac{\hat{\sigma}_0^2}{\hat{\sigma}_1^2} + \hat{\sigma}_1^2 - \hat{\sigma}_0^2}{(\hat{\sigma}_1^2 - \hat{\sigma}_0^2) \ln \frac{\hat{\sigma}_0^2}{\hat{\sigma}_1^2}}\end{aligned}\quad (197)$$

Now let $\rho = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_0^2}$ then

$$\hat{\eta}_* = g(\rho) = \frac{\rho - 1 - \rho \ln \rho}{(1 - \rho) \ln \rho} \quad (198)$$

Note that this function is an indeterminate form for $\rho = 1$, which occurs for example when $\hat{\sigma}_0^2 = \sigma^2, \hat{\sigma}_1^2 = \sigma^2$. Subsequently, if we take the limit as ρ goes to 1 we have

$$\lim_{\rho \rightarrow 1} \hat{\eta}_* = \frac{1}{2} \quad (199)$$

by applying L'Hopital's rule twice. However, it is not that simple, since ρ (and thus $\hat{\eta}_*$) is a function of two random variables. Thus we must show that $\hat{\eta}_*$ converges in probability to $\frac{1}{2}$ if ρ converges in probability to 1. Indeed, since $g(\rho)$ is continuous and ρ converges in probability to 1 (since the sample variances converge in probability to σ^2), we can prove this by taking the Taylor expansion [26] of $g(\rho)$ and show that $g(1)$ converges in probability to $\frac{1}{2}$. That is

$$g(\rho) \xrightarrow{p} \frac{1}{2} \quad (200)$$

Proof. We must prove that

$$g(\rho) = \frac{\rho - 1 - \rho \ln \rho}{(1 - \rho) \ln \rho} \quad (201)$$

converges in probability to $\frac{1}{2}$ when ρ converges in probability to one. Indeed let

$$g(\rho) = \frac{f(\rho)}{h(\rho)} \quad (202)$$

and find the Taylor expansions of numerator and denominator about $\rho = 1$ as

$$\begin{aligned}
f(\rho) &= f(1) + f'(1)(\rho - 1) + f'' \frac{(\rho - 1)^2}{2} + O_p((\rho - 1)^2) \\
&= 0 + (-\ln \rho)|_{\rho=1}(\rho - 1) - \frac{1}{2\rho}|_{\rho=1}(\rho - 1)^2 + O_p((\rho - 1)^2) \\
&= -\frac{1}{2}(\rho - 1)^2 + O_p((\rho - 1)^2)
\end{aligned}$$

where $O_p(\cdot)$ means asymptotically bounded in probability [27].

$$\begin{aligned}
h(\rho) &= h(1) + h'(1)(\rho - 1) + h'' \frac{(\rho - 1)^2}{2} + O_p((\rho - 1)^2) \\
&= 0 + \left(-\ln \rho + \frac{1 - \rho}{\rho}\right)|_{\rho=1}(\rho - 1) - \left(\frac{\rho + 1}{\rho^2}\right)|_{\rho=1}((\rho - 1)^2) + \\
&\quad O_p((\rho - 1)^2)
\end{aligned}$$

$$h(\rho) = -(\rho - 1)^2 + O_p((\rho - 1)^2) \quad (203)$$

Then $g(\rho)$ becomes

$$g(\rho) = \frac{1}{2} + O_p((\rho - 1)^2) \quad (204)$$

Thus $g(\rho)$ converges in probability to $\frac{1}{2}$. $\square$

In addition, since $g(\rho)$ is bounded, then $g(\rho) \xrightarrow{p} \frac{1}{2}$ implies that

$$E\{g(\rho)\} = \frac{1}{2} \quad (205)$$

$$Var\{g(\rho)\} \rightarrow 0 \quad \text{as} \quad m \rightarrow \infty \quad (206)$$

