Singular Transformations of Probability Density Functions

Author: *R. Scott McIntire*

Date: Dec 3, 2003

Overview

For non-singular transformations of variables with distributions, there are standard formulas that describe the probability distribution of the target variable. However, in the case of singular transformations, there is no immediate formula. Often what is done is to add independent variables in a somewhat add-hoc fashion to make the transformation non-singular, and then "integrate out" − if possible − the independent variables from the resulting formula. Below we describe a procedure to determine the target distribution with a more systematic process.

Singular Transformation Density Formula

We provide and prove a theorem for constructing the probability density function of a random variable Y which is a singular function of a random variable X. By singular, we mean that $Y = f(X)$ with f a function such that $f : R^n \mapsto R^m$ and $m < n$.

Theorem 1. *Let $f : \mathcal{D} \subset \mathbf{R}^n \mapsto \mathbf{R}^m$ be a continuously differentiable function with $n > m$. In addition, $Df(\mathbf{x})$ is of rank m $\forall \mathbf{x} \in \mathcal{D}$. If for each $\mathbf{y}$, $f^{-1}(\mathbf{y})$ is a union of $k(\mathbf{y})$ disjoint C^1 $n-m$ dimensional manifolds with parameterizations $\mathbf{x}_i(\in \mu; \mathbf{y})$, $i \in [1, k(\mathbf{y})]$, $\mu \in \Omega_i \subset \mathbf{R}^{n-m}$, with $P_X(\mathbf{x})$ being the continuous probability density function of the random variable X, then the density function $P_Y(\mathbf{y})$ of the random variable Y defined by $Y = f(X)$ exists, is continuous, and is given by*

$$P_Y(\mathbf{y}) = \sum_{i=1}^{k(\mathbf{y})} \int_{\Omega_i} \frac{P_X(\mathbf{x}_i(\mathbf{y};\mu))}{\sqrt{|Df(\mathbf{x}_i(\mathbf{y};\mu))\ Df^T(\mathbf{x}_i(\mathbf{y};\mu))|}} \sqrt{\left|D_\mu \mathbf{x}_i^T(\mathbf{y};\mu)\ D_\mu \mathbf{x}_i(\mathbf{y};\mu)\right|}\, d\mu$$

Prerequisite Results

We recall how to compute the pseudo inverse of a matrix which maps from a high dimension to a low dimension.

Lemma 1. *Let $A : \mathbf{R}^n \mapsto \mathbf{R}^m$ be a linear transformation with $n > m$. If the rank of A is m, then the pseudo inverse of A is $A^T(AA^T)^{-1}$.*

proof : A given $\mathbf{x} \in R^n$ can be written uniquely as $\mathbf{x} = \mathbf{x_1} + \mathbf{x_2}$ where $\mathbf{x_1} \in R(A^T)$ and $\mathbf{x_2} \in N(A)$. For a given $\mathbf{y} \in R^m$, it easy to show that there is a unique $\mathbf{x_p} \in R(A^T)$ such that $A\mathbf{x_p} = \mathbf{y}$. The set of all $\mathbf{x} \in R^n$

such that $A\mathbf{x} = \mathbf{y}$ has the form $\{\mathbf{x_p} + \mathbf{x_n} | \mathbf{x_n} \in N(A)\}$. The pseudo inverse of $\mathbf{y}$ is defined to be $\mathbf{x_p}$. Since $R(A^T) \perp N(A)$, $\mathbf{x_p}$ is the solution to the problem:

$$\begin{aligned} &\min_{\mathbf{x}} \|\mathbf{x}\|^2 \\ \text{subject to}: \ &\mathbf{y} = A\mathbf{x} \end{aligned}$$

The solution of this can be found using Lagrange multipliers. The resulting minimization problem is :

$$\min_{\mathbf{x}} F(\mathbf{x}) = \min_{\mathbf{x}} \|\mathbf{x}\|^2 + \langle \lambda, \mathbf{y} - A\mathbf{x} \rangle$$

