Philip Hougaard

Convex Functions in Exponential Families

Preprint
December
1983
8

Institute of Mathematical Statistics
University of Copenhagen

Philip Hougaard*

CONVEX FUNCTIONS IN EXPONENTIAL FAMILIES

Preprint 1983 No. 8

INSTITUTE OF MATHEMATICAL STATISTICS
UNIVERSITY OF COPENHAGEN

December 1983

Abstract A number of functions of the canonical parameter in an exponential family are shown to be logaritmically convex. In particular the second moment is a convex function of the mean. A consequence is that a mixture of distributions from an exponential family has larger variance than the distribution having the same mean, as shown by Shaked (1980). The same method can be used to construct new exponential families, as is exemplified by a nontrivial family of dependent random variables having gamma distributed sum.

1. INTRODUCTION

Consider a onedimensional natural exponential family, i.e. assume the real-valued random variable X has density of the form $\exp(\theta x)/\phi(\theta)\mu(dx)$. Assume that the family is full, i.e. the canonical parameter θ varies in the set $D = \{\theta \mid \int \exp(\theta x)\mu(dx) < \infty\}$. For a general reference, see Barndorff-Nielsen (1978). By means of Hölder's inequality it can be shown that the normalising constant $\phi(\theta) = \int \exp(\theta x)\mu(dx)$ is log convex, i.e. $\ln \phi(\theta)$ is a convex function. Actually $\ln \phi(\theta)$ is a closed convex function. The log convexity can also be formulated as $\phi(\alpha\theta_1+(1-\alpha)\theta_2) \leq \phi(\theta_1)^{\alpha}\phi(\theta_2)^{1-\alpha}$ for $\alpha \in [0;1]$. In D^o, the interior of D, ϕ is infinitely often differentiable and the k'th derivative is

$$\phi^{(k)}(\theta) = \int x^k \exp(\theta x)\mu(dx) \tag{1.1}$$

If k is even, say $k = 2n$, this defines a new exponential family having density $\exp(\theta x)/\phi^{(2n)}(\theta)x^{2n}\mu(dx)$ defined at least for $\theta \in D^o$. It follows that $\ln \phi^{(2n)}(\theta)$ is a log convex function.

The purpose of the present paper is to prove that a number of other functions are also log convex. The method of proof is similar to the proof of log convexity of $\phi^{(2n)}(\theta)$. A new exponential family is constructed having the function as normalising constant. This is done in Section 2.

One of the simplest functions which are proved log convex is $\phi''\phi - \phi'^2$, where e.g. ϕ'' is short for $\phi''(\theta)$, the second derivative of $\phi(\theta)$. From the log convexity it can be proved that in the original family, the second moment EX^2 is a convex function of the mean EX. A practical consequence of this is that a mixture of distributions from an exponential family has larger variance than the member with the same mean as the mixture. This is proved in Section 3.

Our aim has been to prove log convexity of some functions, but the method also gives new exponential families and can be used to generate such having normalising constants which can be calculated from the original normalising constant. This is discussed in Section 4, exemplified inter alia by a family of twodimensional distributions in which the sum of the variables is gamma distributed.

2. CONSTRUCTION OF CONVEX FUNCTIONS

Each of the functions, which will be proved log convex, is constructed by means of a polynomium in one or several variables. As generalisation is fairly obvious, some calculations will only be performed for specific examples.

Consider first a polynomium $p(x)$ in one variable. The integral $\int p(x)\exp(\theta x)\mu(dx)$ is finite for $\theta \in \tilde{D} \supseteq D^{o}$, because of (1.1), with value, say $\tilde{\phi}(\theta)$. If $p(x) \geq 0$ for all x or at least only negative on a set of measure 0 using $\mu(dx)$, this defines an exponential family density $p(x)\exp(\theta x)\mu(dx)/\tilde{\phi}(\theta)$, proving log convexity of $\tilde{\phi}(\theta)$. This function turns out to be $p(\phi)$, if powers are interpreted as differentiations. For example, let a be a constant, then $p(x) = x^2 - 2ax + a^2 = (x-a)^2 \geq 0$. Each term is integrated using (1.1) yielding