Also since η_* is the MLE of p_{η_*} then it is asymptotically Gaussian, with a variance that vanishes as m gets large. Thus it is not unexpected that the mid-point, i.e. the Chernoff point coincides with the centroid ($\eta = \frac{1}{2}$) as m goes to infinity. The fused variance associated with the Chernoff point in this problem can be written as

$$\hat{\sigma}^2(\hat{\eta}^*) = \frac{\hat{\sigma}_0^2 \hat{\sigma}_1^2 \ln \frac{\hat{\sigma}_1^2}{\hat{\sigma}_0^2}}{\hat{\sigma}_1^2 - \hat{\sigma}_0^2} \quad (207)$$

while the centroid is written as

$$\hat{\sigma}^2(1/2) = \frac{2\hat{\sigma}_0^2\hat{\sigma}_1^2}{\hat{\sigma}_1^2 + \hat{\sigma}_0^2} \quad (208)$$

In general solving the transformation problem for $g(\rho)$, that is the PDF of η_* , could be a daunting task, but could help us better understand this relationship.

3A.3 PDF of $g(\rho)$

From our numerical comparison of the Chernoff point and the Centroid as potential fusion choices, it is clear that the two coincide when the data records are large or if the underlying data are redundant. It will be useful to characterize the PDF of the Chernoff embedding parameter for combining both independent and dependent estimates.

3A.3.1 No Overlap: Independent estimates

Figure 37: PDF of Chernoff embedding parameter (no overlap): $m = 2$

Figure 38: PDF of Chernoff embedding parameter (no overlap): $m = 5$.

In the no overlap case, the PDF of $\rho = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_0^2}$, as the ratio of two independent χ^2 random variables (with m degrees of freedom), is well known and given as an F -distribution. That is F_{mm} to designate the equal degrees of freedom. Then to find the PDF of $y = g(\rho)$ (an invertible function of ρ) we have

$$p_y(y) = f_\rho(g^{-1}(y)) \left| \frac{dg^{-1}(y)}{dy} \right| \quad (209)$$

where

$$f_\rho(\rho; m, m) = \frac{1}{B\left(\frac{m}{2}, \frac{m}{2}\right)} \rho^{m/2-1} (1 + \rho)^{-m} \quad (210)$$

$\rho \in (0, \infty)$, and $B(\cdot, \cdot)$ is the Beta function. The inverse function of $g(\rho)$ does not have an analytical solution, but it is 1-1 and bounded. Thus we can solve (209) numerically to obtain the PDF of y . Figures 37 and 38 show the numerically generated PDF of $g(\rho)$ in the no overlap (or independent) case, for $m = 2$, $m = 5$, respectively. Note that in both cases there is exact agreement with the PDF formed by simulation. In the $m = 2$ case, it is seen that the PDF is nearly Gaussian, while

in the $m = 5$ case it is nearly exact. Thus the transformation $g(\rho)$ is approximately a normalizing transformation of an F distributed random variable (for $m > 3$). Figures 39-40 show the cumulants of $g(\rho)$ as a function of m . Note that the third and fourth cumulants go to zero as expected for a Gaussian distributed PDF, and do so very rapidly.

Figure 39: First two cumulants of $g(\rho)$.

Figure 40: Third and Fourth cumulants of $g(\rho)$.

3A.3.2 Overlap Case

When there is common data between sensors (i.e. overlap), then the PDF of ρ is no longer an F distribution. Thus we want to find the probability density function of the ratio of two variance estimates where the data samples are from the same i.i.d. Gaussian process, but there are some common data samples which

contribute to the two local estimates. Let

$$\rho = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_0^2} \quad (211)$$

where $\hat{\sigma}_0^2, \hat{\sigma}_1^2$ are estimates formed from length m data vectors $\mathbf{x}_0, \mathbf{x}_1$ with $t < m$ common data samples. If we assume that the mean is known, then it is well known that $\frac{m\hat{\sigma}_i^2}{\sigma^2}$ is distributed as chi-squared with m degrees of freedom. Now we can re-write ρ as

$$\rho = \frac{x_{1s} + y_s}{x_{0s} + y_s} \quad (212)$$

where x_{1s} and x_{0s} are the sums of squares of the unique samples from $\mathbf{x}_0, \mathbf{x}_1$, respectively. While y_s is the sum of squares of the common samples. Note that x_{0s} and x_{1s} are then chi-squared with $m - t$ degrees of freedom, and y_s is chi-squared with t degrees of freedom. This is true because the scale terms cancel due to the