At a minimum, $\mathbf{x}^*$, $DF(\mathbf{x}^*)(\mathbf{h}) = \mathbf{0}$. $DF(\mathbf{x}^*)$ can be computed from $F(\mathbf{x} + \mathbf{h}) = F(\mathbf{x}) + DF(\mathbf{x})(h) + o(\mathbf{h})$. The order $\mathbf{h}$ terms are:

$$2\langle \mathbf{x}, \mathbf{h} \rangle - \langle \lambda, A\mathbf{h} \rangle$$

Therefore, $DF(\mathbf{x}^*)(\mathbf{h}) = \langle 2\mathbf{x}^* - A^T\lambda, \mathbf{h} \rangle$. Since $DF(\mathbf{x}^*) \equiv 0$, we have that $2\mathbf{x}^* - A^T\lambda = \mathbf{0}$. Applying A to this last equation and using the fact that A has rank m and $A\mathbf{x}^* = \mathbf{y}$ yields:[1] $\lambda = 2(AA^T)^{-1}\mathbf{y}$. This implies that $\mathbf{x}^* = A^T \left(AA^T\right)^{-1} \mathbf{y}$.

We need the following change of volume formula:

Lemma 2. *Let $A : \mathbf{R}^n \mapsto \mathbf{R}^m$ be a linear transformation with $n < m$, then if A has rank n it transforms the n dimensional unit cube in R^n to a region in R^m with n dimensional volume $\sqrt{|A^TA|}$.*

proof : Since $R(A)$ has rank n, there is a linear map $P : R(A) \mapsto R^n$ which is a linear isomorphism that preserves the inner product. That is, $\langle P\mathbf{x}, P\mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle$. Or, $\langle (P^TP - I)\mathbf{x}, \mathbf{y} \rangle = 0 \ \forall \mathbf{x}, \mathbf{y} \in R^n$. This implies that $P^TP \equiv I$. The change in volume of the unit cube under PA is the same as A as P preserves volumes. However, PA is a map from R^n to R^n, so it changes the volume of the unit cube by $|PA|$. Since for given transformations, $B, C : R^n \mapsto R^n$ we have $|B^T| = B$ and $|BC| = |B||C|$, it follows that $|PA|^2 = |PA||PA| = |A^TP^T||PA| = |A^TP^TPA| = |A^TA|$. The last equality comes from the fact that $P^TP = I$. Therefore, the change in volume is $\sqrt{|A^TA|}$.

Proof of Singular Transformation Formula

We now proceed with the proof of Theorem 1.

proof : Let μ be the induced probability measure defined by[2]

$$\mu(\Omega_y) = \int_{f^{-1}(\Omega_y)} P_X(\mathbf{x})\, d\mathbf{x}$$

Given an arbitrary Borel measurable region Ω_Y and letting $\Omega_X = f^{-1}(\Omega_Y)$ we have that

$$\mu(\Omega_Y) = \int_{\Omega_X} P_X(\mathbf{x})\, d\mathbf{x} \tag{1}$$

[1] Since A has rank m ($\dim(R(A)) = m$) and $R(AA^T) = R(A)$, AA^T is invertible.
[2] The measure extends in a natural way to all Lebesgue measurable sets.

Assume that Ω_X is composed of just one $n-m$ dimensional manifold. In this case there are local coordinates $\mathbf{x_1}, \mathbf{x_2}$, and manifolds Ω_{X_1}, Ω_{X_2}, such that the manifold, Ω_X is a product of Ω_{X_1} and Ω_{X_2} with the component manifolds Ω_{X_1}, Ω_{X_2}, locally tangent to $R(Df^T(\mathbf{x}))$ and $N(Df(\mathbf{x}))$ respectively. Therefore,

$$\mu(\Omega_y) = \int_{\Omega_X} P_X(\mathbf{x}(\mathbf{x_1}, \mathbf{x_2}))\, d\mathbf{x_1} \wedge d\mathbf{x_2}$$

