2-2020

Fusion of Probabilistic Knowledge as Foundation for Sliced-Normal Approach

Michael Beer

Olga Kosheleva

Vladik Kreinovich

Comments:
Technical Report: UTEP-CS-20-09

Fusion of Probabilistic Knowledge as Foundation for Sliced-Normal Approach

Michael Beer[1], Olga Kosheleva[2], and Vladik Kreinovich[3]
[1] *Institute for Risk and Reliability, Leibniz University Hannover*
30167 Hannover, Germany, beer@irz.uni-hannover.de
Departments of [2] *Teacher Education and* [3] *Computer Science*
University of Texas at El Paso
El Paso, TX 79968, USA, {olgak,vladik}@utep.edu

Abstract. In many practical applications, it turns out to be efficient to use Sliced-Normal multi-D distributions, i.e., distributions for which the logarithm of the probability density function (pdf) is a polynomial – – to be more precise, it is a sum of squares of several polynomials. This class is a natural extension of normal distributions, i.e., distributions for which the logarithm of the pdf is a quadratic polynomial.

In this paper, we provide a possible theoretical explanation for this empirical success.

Keywords: sliced-normal distribution, fusion of probabilistic knowledge

1. Formulation of the Problem

Sliced-normal distributions are efficient. In many practical applications, it turns out to be efficient to use Sliced-Normal multi-D distributions, i.e., distributions for which the logarithm of the probability density function (pdf) $\rho(x_1, \ldots, x_n)$ is a polynomial (to be more precise, it is a sum of squares of several polynomials); see, e.g., (Colbert, Crespo, and Peet, 2019; Crespo, 2019; Crespo, Colbert, Kenny, and Giesy, 2019):

$$\ln(\rho(x_1, \ldots, x_n)) = P(x_1, \ldots, x_n)$$

for some polynomial $P(x_1, \ldots, x_n)$, so

$$\rho(x_1, \ldots, x_n) = \exp(P(x_1, \ldots, x_n)).$$

This class is a natural extension of normal distributions, i.e., distributions for which the logarithm of the pdf is a quadratic polynomial; see, e.g., (Sheskin, 2011).

But why? This is what we try to explain. While the sliced-normal distributions have been empirically successful, there seems to be no convincing theoretical explanation for their empirical success. The main goal of this paper is to provide such an explanation.

2. Let Us Formulate This Problem in Precise Terms

Need for a finite-parametric family. In principle, we can have many different probability density functions. The class of all the functions is infinite-dimensional – which means that, to select a single probability density function out of all possible such functions, we need to know the values of infinitely many parameters (e.g., values of the pdf at points with rational coordinates).

In practice, however, at any given moment of time, we only have finitely many observations. Based on these observations, we can determine only finitely many parameters. Thus, it makes sense to looks for families F of probability density functions that depend on finitely many parameters $c_1, \ldots, c_m$, i.e., on families of the type $F = \{\rho(x_1, \ldots, x_n, c_1, \ldots, c_m)\}_{c_1,\ldots,c_m}$.

The dependence should be continuous. All our information about the physical world comes from measurements and from expert estimates. Measurements are never 100% accurate (see, e.g., (Rabinovich, 2005)), expert estimates are even less accurate. So, we can only determine the values x_i and c_j with some accuracy. Based on these approximate values of x_i and c_j, we should make estimates of the corresponding values $\rho(x_1, \ldots, x_n, c_1, \ldots, c_m)$ of the probability density – and the more accurately we perform measurements, the more accurate should be our estimates.

In mathematical terms, this means that the dependence of the function $\rho(x_1, \ldots, x_n, c_1, \ldots, c_m)$ on all its $n + m$ inputs x_i and c_j should be continuous.

Moreover, small inaccuracy in x_i and c_j should lead to proportionally small inaccuracy in the resulting value of $\rho(x_1, \ldots, x_n, c_1, \ldots, c_m)$. Thus, the function $\rho(x_1, \ldots, x_n, c_1, \ldots, c_m)$ should be *differentiable*. Thus, we arrive at the following definition.