Figure 41: Simulation vs Integration for the density of ρ given common samples.

identical data record lengths, and the fact that the samples come from the same random process (identical σ^2). Let

$$u = x_{1s} + y_s \quad (213)$$

Figure 42: PDF of $g(\rho)$ with overlap; $m = 8$, $t = 4$.

$$v = x_{0s} + y_s \quad (214)$$

and assume that $y_s = c$ is fixed. Then $\rho = \frac{u}{v}$ and to find the PDF of ρ , $f_\rho(\rho)$ we must first find the PDFs of u, v (conditioned on c). Indeed these PDFs are just shifted chi-squared functions. That is,

$$f_U(u|c) = f_{x_{1s}}(u - c) = \frac{1}{2^{k/2}\Gamma(\frac{k}{2})} (u - c)^{k/2-1} \exp\left(-\left(\frac{u - c}{2}\right)\right) u_s(u - c) \quad (215)$$

where $u_s(\cdot)$ is the unit step function. Similarly,

$$f_V(v|c) = f_{x_{1s}}(v - c) = \frac{1}{2^{k/2}\Gamma(\frac{k}{2})} (v - c)^{k/2-1} \exp\left(-\left(\frac{v - c}{2}\right)\right) u_s(v - c) \quad (216)$$

where the PDFs are conditioned on c a constant and $k = m - t$ is the degrees of freedom parameter. Now the transformation of a quotient $\rho = \frac{u}{v}$ has a well known form. We introduce the auxiliary variable $w = v$ so that

$$u = \rho w$$

$$v = w$$

The Jacobian is then just w , so that we can write (at a high level) the joint PDF of ρ, w (conditioned on c) as

$$f_{\rho, W}(\rho, w|c) = f_U(\rho w)u_s(\rho w - c) \cdot f_V(w)u_s(w - c) \cdot |w| \quad (217)$$

We then obtain the PDF $f_\rho(\rho)$ as

$$f_\rho(\rho) = \int_0^\infty \int_c^\infty f_{\rho, W}(\rho, w|c) f_c(c) d_w d_c \quad (218)$$

The result of this numerical integration compared to simulation for $m = 6, t = 2$ is shown in figure 41. Note that the integration matches exactly with the simulation. With the formulation for the PDF of ρ given, we can then use the numerical inversion approach for finding the PDF of $g(\rho)$. Figure 42 shows an example with $m = 8, t = 4$. That is, when we have two eight point data records with four samples in common (50 percent overlap). Note that the PDF is now clearly non-Gaussian but has higher concentration about $\frac{1}{2}$, thus is overall closer to the Centroid.

3A.4 Conclusions

We have shown that when using the EEF to combine local PDF estimates the best criteria to apply is due to Chernoff. In particular, for small samples and no overlap, there is appreciable gain using the KL-divergence from the truth as the performance measure. As the data records get large the Chernoff and Centroid fusion approaches coincide, independent of overlap. With increasing overlap, the PDF of the Chernoff embedding parameter becomes more and more non-Gaussian, and becomes sharply peaked about $\frac{1}{2}$, coinciding with the Centroid. Thus the performance gain of Chernoff decreases both as data becomes more available, and when there is more common data. Therefore in many applications the Centroid may be a better choice for fusion, as there is less computation required. Note that

all these results assume that the local estimates are based on the same amount of data. For the more general case more details remain to be studied and understood.