$$\tilde{\phi}(\theta) = \phi'' - 2a\phi' + a^2\phi$$

More interesting functions can be found by considering polynomials of several variables, which will be demonstrated for two variables (x,y). Assume the polynomium $p(x,y)$ is nonnegative for all (x,y) or at least only negative on a set of measure 0 using product measure $\mu \times \mu$. The integral

$$\int_x \int_y p(x,y)e^{\theta(x+y)}\mu(dx)\mu(dy)$$

is finite for $\theta \in \tilde{D} \supseteq D^{o}$, with value, say $\tilde{\phi}(\theta)$. Because $p(x,y)$ is nonnegative this defines a density of (x,y). Thus the density of $z = x+y$ is

$$e^{\theta z}/\tilde{\phi}(\theta) \qquad \int_{x+y=z} p(x,y)\mu(dx)\mu(dy) ,$$

proving that $\tilde{\phi}(\theta)$ is log convex. As the underlying measure for z is not needed for the log convexity, the only problem is to calculate $\tilde{\phi}(\theta)$, which will be done for an example. Suppose $p(x,y) = (x-y)^2 = x^2 - 2xy + y^2$, which is clearly nonnegative. Integrating termvise give products of integrals. Thus

$$\tilde{\phi}(\theta) = \int_x x^2 e^{\theta x}\mu(dx) \int_y e^{\theta y}\mu(dy) - 2\int_x xe^{\theta x}\mu(dx) \int_y ye^{\theta y}\mu(dy)$$

$$+ \int_x e^{\theta x}\mu(dx) \int_y y^2 e^{\theta y}\mu(dy) = \phi''\phi - 2\phi'^2 + \phi\phi''$$

$$= 2(\phi''\phi - \phi'^2),$$

from which the log convexity of $\phi''\phi - \phi'^2$ follows. Interpreting powers as differentiations, e.g. x^2y as $\phi''\phi'$, the integral is $\tilde{\phi}(\theta) = p(\phi,\phi)$, showing that $\tilde{\phi}$ only depends on a symmetrized version of $p(x,y)$, e.g. the two squared terms in the example give rise to identical terms in $\tilde{\phi}(\theta)$.

The integral can still be calculated if we allow $p(x,y)$ to be a polynomium in x,y and $\exp(a_i x)$, $\exp(b_i y)$, $i = 1,\dots,I$, where a_i, b_i are constants. For example

$$\phi^{(k)}(\theta+a) = \int x^k \exp(ax)\exp(\theta x)\mu(dx)$$

If the polynomium has only one term, this is just a change of parameter, but if there are more terms this does give new results. For example if the polynomium is $\{\exp(ax) - 1\}^2$, we prove log convexity of $\phi(\theta + 2a) - 2\phi(\theta + a) + \phi(\theta)$.

Any nonnegative function $p(x)$ will by this method give rise to an exponential family, thereby proving log convexity of some function on the set for which the integral is finite. The advantage of polynomials in x and exponentials of x is that the integral can easily be calculated using difference and differential operators on ϕ and it is finite on D^o.

From the proof it is trivial that if $\tilde{\phi}_1(\theta)$ and $\tilde{\phi}_2(\theta)$ are created in this way, also $\tilde{\phi}_1(\theta) + \tilde{\phi}_2(\theta)$ is OK, proving log convexity of this function. It is, however, true in general that if $f_1(\theta)$ and $f_2(\theta)$ are both log convex, also their sum is log convex. To see this evaluate $(f_1 + f_2)(\alpha\theta_1 + (1-\alpha)\theta_2) \le f_1(\theta_1)^\alpha f_1(\theta_2)^{1-\alpha} + f_2(\theta_1)^\alpha f_2(\theta_2)^{1-\alpha}$, because of the log convexity of f_1 and f_2. Using Hölder's inequality that is $\le$ $\{f_1(\theta_1) + f_2(\theta_1)\}^\alpha \{f_1(\theta_2) + f_2(\theta_2)\}^{1-\alpha}$, which is the desired result. As we can multiply by a positive constant without changing log convexity, it is clear that the classes of functions we consider are all convex cones, i.e. the class of log convex functions, the class of log convex functions derived from positive polynomials, the class of positive polynomials etc. Therefore it is sufficient to consider minimal examples, i.e. a minimal class of functions generating the same as all polynomials. For example if the essential infimum of $p(x)$ using μ is $c > 0$, a stronger result follows from using $p(x) - c$ as polynomium. More generally, if nonproportional polynomials p_1, p_2 exist with $p_1(x) \ge 0$, $p_2(x) \ge 0$, $p(x) = p_1(x) + p_2(x)$, stronger results follow using $p_1(x)$ and $p_2(x)$ instead of $p(x)$. A positive polynomium in one variable, not including exponential terms, is a sum of squares of polynomials, i.e. if $p(x) \ge 0$, $\exists k, p_1,\dots,p_k$: $p(x) = \Sigma \{p_i(x)\}^2$, such that we need only consider squares of polynomials. For polynomials of several variables or including exponential terms, this is not true. As noted by Berg, Christensen & Jensen (1979) the positive polynomium $p(x,y) = x^2y^2(x^2 + y^2 - 1) + 1$ is not a sum of squares of polynomials. As the infimum of this polynomium is 26/27, it follows that $2\phi''''\phi'' - \phi''^2 + \phi/27$ is log convex. Taking $y = \exp(ax/2)$ for some constant a, yields log convexity of