Definition 1. *Let n and m be positive integers. By an m-*parametric family of probability density functions on $\mathbb{R}^n$ *(or simply a* family*, for short), we mean a differentiable function $\rho(x_1, \ldots, x_n, c_1, \ldots, c_m)$ of $n+m$ variables for which, for each tuple $(c_1, \ldots, c_m)$, the corresponding function $x_1, \ldots, x_n \to \rho(x_1, \ldots, x_n, c_1, \ldots, c_m)$ is a probability density function, i.e.:*

- *we have $\rho(x_1, \ldots, x_n, c_1, \ldots, c_m) \geq 0$ for all x_i and c_j, and*
- *we have $\int \rho(x_1, \ldots, x_n, c_1, \ldots, c_m)\, dx_1 \ldots dx_n = 1$ for all tuples $(c_1, \ldots, c_m)$.*

3. The Class of Distributions Should Be Closed Under Fusion

Need for fusion. The very fact that we only know the *probability* of different tuples $x = (x_1, \ldots, x_n)$ means that we do not know which of the tuples describe the corresponding real-life situation. In other words, the fact that we have a probabilistic knowledge means that our knowledge is incomplete. It is therefore desirable to gain additional knowledge about the situation – either by performing additional measurements, or by requesting additional expert estimates.

This additional knowledge usually comes in the form of a probability distribution. Once we have this probability distribution, we need to fuse it with the distribution describing our original knowledge.

The class of distributions should be closed under fusion. The main objective in selecting a finite-parametric family of distributions is to come up with a reasonable family, a family that

describes reasonably well all possible states of our knowledge. From this viewpoint, it is reasonable to require that:

- if both fused pieces of knowledge are described by distributions from our family,
- then the result of fusing these two pieces of knowledge should also be described by distributions from our family.

In mathematical terms, this means that the desired family of probability distributions should be *closed* under fusion.

To describe this requirement in precise terms, let us describe fusion in precise terms.

How to describe fusion: a natural idea. In probability theory, if we have two independent events with probabilities p_1 and p_2, then the probability that both events will happen is equal to the product of these probabilities. Similarly, if we have two independent sources of information, so that:

- based on the information from the first source, we assign, to each of N alternatives $a_1, \ldots, a_N$, the probabilities $p_{11}, \ldots, p_{1N}$,
- based on the information from the second source, we assign, to each of N alternatives $b_1, \ldots, b_N$, the probabilities $p_{21}, \ldots, p_{2N}$,

then the probability that we have alternative a_i in the first case and alternative b_j in the second case is equal to the product of the corresponding probabilities $p_{1i} \cdot p_{2j}$.

If it turns out that in both cases, we have the exact same set of alternatives, then we need to consider *conditional* probabilities, namely probabilities under the condition that $i = j$. In general, the conditional probability $P(A \,|\, B)$ of an event A under the condition B can be obtained by dividing the probability $P(A \,\&\, B)$ of $A \,\&\, B$ by the probability $P(B)$ that the condition B is satisfied. In our case, this means that after the fusion, the probability of the i-th alternative is equal to $p_i = C \cdot p_{1i} \cdot p_{2i}$, where the coefficient $C \stackrel{\text{def}}{=} \dfrac{1}{P(B)}$ can be obtained from the requirement that the resulting probabilities add up to 1, i.e., that $\sum\limits_{i=1}^{N} p_i = C \cdot \sum\limits_{i=1}^{N} p_{1i} \cdot p_{2i} = 1$, so that $C = \dfrac{1}{\sum\limits_{i=1}^{N} p_{1i} \cdot p_{2i}}$.

Similar formulas can be obtained for continuous distributions: if we have two independent sources of information that lead to distributions $\rho_1(x)$ and $\rho_2(x)$, then the fusion of these two pieces of information is a probability distribution $\rho(x) = C \cdot \rho_1(x) \cdot \rho_2(x)$, where C is a normalization constant selected so as to guarantee that $\int \rho(x)\, dx$, i.e., $C = \dfrac{1}{\int \rho_1(x) \cdot \rho_2(x)\, dx}$.