List of References

- [1] R. T. Clemen and R. L. Winkler, “Combining probability distributions from experts in risk analysis,” *Risk analysis*, vol. 19, no. 2, pp. 187–203, 1999.
- [2] R. A. Jacobs, “Methods for combining experts’ probability assessments,” *Neural computation*, vol. 7, no. 5, pp. 867–888, 1995.
- [3] N. C. Lind and A. S. Nowak, “Pooling expert opinions on probability distributions,” *Journal of engineering mechanics*, vol. 114, no. 2, pp. 328–341, 1988.
- [4] G. E. Hinton, “Products of experts,” in *Artificial Neural Networks, 1999. ICANN 99. Ninth International Conference on (Conf. Publ. No. 470)*, vol. 1. IET, 1999, pp. 1–6.
- [5] M. P. Clements and D. I. Harvey, “Combining probability forecasts,” *International Journal of Forecasting*, vol. 27, no. 2, pp. 208–223, 2011.
- [6] M. Rufo, J. Martin, C. Pérez, *et al.*, “Log-linear pool to combine prior distributions: A suggestion for a calibration-based approach,” *Bayesian Analysis*, vol. 7, no. 2, pp. 411–438, 2012.
- [7] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, “Wireless sensor networks: a survey,” *Computer networks*, vol. 38, no. 4, pp. 393–422, 2002.
- [8] Y. Wang and X. R. Li, “Distributed estimation fusion with unavailable cross-correlation,” *Aerospace and Electronic Systems, IEEE Transactions on*, vol. 48, no. 1, pp. 259–278, 2012.
- [9] D. M. Tax, M. Van Breukelen, R. P. Duin, and J. Kittler, “Combining multiple classifiers by averaging or by multiplying?” *Pattern recognition*, vol. 33, no. 9, pp. 1475–1485, 2000.
- [10] R. Viswanathan and P. K. Varshney, “Distributed detection with multiple sensors i. fundamentals,” *Proceedings of the IEEE*, vol. 85, no. 1, pp. 54–63, 1997.
- [11] S. Kay, “Exponentially embedded families-new approaches to model order estimation,” *Aerospace and Electronic Systems, IEEE Transactions on*, vol. 41, no. 1, pp. 333–345, 2005.

- [12] M. B. Hurley, “An Information-Theoretic justification for covariance Intersection and its generalization,” in *Proceedings of the Fifth International Conference on Information Fusion*, vol. 1, 2002, pp. 505–511.
- [13] M. Uney, D. E. Clark, and S. J. Julier, “Distributed fusion of PHD filters via exponential mixture densities,” *Selected Topics in Signal Processing, IEEE Journal of*, vol. 7, no. 3, pp. 521–531, 2013.
- [14] T. Bailey, S. Julier, and G. Agamennoni, “On conservative fusion of information with unknown non-gaussian dependence,” in *15th International Conference on Information Fusion*. IEEE, 2012, pp. 1876–1883.
- [15] G. Battistelli and L. Chisci, “Kullback–Leibler average, consensus on probability densities, and distributed state estimation with guaranteed stability,” *Automatica*, vol. 50, no. 3, pp. 707–718, 2014.
- [16] S. Zacks, “Unbiased estimation of the common mean of two normal distributions based on small samples of equal size,” *Journal of the American statistical Association*, vol. 61, no. 314, pp. 467–476, 1966.
- [17] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall, 1993.
- [18] F. Graybill and R. Deal, “Combining unbiased estimators,” *Biometrics*, no. 15, pp. 543–550, 1959.
- [19] N. Pal, J.-J. Lin, C.-H. Chang, and S. Kumar, “A revisit to the common mean problem: Comparing the maximum likelihood estimator with the Graybill–Deal estimator,” *Computational Statistics & Data Analysis*, vol. 51, no. 12, pp. 5673–5681, 2007.
- [20] J. Ajgl and M. Simandl, “On conservativeness of posterior density fusion,” in *Information Fusion (FUSION), 2013 16th International Conference on*. IEEE, 2013, pp. 85–92.
- [21] S. Amari, “Integration of stochastic models by minimizing alpha-divergence,” *Neural Computation*, vol. 19, pp. 2780–2796, 2007.
- [22] T. M. Cover and J. A. Thomas, *Elements of information theory*, 2nd ed. John Wiley and Sons, 2006.
- [23] F. Nielsen, “An Information-Geometric characterization of Chernoff information,” *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 269–272, 2013.
- [24] K.-C. Chang, C.-Y. Chong, and S. Mori, “On scalable distributed sensor fusion,” in *Information Fusion, 2008 11th International Conference on*. IEEE, 2008, pp. 1–8.