$$\phi''''(\theta + a) + \phi''(\theta + 2a) - \phi''(\theta + a) + \phi(\theta)/27.$$

It seems not possible to give a simple characterization of a minimal generating class for the log convex functions derived by polynomials.

Formally this procedure constructs a mixture of distributions generated as translations of (1.1) for $k = 0,1,\ldots$ and for a polynomium in, say, n variables, a mixture of n-dimensional distributions each consisting of independent variables generated by translations of (1.1). It is, however, not a true mixture, because the weights might be negative, and for k odd the integrand in (1.1) is not necessarily positive. The most interesting results follow from choosing the weights as negative as possible.

It is not necessary to use product measure or the same θ for x and y. By choosing different measures another class of log convex functions can be found. The results actually generalize to multidimensional families, which can be exemplified by the two-dimensional family $\exp(\theta_1 x_1 + \theta_2 x_2)\mu(dx_1, dx_2)/\phi(\theta_1,\theta_2)$. By use of the polynomium $(x_1 - x_2)^2$ we get a family with normalising constant $d^2\phi/d\theta_1^2 - 2\,d^2\phi/d\theta_1 d\theta_2 + d^2\phi/d\theta_2^2$, proving that this function is log convex.

If μ is concentrated on the nonnegative numbers $\phi^{(n)}(\theta)$ is log convex for all integers n. Using the method in this paper also many other log convex functions can be found in this case as well as in other cases with observations concentrated on a subset of $\mathbb{R}$. For those cases there seems not to be results concerning positive polynomials in general, but there are some results for polynomials in one variable and excluding exponential terms. Polynomials of one variable, which are nonnegative on the nonnegative numbers can be written on form $p(x) = p_1(x) + x\,p_2(x)$, where p_1, p_2 are polynomials, which are nonnegative on the whole real axis. Thus compared to the real case the only extra positive polynomium of degree less than or equal 2 is $p(x) = x$, saying that $\phi'(\theta)$ is log convex. For μ concentrated on $[0;1]$ the positive polynomials of one variable can be represented by $p(x) = xp_1(x) + (1-x)p_2(x)$, where $p_1(x)$ and $p_2(x)$ are nonnegative on the real axis. Compared to the real case we get x and $1-x$ of degree 1, yielding log convexity of $\phi'(\theta)$ and $\phi(\theta) - \phi'(\theta)$. Of degree 2 the extra polynomials have the form $p_1(x) = (x-a)^2$, $p_2(x) = (x-b)^2$ for some a,b, i.e. $p(x) = \{1 + 2(b-a)\}\ x^2 + (a^2 - b^2 - 2b)\ x + b^2$. For general polynomials it is very limited what can be said to characterise them. A simple example for μ concentrated on $[0;1]$ stems from the convexity of e^{ax}, which implies that $x\{\exp(a) - 1\} + 1 - \exp(ax) \geq 0$, giving log convexity of $\phi'(\theta)\{\exp(a) - 1\} + \phi(\theta) - \phi(\theta + a)$.

If the original distribution is continuous and p not identically 0, the function $\ln \tilde{\phi}(\theta)$ is strictly convex. If it is discrete, the derived distribution might be degenerate in which case the convexity is not strict.