Similarly, we can define the result of fusing several probability distributions.

Definition 2. *Let $\rho_1(x), \ldots, \rho_k(x)$ be probability density functions (pdfs) on $\mathbb{R}^n$. By the* result of fusing *these pfds, we mean a pdf $\rho(x) = C \cdot \rho_1(x) \cdot \ldots \cdot \rho_k(x)$, where $C = \dfrac{1}{\int \rho_1(x) \cdot \ldots \cdot \rho_k(x)\, dx}$.*

Definition 3. *We say that the family* $\rho(x, c)$ *is* closed under fusion *if for every* k *pdfs* $\rho\left(x, c^{(1)}\right), \ldots, \rho\left(x, c^{(k)}\right)$ *from this family, the result of fusing these pdfs also belongs to the same family, i.e., has the form* $\rho(x, c)$ *for some tuple* c.

4. Every Piece of Knowledge Can Be Obtained by Fusing Several "Smaller" Pieces of Information

Main idea. Sometimes, knowledge comes in one big step. However, more typically, to gain the knowledge, we must acquire it piece by piece, sometimes in two steps, sometimes in three steps, sometimes in four steps, etc. So, it is natural to come up with the following definition.

Definition 4. *We say that in a family* $\rho(x, c)$, every piece of knowledge can be obtained by fusing small pieces of information *if for every pdf* $\rho(x, c)$ *from this family and for every integer* $M \geq 2$, *there exists another pdf* $\rho(x, c')$ *from this family so that fusing* M *copies of* $\rho(x, c')$ *leads to* $\rho(x, c)$.

5. The Family of Distributions Should Not Depend on the Choices of Starting Points and Measuring Units for x_i

Possibility to change measuring unit and a starting point. We want to deal with physical quantities, but in reality, we deal with their numerical values. These numerical values depend on what measuring unit we use for measuring the quantity, and what starting point we select for this measurement. When we change the measuring unit and/or the starting point, the numerical values change.

For example, if we change the measuring unit from meters to centimeters, all the numerical values are multiplied by 100, so that, 2 m becomes 200 cm. In general, if we change from the original measuring unit to a new one which is a times smaller, then all the numerical values are multiplied by a: $x \to a \cdot x$. This transformation is known as *scaling*.

Similarly, if we change the starting point to the one which is b units before – as we can do for time, temperature, and many other quantities – then b is added to all the numerical values $x \to x + b$. This transformation is known as *shift*. A shift can also be viewed as a kind of re-scaling.

If we change both the measuring unit and the starting point, then numerical value change as $x \cdot a \cdot x + b$. These transformations change the pdf: if we had a pdf $\rho(x_1, \ldots, x_n)$, and we apply such transformation $x_i \to x_i' \stackrel{\text{def}}{=} a_i \cdot x_i + b_i$ to each of inputs, then in terms of the new numerical values $x_1', \ldots, x_n'$, the corresponding pdf takes a different form.

Definition 5. *Let* $\rho(x_1, \ldots, x_n)$ *be a pdf, let* $a = (a_1, \ldots, a_n)$ *be a tuple of positive numbers, and let* $b = (b_1, \ldots, b_n)$ *be a tuple of real numbers. By a* (a, b)-re-scaling *of the pdf* ρ, *we mean a pdf*

$$\rho'(x_1', \ldots, x_n') = \frac{1}{\prod\limits_{i=1}^{n} a_i} \cdot \rho\left(\frac{x_1' - b_1}{a_1}, \ldots, \frac{x_n' - b_n}{a_n}\right).$$

A natural invariance requirement. We want to come up with a universal family of probability distributions, a family that would be applicable no matter what measuring units and what starting points we select for all the inputs. Thus, it is reasonable to require that our family is invariant with respect to the corresponding transformations.

Definition 6. *We say that a family F is* scale- and shift-invariant *if every pdf $\rho(x, c)$ from this family and for every two tuples a and b, the (a, b)-re-scaling of the pdf $\rho(x, c)$ also belongs to the family F.*

Now, we are ready for formulate and prove our main result.