- [25] N. R. Ahmed and M. Campbell, “Fast consistent Chernoff Fusion of Gaussian mixtures for Ad Hoc sensor networks,” *IEEE Transactions on Signal Processing*, vol. 60, no. 12, pp. 6739–6745, 2012.
- [26] P. J. Brockwell and R. A. Davis, *Time series: theory and methods*. Springer Science & Business Media, 2013.
- [27] E. L. Lehmann and G. Casella, *Theory of point estimation*. Springer, 1998.

CHAPTER 4

Future Work

The research pursued in this dissertation has a number of potential interesting future directions.

1. *Formal proof of Chernoff distance as a squared distance in global PDF space*

Other than the Fisher Information metric, defining a true metric in an information space can be elusive or ad hoc. Most information measures that have been shown to be useful in estimation and detection are directional in nature, i.e. are not symmetric, or violate the triangle inequality. In Chapter 2 we discuss that pairwise Chernoff distance behaves as a squared distance in PDF space, but are yet to prove this formally. A formal proof will allow a true information metric to be derived from Chernoff distance (by a square root transformation) which will be very useful in numerous applications.

2. *Conditional estimation along orthogonal bisectors of PDFs within Three component PDF*

In Chapter 2 we determined within the three component EEF that the distance along the orthogonal bisector of one pairwise EEF (e.g. connecting PDFs p_0 and p_1) depends only on $E[T_2|T_1]$, the conditional expectation of the sufficient statistic T_2 of the EEF (connecting p_2 and p_0) given the sufficient statistic T_1 . The length squared of this quantity was derived as the KL-divergence of component p_2 from the orthogonal bisector of the EEF of p_0 and p_1 . Note that this is a fundamental decomposition of a set of joint statistics (T_1, T_2) into its two orthogonal components. Hence the “path” in PDF space appears to be due only to the factor $T_2|T_1$, which in the EEF is the log likelihood of $\frac{p_2}{p_0}$. We are studying this further as it relates to the statistical curvature concept of Efron [1], and may have further utility in distributed estimation.

3. *The extension of Chernoff Information as a bounding measure*

Our extension of Chernoff Information results from minimizing the cumulant generating function (CGF) of the M -component EEF. We show that the negative of the CGF can be written as

$$-\mathcal{K}_0(\boldsymbol{\eta}) = \sum_{j=0}^{M-1} \eta_j D(p_{\boldsymbol{\eta}} \| p_j) \quad (219)$$

the weighted sum of the KL-divergence of the EEF to all of its component densities. Similar results exist in Amari [2] and Cencov [3]. Thus the extension of Chernoff Information is a weighted sum of KL-divergences. In [4] this is offered as a general information measure related to the Hellinger transform [5]. This extension as a measure of information requires more investigation, particularly in light of the importance of Chernoff information in general.

4. *In the EEF: uncorrelated implies independence*

Since for Gaussian we use uncorrelatedness in principal component analysis as well as other areas, this property could be very significant. This also extends the usual theoretical result, which is heavily relied upon, that decorrelating Gaussian random variables makes them independent. Note that a mixed parametrization [2] gives a diagonal Fisher information matrix which is hence uncorrelated and therefore independent.

5. *The embedding of multiple potential models*

We will investigate embedding of multiple models into EEF, of which only one or more are valid. Develop a procedure for picking significant models based on the estimated embedding parameters from data.

List of References

- [1] B. Efron, "Defining the curvature of a statistical problem (with applications to second order efficiency)," *The Annals of Statistics*, vol. 3, no. 6, pp. 1189–1242, November 1975.

- [2] S. Amari and H. Nagaoka, *Methods of Information Geometry*. Oxford University Press, 2000.
- [3] N. Cencov, *Statistical decision rules and optimal inference*. American Mathematical Society, 1982.
- [4] J. Ginebra, “On the measure of the information in a statistical experiment,” *Bayesian Analysis*, vol. 2, no. 1, pp. 167–211, 2007.
- [5] I. Fazekas and F. Liese, “Some properties of the Hellinger transform and its application in classification problems,” *Computers & Mathematics with Applications*, vol. 31, no. 8, pp. 107–116, 1996.