The functions without exponentialterms can be written in terms of moments in the original distribution by isolating a power of ϕ. The examples $(x-a)^2$ and $(x-y)^2$ above are the functions $\phi\ E(X-a)^2$ and $\phi^2\ \mathrm{Var}(X)$.

The table lists some simple examples and the polynomials used to prove the log convexity. The ranges for the constants a and b are described. The constants c and d are (possibly complicated) functions of a and b.

3. A CONVEX FUNCTION OF THE MEAN

In the preceding section only functions of the canonical parameter were considered. This section focuses on the following theorem and its implications.

Theorem In an exponential family the second moment is a convex function of the mean.

This result has been proven earlier by Shaked (1980) and Schweder (1982). A particularly simple class of models is those, for which the second moment is a quadratic function of the mean. A unified theory for these models has been developed in Morris (1982, 1983).

Introducing the meanvalue parameter $\tau = E_\theta X$ and $\chi(\theta) = \ln \phi(\theta)$, the conjecture concerns the convexity of $f(\tau) = \tau^2 + \chi''(\theta(\tau))$, the terms describing $(EX)^2$ and $Var(X)$ respectively. The theorem could be formulated in terms of the second moment around any given value instead of 0. This is however, not more general, as the difference is an affine function, such that convexity is preserved.

Consider a mixture of distributions from the exponential family. More precisely let T be a continuous or discrete random variable concentrated on $\tau(D^o)$, such that each value of T chooses the (unique) distribution from the exponential family having mean T. Let X follow that distribution given T. The unconditional mean is $EX = E(EX|T) = ET$, say τ_0. It follows from the theorem that $Var(X) \geq Var(X|T=\tau_0)$, i.e. the variance in a mixture is larger than the variance in the member of the family having the same mean. To see this, choose $f(\tau) = E_\tau(X-\tau_0)^2$, the second moment around τ_0. Then, using $\upsilon(dt)$ as the measure for T, $Var(X) = E(X-\tau_0)^2 = \int f(\tau)\upsilon(dt)$, which because of the convexity is greater than or equal to $f(\tau_0) = Var(X|T = \tau_0)$. This result is not surprising as in practicemixture distributions are chosen, when there is some theoretical belief that the distribution is from a certain family, but the variation is larger than expected in that family. In short, mixtures are used to increase the variation. However in general it is not true that the variance is increased. In location models and in scale models the variance is a convex function of the mean, from which it trivially follows that the second moment is strictly convex. In those cases and in exponential families as shown above, the variance is actually increased.

For the normal distribution case $(T \sim N(\tau_0, \sigma_t^2),\ X|T \sim N(T, \delta_x^2))$ the value of $\mathrm{Var}\,X/\mathrm{Var}(X|T = \tau_0)$ is $(1-\rho^2)^{-1}$, where ρ is the correlation coefficient. In other cases that ratio is a simple measure of how large the variance is compared to the variance in the exponential family.

<u>Proof of the theorem</u> By twice differentiation of $f(\tau) = \tau^2 + \chi''(\theta(\tau))$ it is found that $f''(\tau) = 2 + \chi''''(d\tau/d\theta)^{-2} - \chi''' \, d^2\tau/d\theta^2 \, (d\tau/d\theta)^{-3}$. Using $\tau(\theta) = \chi'$ it is found that $f''(\tau) = 2 + \chi''''\chi''^{-2} - \chi'''^2 \chi''^{-3}$. Differentiation of $g(\theta) = \ln(\phi''\phi - \phi'^2) = 2 \ln \phi + \ln \chi''$ yields $g''(\theta) = 2\chi'' + \chi''''\chi''^{-1} - \chi'''^2 \chi''^{-2}$. The convexity of g, shown in Section 2, implies $g''(\theta) \geq 0$, which in turn implies $f''(\tau) = g''(\theta)/\chi'' \geq 0$ and thereby convexity of $f(\tau)$, using that $\chi'' = V X > 0$. If the variance is not strictly positive, X is degenerate; independent of θ and T varies in a one-element set, in which case the result is trivially correct.

4. CONSTRUCTION OF NEW EXPONENTIAL FAMILIES

The method of Section 2 yields new exponential families. Some of them have a natural interpretation. For example suppose that for a positive random variable the probability of sampling is proportional to the variable. This is often the case in stereological applications. The distribution in the sample is then the family derived by choosing $p(x) = x$. Another example of the same effect is in Hougaard (1984), where the hazard of death is proportional to x and the distribution of x among deaths is thus the derived family.