6. Main Result

Proposition. *For every family F:*

- *which is closed under fusion,*
- *for which every piece of knowledge can be obtained by fusing small pieces of information, and*
- *which is scale- and shift-invariant,*

there exists an integer $d \le m + 1$ such that every probability density function from this family has the form $\rho(x, c) = \exp(P(x_1, \ldots, x_n))$ for some polynomial $P(x_1, \ldots, x_n)$ of degree $\le d$ with respect to each of its variables.

Comment. This result explains the empirical success of sliced-normal distributions.

Proof.

1°. Let F be the family that satisfies all the conditions described in the formulation of the Proposition. By a *log-function*, we will mean a function of the type $L(x, c, s) = \ln(\rho(x, c)) + s$ for some tuple c and some real number s. Let us denote the class of all log-functions by $\mathcal{L}$.

2°. Let us prove that the class of all log-functions is closed under addition, i.e., that for every two log-functions $L(x, c;, s')$ and $L(x, c'', s'')$, their sum is also a log-function.

Indeed, by definition, $L(x, c', s') = \ln(\rho(x, c')) + s'$ and $L(x, c'', s'') = \ln(\rho(x, c'')) + s''$. Since the family F is closed under fusion, the result of fusing the corresponding pdfs is also a pdf from the same family, i.e., $C \cdot \rho(x, c') \cdot \rho(x, c'') = \rho(x, c)$ for some tuple c. By taking logarithms of both sides of this equality, we conclude that

$$\ln(C) + \ln(\rho(x, c')) + \ln(\rho(x, c'')) = \ln(\rho(x, c)).$$

If we add $s' + s'' - \ln(C)$ to both sides of the resulting equality, we conclude that

$$(\ln(\rho(x, c')) + s') + (\ln(\rho(x, c'')) + s'') = \ln(\rho(x, c)) + (s' + s'' - \ln(C)),$$

i.e., that the sum of the two given log-functions is indeed a log-function:

$$L(x, c', s') + L(x, c'', s'') = L(x, c, s' + s'' - \ln(C)).$$

The statement is proven.

3°. Let us now prove that for each log-function $L(x, c, s)$ and for every integer $M \geq 2$, the function $M^{-1} \cdot L(x, c, s)$ is also a log-function.

By definition, $L(x, c, s) = \ln(\rho(x, c)) + s$. Since for the family F, every piece of knowledge can be obtained by fusing small pieces of information, we conclude that the pdf $\rho(x, c)$ can be obtained by fusing M instances of some other pdf $\rho(x, c')$, i.e., that $\rho(x, c) = C \cdot (\rho(x, c'))^M$. By taking logarithms of both sides of this equality, we get $\ln(\rho(x, c)) = M \cdot \ln(\rho(x, c')) + \ln(C)$, thus

$$M^{-1} \cdot \ln(\rho(x, c)) = \ln(\rho(x, c')) + M^{-1} \cdot \ln(C).$$

By adding $M^{-1} \cdot s$ to both sides, we get

$$M^{-1} \cdot (\ln(\rho(x, c)) + s) = \ln(\rho(x, c')) + M^{-1} \cdot (\ln(C) + s).$$

The left-hand side of this formula is exactly $M^{-1} \cdot L(x, c, s)$, and the right-hand side is a log-function. So, the statement is proven.

4°. Let us now consider the closure $\mathcal{C}$ of the set $\mathcal{L}$ of all log-functions – the closure in the usual topological sense, i.e., the set of all limit functions with respect to some natural topology on the class of all differentiable functions. Since the set $\mathcal{L}$ is closed under addition, its closure $\mathcal{C}$ is also closed under addition.

Let us prove that this closure is closed under multiplication by positive numbers. In other words, let us prove that for each function $f(x) \in \mathcal{C}$ and for every positive real number $r > 0$, the function $r \cdot f(x)$ also belongs to $\mathcal{C}$. Since $\mathcal{C}$ is the closure of the set of all log-functions, it is sufficient to prove that for each log-function $L(x, c, s)$ and for every positive real number $r > 0$, the function $r \cdot L(x, c, s)$ is a limit of log-functions.