Since for a given transformation A, $R(A^T) \perp N(A)$, it follows that the product measure $d\mathbf{x_1} \wedge d\mathbf{x_2}$ is equal to the product of the component measures. Thus, we may write (1) as

$$\mu(\Omega_y) = \int_{\Omega_{X_1}} \int_{\Omega_{X_2}} P_X(\mathbf{x}(\mathbf{x_1}, \mathbf{x_2}))\, d\mathbf{x_1}\, d\mathbf{x_2}$$

For a given $\mathbf{x_2}$, $\mathbf{y}$ maps to the manifold Ω_{X_1} in an invertible way (locally). The linear approximation to the map f is Df, its pseudo inverse is by lemma 1: $\mathbf{x} = Df^T(\mathbf{x})(Df(\mathbf{x})Df^T(\mathbf{x}))^{-1}\mathbf{y}$. Using $\mathbf{y}$ to express $\mathbf{x_1}$ combined with lemma 2 we have

$$\mu(\Omega_y) = \int_{\Omega_Y} \int_{\Omega_{X_2}} \frac{P_X(\mathbf{x}(\mathbf{y}, \mathbf{x_2}))}{\sqrt{|Df(\mathbf{x}(\mathbf{y}, \mathbf{x_2}))Df^T(\mathbf{x}(\mathbf{y}, \mathbf{x_2}))|}}\, d\mathbf{x_2}\, d\mathbf{y}$$

We argue now that the measure μ is absolutely continuous with respect to Lebesgue measure[1] so that there exists a measurable function, $P_Y(y)$ such that $\mu(\Omega_Y) = \int_{\Omega_Y} P_Y(y)\, d\mathbf{y}$. So that

$$\int_{\Omega_Y} P_Y(\mathbf{y})\, d\mathbf{y} = \int_{\Omega_Y} \int_{\Omega_{X_2}} \frac{P_X(\mathbf{x}(\mathbf{y}, \mathbf{x_2}))}{\sqrt{|Df(\mathbf{x}(\mathbf{y}, \mathbf{x_2}))Df^T(\mathbf{x}(\mathbf{y}, \mathbf{x_2}))|}}\, d\mathbf{x_2}\, d\mathbf{y}$$

Or,

$$\int_{\Omega_Y} \left(P_Y(\mathbf{y}) - \int_{\Omega_{X_2}} \frac{P_X(\mathbf{x}(\mathbf{y}, \mathbf{x_2}))}{\sqrt{|Df(\mathbf{x}(\mathbf{y}, \mathbf{x_2}))Df^T(\mathbf{x}(\mathbf{y}, \mathbf{x_2}))|}}\, d\mathbf{x_2} \right) d\mathbf{y} = 0$$

Since this is true for all Borel measurable sets, Ω_Y, we have

$$P_Y(\mathbf{y}) = \int_{\Omega_{X_2}} \frac{P_X(\mathbf{x}(\mathbf{y}, \mathbf{x_2}))}{\sqrt{|Df(\mathbf{x}(\mathbf{y}, \mathbf{x_2}))Df^T(\mathbf{x}(\mathbf{y}, \mathbf{x_2}))|}}\, d\mathbf{x_2}$$

But in this case the component manifold, Ω_{X_2}, is just $f^{-1}(\mathbf{y})$. If we parameterize this as $\mathbf{x}(\mathbf{y}, \mu)$, then by lemma 2 we have

$$P_Y(\mathbf{y}) = \int_{f^{-1}(y)} \frac{P_X(\mathbf{x}(\mathbf{y}, \mu))}{\sqrt{|Df(\mathbf{x}(\mathbf{y}, \mu))Df^T(\mathbf{x}(\mathbf{y}, \mu))|}} \sqrt{|D_\mu \mathbf{x}^T(\mathbf{y}, \mu))D_\mu \mathbf{x}(\mathbf{y}, \mu))|}\, d\mu$$

In the general case, we may have a situation like $y = x^2$. Here, there are two manifolds that contribute to the probability distribution of y. More generlly, there may be some number of manifolds which map onto the target variable. In this case we add the contributions of each to the probability distribution of the target variable, Y.