In most cases, however, the derived distributions do not have a natural interpretation in terms of the original distribution, but they might be interesting in their own right. It is for example possible to construct distributions with several modes. We will consider two examples of a different kind.

Starting with the inverse Gaussian distribution, for which a for our purpose natural version of the density is

$$f(x;\theta,\psi) = (-\psi/\pi)^{1/2} \exp\{2(\psi\theta)^{1/2}\} \, x^{-3/2} \exp(\theta x + \psi/x),$$

for $x > 0$ and θ and ψ are parameters, $\theta \leq 0$, $\psi < 0$. Letting primes denote differentiations with respect to θ and with

$$\phi(\theta,\psi) = \exp\{-2(\psi\theta)^{1/2}\}(-\psi)^{-1/2},$$

we find

$$2(\phi''\phi - \phi'^2) = \exp\{-4(\psi\theta)^{1/2}\}(-\psi)^{-1/2} (-\theta)^{-3/2},$$

which is fairly simple. From this we find that the density of the distribution derived by the polynomium $(x-y)^2$ is

$$(-\theta)^{3/2}(-\psi)^{1/2} \pi^{-1} \exp\{4(\psi\theta)^{1/2}\} x^{-3/2} y^{-3/2} (x-y)^2$$

$$\cdot \exp\{\theta(x+y) + \psi(1/x + 1/y)\}$$

Also starting with the gamma distribution we can find a two dimensional distribution with x and y dependent, for which the sum $z = x + y$ is gamma distributed.

Besides assuming $p(x,y) \geq 0$ for all $x,y \geq 0$ it is necessary to assume that all terms in $p(x,y)$ are of the same order, say k. In Section 2 it is described how to calculate $\tilde{\phi}(\theta)$ such that

$$p(x,y)e^{\theta(x+y)} x^{\delta-1} y^{\gamma-1} / \tilde{\phi}(\theta)$$

is a probability density, using the normalising constant in the gamma density ($\phi(\theta) = \Gamma(\delta)/\theta^{\delta}$ for $\mu(dx) = x^{\delta-1}dx$). Splitting up after the terms in $p(x,y)$ the integral giving the density of $z = x+y$ will consist of terms each corresponding to a sum of two independent gammas, for which the sum of shape parameters is $\delta + \gamma + k$. For example $p(x,y) = (x-y)^2$ yields the normalising constant $\tilde{\phi}(\theta) = \Gamma(\delta)\Gamma(\gamma)\{(\delta-\gamma)^2 + \delta + \gamma\}\ \theta^{-(\delta+\gamma+2)}$ and the distribution of z is then gamma with shape parameter $\delta + \gamma + 2$. In this distribution x and y tend to be different; the density is 0 along the line $x = y$.

Using the mixture interpretation, cf. Section 2, the result is not surprising. It can be considered as a mixture of twodimensional distributions consisting of independent gammas such that the sum of the shape parameters constantly is $\delta + \gamma + k$. However this procedure allows negative weights, and it is easy to check that the weights are allowable; the only criterion being that the polynomium and thereby the density for (x,y) is nonnegative.

ACKNOWLEDGEMENTS

The comments from Poul S. Eriksen, Søren Johansen and Steffen Lauritzen are greatly appreciated.

REFERENCES

Barndorff-Nielsen, O. (1978): Information and exponential families. Wiley.

Berg, C.,Christensen, J.P.R. & Jensen, C.U. (1979): A remark on the multidimensional moment problem. Math.Ann. 243, 163 - 169.

Hougaard, P. (1984): Life table methods for heterogenous populations: Distributions describing the heterogeneity. To appear in Biometrika, 71.

Morris, C.N. (1982): Natural exponential families with quadratic variance functions. Ann. Stat. 10, 65 - 80.

Morris, C.N. (1983): Natural exponential families with quadratic variance functions: Statistical theory. Ann.Stat. 11, 515 - 529.

Schweder, T. (1982): On the dispersion of mixtures. Scand. J. Statist. 9, 165 - 169.

Shaked, M. (1980): On mixtures from exponential families. J. Roy. Statist. Soc. Ser. B 42, 192 - 198.