Indeed, for every possible accuracy $\varepsilon > 0$, we can approximate, with this accuracy, the real number r by a rational number $\dfrac{N}{M}$. By Part 3 of this proof, the function $M^{-1} \cdot L(x, c, s)$ is also a log-function. Now, by Part 2 of this proof, the function $\dfrac{N}{M} \cdot L(x, c, s)$ is also a log-function – as the sum of N log-functions $M^{-1} \cdot L(x, c, s)$. When $\dfrac{N}{M}$ tends to r, the corresponding function $\dfrac{N}{M} \cdot L(x, c, s)$ tends to $r \cdot L(x, c, s)$. Thus, the function $r \cdot L(x, c, s)$ is indeed a limit of log-functions. The statement is proven.

5°. By combining Parts 2 and 4, we conclude that for every finite set of functions $C_1(x), \ldots, C_k(x)$ from the set $\mathcal{C}$, and for every tuple of positive numbers $r_1, \ldots, r_k$, the linear combination

$$r_1 \cdot C_1(x) + \ldots + r_k \cdot C_k(x)$$

also belongs to $\mathcal{L}$.

6°. Let us now prove that the set $\mathcal{C}$ cannot contain more than $m+1$ linearly independent functions.

Indeed, if this was the case, and we would have more than $m+1$ linearly independent functions, then we would have at least $m+2$ of them $C_1(x), \ldots, C_{m+2}(x)$ in the class $\mathcal{C}$. Then, due to Part 5 of this proof, the class $\mathcal{C}$ will contain a $(m+2)$-parametric family of functions

$$r_1 \cdot C_1(x) + \ldots + r_{m+2} \cdot C_{m+2}(x)$$

of different functions. However, the class $\mathcal{C}$ is the closure of the class $\mathcal{L}$ of functions of the type $\ln(\rho(x, c)) + s$ that depend on $m+1$ parameters:

- we have m parameters $c_1, \ldots, c_m$ and
- we have an additional parameter s.

So, the closure of this set is also of dimension $m+1$ (or less) – and thus, cannot contain more-dimensional subfamilies. The statement is proven.

7°. Let us denote by $\mathcal{S}$ the class of all linear combinations of functions from the class $\mathcal{C}$. Clearly, $\mathcal{C} \subseteq \mathcal{S}$, and, due to Part 6, the dimension d of the linear space $\mathcal{S}$ cannot exceed $m+1$. So, if we pick any basis $e_1(x), \ldots, e_d(x)$ in this class, then each function $f(x)$ from the class $\mathcal{S}$ can be represented as a linear combination of functions from this basis: $f(x) = C_1 \cdot e_1(x) + \ldots + C_d \cdot e_d(x)$, for some values $C_1, \ldots, C_d$.

We can pick the basis from the set $\mathcal{C}$. Moreover, since the closure does not change the dimension, we can pick it from the original class $\mathcal{L}$ of log-functions. All the pdf functions from the family F are, by definition of a family, differentiable. Thus, every log-function is also differentiable. Hence, we can choose the basis of differentiable functions.

8°. Let us prove that the class $\mathcal{L}$ is closed under arbitrary re-scalings, i.e., if a function $f(x_1, \ldots, x_n)$ is in this class, then for each tuple $a = (a_1, \ldots, a_n)$ of positive numbers and for each tuple $b = (b_1, \ldots, b_n)$ of real numbers, the function $f(a \cdot x_1 + b_1, \ldots, a_n \cdot x_n + b_n)$ also belongs to the class $\mathcal{L}$.

This follows from the requirement that the family F is scale- and shift-invariant, if we take logarithms of both sides and add appropriate constants to both sides.

9°. From Part 8, we can conclude that the closure class $\mathcal{C}$ is also invariant with respect to arbitrary re-scalings. Thus, the class $\mathcal{S}$ of all linear combinations of functions from $\mathcal{C}$ is also thus invariant.