Applications of Theorem 1.

We apply Theorem 1 with various types of singular mappings.

[1] Since the Lebesgue measure is regular, any Lebesgue measurable set can be approximated by a Borel measurable set.

Example 1: $y = f(\mathbf{x}) = x_1 + x_2 \quad (f : \mathbf{R}^2 \mapsto R)$

The set, $f^{-1}(y)$ can be parameterized as $\mathbf{x}_1(\mu; y) = (\mu, y - \mu)^T$ with $\mu \in (-\infty, \infty)$.

$\sqrt{|D_\mu \mathbf{x}_1^T(\mu; y)\ D_\mu \mathbf{x}_1(\mu; y)|} = \|(1, -1)\| = \sqrt{2}$.

Also, $\sqrt{|Df(\mathbf{x}_1(\mu; y))\ Df^T(\mathbf{x}_1(\mu; y))|} = \|\nabla_{\mathbf{x}} f(\mathbf{x}_1(\mu; y))\| = \|(1, 1)\| = \sqrt{2}$.

Therefore, by Theorem 1 we have:

$$P_Y(y) = \int_{-\infty}^{\infty} P_{X_1 X_2}(\mu, y - \mu)\, d\mu$$

Example 2: $y = f(\mathbf{x}) = x_1/x_2 \quad (f : \mathbf{R}^2 \mapsto R)$

The set $f^{-1}(y)$ can be parameterized as $\mathbf{x}_1(\mu; y) = (y\mu, \mu)^T$ with $\mu \in (-\infty, 0) \cup (0, \infty)$. We label the first and second components of $\mathbf{x}_1$ as $x_{1,1}$ and $x_{1,2}$ respectively.

$\sqrt{|D_\mu \mathbf{x}_1^T(\mu; y)\ D_\mu \mathbf{x}_1(\mu; y)|} = \|(y, 1)\| = \sqrt{1 + y^2}$. And,

$$\begin{aligned}
\sqrt{|Df(\mathbf{x}_1(\mu; y))\ Df^T(\mathbf{x}_1(\mu; y))|} &= \|\nabla_{\mathbf{x}} f(\mathbf{x}_1(\mu; y))\| \\
&= (1/x_{1,2}(\mu; y), -x_{1,1}(\mu; y)/x_{1,2}(\mu; y)^2)\| \\
&= \sqrt{1 + y^2}/|\mu|
\end{aligned}$$

Therefore, Theorem 1 gives:

$$P_Y(y) = \int_{-\infty}^{\infty} |\mu|\, P_{X_1 X_2}(y\mu, \mu)\, d\mu$$

Example 3: $y = f(\mathbf{x}) = x_1^2 + x_2^2 \quad (f : \mathbf{R}^2 \mapsto R)$

The set $f^{-1}(y)$ is the union of the parameterizations: $\mathbf{x}_1(\mu; y) = (\mu, \sqrt{y - \mu^2})^T$

and $\mathbf{x}_2(\mu; y) = (\mu, -\sqrt{y - \mu^2})^T$ with $\mu \in [-\sqrt{y}, \sqrt{y}]$.

$\sqrt{|D_\mu \mathbf{x}_1^T(\mu; y)\ D_\mu \mathbf{x}_1(\mu; y)|} = \|(1, -\mu/\sqrt{y - \mu^2}\| = \sqrt{y/(y - \mu^2)}$.

$\sqrt{|D_\mu \mathbf{x}_2^T(\mu; y)\ D_\mu \mathbf{x}_2(\mu; y)|} = \|(1, \mu/\sqrt{y - \mu^2}\| = \sqrt{y/(y - \mu^2)}$.

$\sqrt{|Df(\mathbf{x}_1(\mu; y))\ Df^T(\mathbf{x}_1(\mu; y))|} = 2\|(\mu, \sqrt{y - \mu^2})\| = 2\sqrt{y}$.

$\sqrt{|Df(\mathbf{x}_2(\mu; y))\ Df^T(\mathbf{x}_2(\mu; y))|} = 2\|(\mu, -\sqrt{y - \mu^2})\| = 2\sqrt{y}$.