TABLE OF LOG CONVEX FUNCTIONS

function	range for a(b)	originating polynomium
$\phi(\theta+2a) - 2\phi(\theta+a) + \phi(\theta)$	$\mathbb{R}$	$\{1-\exp(ax)\}^2$
$\phi'' - 2a\phi' + a^2\phi$	$\mathbb{R}$	$(x-a)^2$
$\phi''(\theta) - 2\phi'(\theta+a) + \phi(\theta+2a)$	$\mathbb{R}$	$\{x-\exp(ax)\}^2$
$\phi''\phi - \phi'^2$		$(x-y)^2$
$\phi''\phi - \phi'^2 + a(\phi'+b\phi)^2$	$[0,2](\mathbb{R})$	$(x-cy+d)^2$
$\phi'''' - 2a\phi'' + a^2\phi$	$\mathbb{R}$	$(x^2-a)^2$
$\phi''''\phi - \phi''^2$		$(x^2-y^2)^2$
$\phi''''\phi - a\phi'''\phi' + (a-1)\phi''^2$	$[0;4]$	$(x^2-cy^2-dxy)^2$
$\phi''''\phi - 2a\phi''\phi' + a^2\phi''\phi$	$\mathbb{R}$	$(x^2-ay)^2$
$\phi''^2\phi + \phi''\phi^2 - 2\phi'^3$		$(x-yz)^2$
$\phi^{(2n)}\phi - \phi^{(n)2}$		$(x^n-y^n)^2$
$\phi^{(2n)} - 2a\phi^{(n)} + a^2\phi$	$\mathbb{R}$	$(x^n-a)^2$
$\phi^{(4n)}\phi - a\phi^{(3n)}\phi^{(n)} + (a-1)\phi^{(2n)2}$	$[0,4]$	$(x^{2n}-cy^{2n}-dx^ny^n)^2$

PREPRINTS 1982

COPIES OF PREPRINTS ARE OBTAINABLE FROM THE AUTHOR OR FROM THE INSTITUTE OF MATHEMATICAL STATISTICS, UNIVERSITETSPARKEN 5, 2100 COPENHAGEN Ø, DENMARK.

No. 1 Holmgaard, Simon and Yu, Song Yu: Gaussian Markov Random Fields Applied to Image Segmentation.

No. 2 Andersson, Steen A., Brøns, Hans K. and Jensen, Søren Tolver: Distribution of Eigenvalues in Multivariate Statistical Analysis.

No. 3 Tjur, Tue: Variance Component Models in Orthogonal Designs.

No. 4 Jacobsen, Martin: Maximum-Likelihood Estimation in the Multiplicative Intensity Model.

No. 5 Leadbetter, M.R.: Extremes and Local Dependence in Stationary Sequences.

No. 6 Henningsen, Inge and Liestøl, Knut: A Model of Neurons with Pacemaker Behaviour Recieving Strong Synaptic Input.

No. 7 Asmussen, Søren and Edwards, David: Collapsibility and Response Variables in Contingency Tables.

No. 8 Hald, A. and Johansen, S.: On de Moivre's Recursion Formulae for Duration of Play.

No. 9 Asmussen, Søren: Approximations for the Probability of Ruin Within Finite Time.

PREPRINTS 1983

COPIES OF PREPRINTS ARE OBTAINABLE FROM THE AUTHOR OR FROM THE INSTITUTE OF MATHEMATICAL STATISTICS, UNIVERSITETSPARKEN 5, 2100 COPENHAGEN Ø, DENMARK.

No. 1 Jacobsen, Martin: Two Operational Characterizations of Cooptional Times.

No. 2 Hald, Anders: Nicholas Bernoulli's Theorem.

No. 3 Jensen, Ernst Lykke and Rootzén, Holger: A Note on De Moivre's Limit Theorems: Easy Proofs.

No. 4 Asmussen, Søren: Conjugate Distributions and Variance Reduction in Ruin Probability Simulation.

No. 5 Rootzén, Holger: Central Limit Theory for Martingales via Random Change of Time.

No. 6 Rootzén, Holger: Extreme Value Theory for Moving Average Processes.

No. 7 Jacobsen, Martin: Birth Times, Death Times and Time Substitutions in Markov Chains.

No. 8 Hougaard, Philip: Convex Functions in Exponential Families.