10°. Let us first study the consequences of shift-invariance of the class $\mathcal{S}$ with respect to the first variable. This shift-invariance implies, in particular, that for each basis function $e_i(x_1, x_2, \ldots, x_n)$, the result of its shift $e_i(x_1 + b_1, x_2, \ldots, x_n)$ is also a function from the class $\mathcal{S}$, i.e., that

$$e_i(x_1 + b_1, x_2, \ldots, x_n) = \sum_{j=1}^{d} C_{ij}(b_1) \cdot e_j(x_1, x_2, \ldots, x_n),$$

for some coefficients C_{ij} that, in general, depend on b_1.

For a while, let us fix the values $x_2, \ldots, x_n$ and only consider the dependence on x_1. In other words, let us consider auxiliary functions $E_i(x_1) \stackrel{\text{def}}{=} e_i(x_1, x_2, \ldots, x_n)$. Foe these auxiliary functions, the above formula takes the form

$$E_1(x_1 + b_1) = C_{11}(b_1) \cdot E_1(x_1) + \ldots + C_{1d}(b_1) \cdot E_d(x_1);$$

$$\dots$$

$$E_d(x_1 + b_1) = C_{d1}(b_1) \cdot E_1(x_1) + \ldots + C_{dd}(b_1) \cdot E_d(x_1);$$

Here, all the functions $E_1(x_1) \ldots, E_d(x_1)$ are differentiable – since they come by fixing some values from the basis functions $e_i(x_1, \ldots, x_n)$, and the basis functions are differentiable.

Let us prove that the dependencies $C_{ij}(b_1)$ are also differentiable. Indeed, for each i, let us pick d different values $x_{11}, \ldots, x_{1d}$ of x_1, then we get the following d linear equations for d unknowns $C_{i1}(b_1), \ldots, C_{in}(b_1)$:

$$E_i(x_{11} + b_1) = C_{i1}(b_1) \cdot E_1(x_{11}) + \ldots + C_{id}(b_1) \cdot E_d(x_{11});$$

$$\dots$$

$$E_i(x_{1d} + b_1) = C_{i1}(b_1) \cdot E_1(x_{1d}) + \ldots + C_{id}(b_1) \cdot E_d(x_{1d}).$$

Each element $C_{ij}(b_1)$ solution to a system of linear equations can be described, by using the Cramer rule, as the ratio of two determinants, i.e., as a smooth function of all the coefficients. Since the coefficients $E_i(x_{1k} + b_1)$ smoothly depend on b_1, we conclude that the solutions $C_{ij}(b_1)$ are also differentiable functions of b_1.

Since all the functions $E_i(x_1)$ and $C_{ij}(b_1)$ are differentiable, we can differentiate both sides of all equalities describing $E_i(x_1 + b_1)$ with respect to b_1, and take $b_1 = 0$. Then, we get the following system of equations:

$$E_1'(x_1) = c_{11} \cdot E_1(x_1) + \ldots + c_{1d} \cdot E_d(x_1);$$

$$\dots$$

$$E_d'(x_1) = c_{d1} \cdot E_1(x_1) + \ldots + c_{dd} \cdot E_d(x_1),$$

where $E_i'(x_1)$ denotes the derivative, and $c_{ij} \stackrel{\text{def}}{=} C_{ij}'(0)$.

In other words, for the functions $E_1(x), \ldots, E_d(x)$, we get a system of linear differential equations with constant coefficients. It is known that a general solution to such system of equations is a linear combination of functions of the type $x_1^k \cdot \exp((p + \mathrm{i} \cdot q) \cdot x_1)$, i.e., functions of the type $x_1^k \cdot \exp(p \cdot x_1) \cdot \cos(q \cdot x_1)$ and $x_1^k \cdot \exp(p \cdot x_1) \cdot \sin(q \cdot x_1)$, where $p + \mathrm{i} \cdot q$ are eigenvalues of the matrix c_{ij}, and k is a non-negative integer corresponding to duplicate eigenvalues.