Therefore, Theorem 1 gives:

$$P_Y(y) = \begin{cases} \int_{-\sqrt{y}}^{\sqrt{y}} \frac{P_{X_1 X_2}(\mu, \sqrt{y - \mu^2}) + P_{X_1 X_2}(\mu, -\sqrt{y - \mu^2})}{2\sqrt{y - \mu^2}}\, d\mu & y \geq 0\ ; \\ 0 & \text{otherwise.} \end{cases}$$

Example 4: $y_1 = x_1 + x_2 + x_3\ ;\ y_2 = x_1 - x_2 \quad (f : \mathbf{R}^3 \mapsto \mathbf{R}^2)$

The set $f^{-1}(\mathbf{y})$ can be parameterized as $\mathbf{x}_1(\mu; y_1, y_2) = (\mu, \mu - y_2, y_1 + y_2 - 2\mu)^T$ with $\mu \in (-\infty, \infty)$.

$\sqrt{|D_\mu \mathbf{x}_1^T(\mu; \mathbf{y})\ D_\mu \mathbf{x}_1(\mu; \mathbf{y})|} = \|(1, 1, -2)\| = \sqrt{6}$

And, $\sqrt{|Df(\mathbf{x}_1(\mu; \mathbf{y}))\ Df^T(\mathbf{x}_1(\mu; \mathbf{y}))|} = \sqrt{\left| \begin{pmatrix} 1 & 1 & 1 \\ 1 & -1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \\ 1 & 0 \end{pmatrix} \right|} = \sqrt{6}$

Therefore, Theorem 1 gives:

$$P_{Y_1 Y_2}(y_1, y_2) = \int_{-\infty}^{\infty} P_{X_1 X_2 X_3}(\mu, \mu - y_2, y_1 + y_2 - 2\mu)\, d\mu$$

Example 5a: $y_1 = x_1 + x_2 + x_3\ ;\ y_2 = x_1 - x_2 + x_4 \quad (f : \mathbf{R}^4 \mapsto \mathbf{R}^2)$

The set $f^{-1}(\mathbf{y})$ can be parameterized as $\mathbf{x}_1(\mu_1, \mu_2; y_1, y_2) = (\mu_1, \mu_2, y_1 - \mu_1 - \mu_2, y_2 - \mu_1 + \mu_2)^T$ with $\mu_1, \mu_2 \in (-\infty, \infty)$.

$\sqrt{|D_\mu \mathbf{x}_1^T(\mu; \mathbf{y})\ D_\mu \mathbf{x}_1(\mu; \mathbf{y})|} = \sqrt{\left| \begin{pmatrix} 1 & 0 & -1 & -1 \\ 0 & 1 & -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \\ -1 & 1 \end{pmatrix} \right|} = 3$

And, $\sqrt{|Df(\mathbf{x}_1(\mu; \mathbf{y}))\ Df^T(\mathbf{x}_1(\mu; \mathbf{y}))|} = \sqrt{\left| \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \right|} = 3$

Therefore, Theorem 1 gives:

$$P_{Y_1 Y_2}(y_1, y_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P_{X_1 X_2 X_3 X_4}(\mu_1, \mu_2, y_1 - \mu_1 - \mu_2, y_2 - \mu_1 + \mu_2)\, d\mu_1\, d\mu_2$$

Example 5b: $y_1 = x_1 + x_2 + x_3\ ;\ y_2 = x_1 - x_2 + x_4 \quad (f : \mathbf{R}^4 \mapsto \mathbf{R}^2)$

This is the same function as in Example 5a. The difference in this example is how we parameterize the set $f^{-1}(\mathbf{y})$. In this example we parameterize this set in the following way: $\mathbf{x}_1(\mu_1, \mu_2; y_1, y_2) = (2\mu_1, \mu_2, y_1 - 2\mu_1 - \mu_2, y_2 - 2\mu_1 + \mu_2)^T$ with $\mu_1, \mu_2 \in (-\infty, \infty)$.