For a $d \times d$ matrix, the multiplicity of an eigenvalue cannot exceed d, so $k \le d$.

11°. Let us now study the consequences of *scale*-invariance of the class $\mathcal{S}$ with respect to the first variable. This scale-invariance implies, in particular, that for each basis function $e_i(x_1, x_2, \ldots, x_n)$, the result of its re-scaling $e_i(a_1 \cdot x_1, x_2, \ldots, x_n)$ is also a function from the class $\mathcal{S}$, i.e., that

$$e_i(a_1 \cdot x_1, x_2, \ldots, x_n) = \sum_{j=1}^{d} D_{ij}(a_1) \cdot e_j(x_1, x_2, \ldots, x_n),$$

for some coefficients D_{ij} that, in general, depend on a_1. Thus,

$$E_1(a_1 \cdot x_1) = D_{11}(a_1) \cdot E_1(x_1) + \ldots + D_{1d}(a_1) \cdot E_d(x_1);$$

$$\ldots$$

$$E_d(a_1 \cdot x_1) = D_{d1}(a_1) \cdot E_1(x_1) + \ldots + D_{dd}(a_1) \cdot E_d(x_1);$$

Similarly to Part 10 of this proof, we can prove that the dependencies $D_{ij}(a_1)$ are also differentiable. By differentiating both sides of the above equations with respect to a_1 and taking $a_1 = 1$, we conclude that

$$x_1 \cdot E_1'(x_1) = d_{11} \cdot E_1(x_1) + \ldots + d_{1d} \cdot E_d(x_1);$$

$$\ldots$$

$$x_1 \cdot E_d'(x_1) = d_{d1} \cdot E_1(x_1) + \ldots + d_{dd} \cdot E_d(x_1),$$

where $d_{ij} \stackrel{\text{def}}{=} D_{ij}'(1)$.

In each equation, the left-hand side $x_1 \cdot \dfrac{dE_i}{dx_1}$ can be reformulated as $\dfrac{dE_i}{dx_1/x_1} = \dfrac{dE_i}{d(\ln(x_1))}$. Thus, for the new variable $X_1 \stackrel{\text{def}}{=} \ln(x_1)$, we get the system of linear differential equations with constant coefficients:

$$\frac{dE_1}{dX_1} = d_{11} \cdot E_1(X_1) + \ldots + d_{1d} \cdot E_d(X_1);$$

$$\ldots$$

$$\frac{dE_d}{dX_1} = d_{d1} \cdot E_1(X_1) + \ldots + d_{dd} \cdot E_d(X_1).$$

We already know that a general solution to this equation is a linear combination of functions $X_1^k \cdot \exp(p \cdot X_1) \cdot \cos(q \cdot X_1)$ and $x_1^k \cdot \exp(p \cdot X_1) \cdot \sin(q \cdot X_1)$. Substituting $X_1 = \ln(x_1)$ into these formulas and taking into account that $\exp(p \cdot \ln(x_1)) = (\exp(\ln(x_1))^p = x_1^p$, we conclude that a general solution is a linear combination of functions of the type $(\ln(x_1))^k \cdot x_1^p \cdot \cos(q \cdot \ln(x_1))$ and $(\ln(x_1))^k \cdot x_1^p \cdot \sin(q \cdot \ln(x_1))$.

12°. From Parts 10 and 11 of this proof, we get two different expressions for the functions $E_i(x_1)$. By comparing these expressions, one can easily see that the only functions that can be described in both forms are functions of the form x^k for some non-negative integer $k \le d$ – or their linear combinations. So, each function $E_i(x_1)$ is a linear combination of such functions – i.e., a polynomial.

13°. We have shown that for each combination of values of $x_2, \ldots, x_n$, the dependence of each function $e_i(x_1, x_2, \ldots, x_n)$ on x_1 can be described by a polynomial of degree $\le d$. Similarly, we can prove that for each combination of values $x_1, x_3, \ldots, x_n$), the dependence on x_2 is also described by a polynomial. Let us combine these two conclusions and prove that each i, and for all possible values of $x_3, \ldots, x_n$, the dependence of $e_i(x_1, x_2, x_2, \ldots, x_n)$ on x_1 *and* on x_2 can be described by a polynomial of two variables.

Indeed, let us denote $T(x_1, x_2) \stackrel{\text{def}}{=} e_i(x_1, x_2, x_3, \ldots, x_n)$. We know that:

- for each x_2, this expression is a polynomial is x_1, and
- for each x_1, this expression is a polynomial is x_2.

Let us prove that $T(x_1, x_2)$ is a polynomial of two variables.

Indeed, the fact that the dependence of e_i on x_1 can be described by a polynomial of order $\le d$ can be rewritten, in terms of the function $T(x_1, x_2)$, as

$$T(x_1, x_2) = a_0(x_2) + a_1(x_2) \cdot x_1 + \ldots + a_d(x_2) \cdot x_1^d.$$

In writing this expression, we took into account that, in general, for different values of x_2, the coefficients $a_0, \ldots, a_d$ of this polynomial may be different.

Let us substitute d_1 different values $x_{10}, \ldots, x_{1d}$ of x_1 into this formula. As a result, we have $d+1$ linear equations for $d+1$ unknowns $a_0(x_2), \ldots, a_d(x_1)$, with constant coefficients:

$$T(x_{10}, x_2) = a_0(x_2) + a_1(x_2) \cdot x_{10} + \ldots + a_d(x_2) \cdot x_{10}^d;$$

$$\ldots$$

$$T(x_{1d}, x_2) = a_0(x_2) + a_1(x_2) \cdot x_{1d} + \ldots + a_d(x_2) \cdot x_{1d}^d.$$

In general, each component in a solution to a system of linear equations is a linear combination of the right-hand sides. The right-hand sides $T(x_{1i}, x_2)$ are polynomials of x_2. Thus, each coefficient $a_i(x_2)$ is a linear combination of polynomials – thus, a polynomial itself. Since all the expressions $a_i(x_2)$ are polynomials, the whole above expression for $T(x_1, x_2)$ becomes a polynomial in two variables x_1 and x_2.

By adding variables one by one, we can prove that the dependence on x_1, x_2, and x_3 is a polynomial, etc. – all the way to proving that the dependence of each of the basis function $e_i(x_1, \ldots, x_n)$ on all n variables $x_1, \ldots, x_n$ is a polynomial. Thus, each element of the class $\mathcal{S}$ – which is a linear combination of the basis functions – is also a polynomial.

For each tuple of parameters c, the function $\ln(\rho(x, c))$ belongs to the class $\mathcal{L} \subseteq \mathcal{S}$ and is, thus, also a polynomial. So, indeed, each pdf $\rho(x, c)$ from the family F has the form $\exp(P(x_1, \ldots, x_n))$ for some polynomial $P(x_1, \ldots, x_n)$. The proposition is proven.

Acknowledgements

This work was supported in part by the US National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science) and HRD-1242122 (Cyber-ShARE Center of Excellence).

References

Colbert, B. K., L. G. Crespo, and M. M. Peet, A sum of squares optimization approach to uncertainty quantification, In: *Proceedings of the 2019 American Control Conference ACC*, Philadelphia, Pennsylvania, USA, July 2019, pp. 5378–5384.

Crespo, L., An introduction to sliced-normal distributions, In M. Beer and E. Zio, editors, *Proceedings of the 29th European Safety and Reliability Conference ESREL'2019*, Hannover, Germany, September 2019. Research Publishing, Singapore, p. 28.

Crespo, L. G., B. K. Colbert, S. P. Kenny, and D. P. Giesy, On the quantification of aleatory and epistemic uncertainty using Sliced-Normal distributions, *Systems and Control Letters*, 134, Paper 104560, 2019.

Rabinovich, S. G., *Measurement Errors and Uncertainties: Theory and Practice*, Springer, New York, 2005.

Sheskin, D. J., *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman and Hall/CRC, Boca Raton, Florida, 2011.