Continuing as in 5a we have: $\sqrt{|D_\mu \mathbf{x}_1^T(\mu; \mathbf{y})\ D_\mu \mathbf{x}_1(\mu; \mathbf{y})|} = \sqrt{\left| \begin{pmatrix} 2 & 0 & -2 & -2 \\ 0 & 1 & -1 & 1 \end{pmatrix} \begin{pmatrix} 2 & 0 \\ 0 & 1 \\ -2 & -1 \\ -2 & 1 \end{pmatrix} \right|} = 6$

And, $\sqrt{|Df(\mathbf{x}_1(\mu; \mathbf{y}))\ Df^T(\mathbf{x}_1(\mu; \mathbf{y}))|} = \sqrt{\left| \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & -1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \right|} = 3$

Therefore, Theorem 1 gives:

$$P_{Y_1Y_2}(y_1,y_2) = 2\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} P_{X_1X_2X_3X_4}(2\mu_1,\mu_2,y_1-2\mu_1-\mu_2,y_2-2\mu_1+\mu_2)\,d\mu_1\,d\mu_2$$

Application to the Distribution of the Ratio of Normals

We use Theorem 1 to compute the distribution of the ratio of two correlated normals.

If x_1, x_2 are Gaussian random variables with joint density function, $P_{X_1X_2}$, means: μ_1, μ_2, and positive definite covariance matrix:[1] $\Sigma = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$, then the density function of the ratio $y = x_1/x_2$ is given by the formula in example 2:

$$P_Y(y) = \int_{-\infty}^{\infty} |s| P_{X_1X_2}(sy,s)\,ds = \frac{1}{2\pi\sqrt{ac-b^2}}\int_{-\infty}^{\infty} |s| e^{-\left([sy-\mu_1,s-\mu_2]\Sigma^{-1}[sy-\mu_1,s-\mu_2]^T\right)/2}\,ds$$

This may be reduced to:

$$P_Y(y) = \frac{\sqrt{ac-b^2}e^{-\gamma}}{\pi(cy^2-2by+a)}\left[1+\sqrt{\pi}\tau(y)e^{\tau(y)^2}\operatorname{erf}(\tau(y))\right]$$

where

$$\gamma = \frac{c\mu_1^2-2b\mu_1\mu_2+a\mu_2^2}{2(ac-b^2)}$$

$$\tau(y) = \frac{|y(c\mu_1-b\mu_2)+a\mu_2-b\mu_1|}{\sqrt{cy^2-2by+a}\sqrt{2(ac-b^2)}}$$

The function $\tau(y)$ is bounded on the interval $(-\infty,\infty)$. This follows since by the positive definiteness of Σ, the denominator is bound away from 0; while, for large values of y, $\tau(y)$ is $O(1)$.

Note: As a consequence, this distribution does not have *any* moments. This follows since the function $P_Y(y)$ is $O(1/y^2)$, when $|y|$ is large. *Therefore, one cannot talk about the mean or standard deviation of this distribution.*

This distribution is also not unimodal in general.

Special Cases

The density function simplifies under the condition: $\mu_1 = \mu_2 = 0$. $P_Y(y)$ is given by:

$$P_Y(y) = \frac{\sqrt{ac-b^2}}{\pi(cy^2-2by+a)} = \frac{\sqrt{ac-b^2}}{\pi\left((\sqrt{c}y-\frac{b}{\sqrt{c}})^2+(a-\frac{b^2}{c})\right)}$$

If, in addition, x_1 and x_2 are uncorrelated, then $P_Y(y)$ becomes:

$$P_Y(y) = \frac{\sqrt{ac}}{\pi(cy^2+a)}$$

Finally, if we further set, $a = c = 1, b = 0$; that is, x_1 and x_2 are independent Gaussians each with mean 0 and variance 1, then $P_Y(y)$ becomes:

$$P_Y(y) = \frac{1}{\pi(y^2+1)}$$

[1] If Σ is positive definite we have: $a > 0,\ c > 0,\ ac-b^2 > 0$.