Supplementary Information for

Divisive normalization is an efficient code for multivariate Pareto-distributed environments

Stefan Bucher and Adam Brandenburger

Corresponding author: Stefan Bucher.
E-mail: stefan.bucher@nyu.edu

This PDF file includes:

- Supplementary text
- Figs. S1 to S9
- SI References
Supporting Information Text

Equation numbers without an “A” prefix refer to the main text.

**Proof of Proposition 1.** To see that \( r(x) \in \Delta \) for all \( x \in \mathbb{R}_+^n \) (non-negative real vectors), note that

\[
\sum_{i=1}^{n} \lambda_i r_i(x) = \sum_{i=1}^{n} \lambda_i \frac{\gamma}{b^\alpha} \frac{x_i^\alpha}{\sum_{j=1}^{n} \lambda_j x_j^\alpha} = \gamma \frac{\sum_{i=1}^{n} \lambda_i x_i^\alpha}{b^\alpha + \sum_{j=1}^{n} \lambda_j x_j^\alpha} < \gamma.
\]

We next show that for all \( y \in \Delta \) there is a unique \( x \in \mathbb{R}_+^n \) such that \( r(x) = y \), i.e., such that

\[
\frac{\gamma}{b^\alpha + \sum_{j=1}^{n} \lambda_j x_j^\alpha} = \frac{y_i}{y},
\]

for all \( i \). Letting \( z_i = \lambda_i x_i^\alpha \), \( w_i = \lambda_i y_i / \gamma \), and \( \bar{b} = b^\alpha \), this is equivalent to showing that there is a unique \( z \in \mathbb{R}_+^n \) such that

\[
\frac{z_i}{\bar{b} + \sum_{j=1}^{n} z_j} = w_i,
\]

for all \( i \), or, equivalently, that the system of equations

\[
(I_n - w 1^T) z = \bar{b} w
\]  

has a unique solution \( z \in \mathbb{R}_+^n \). The matrix determinant lemma (1, Theorem 18.1.1) states that for an invertible square matrix \( A \) and column vectors \( u \) and \( v \),

\[
\det(A + uv^T) = (1 + v^T A^{-1} u) \det(A).
\]

Setting \( A = I_n \), \( u = -w \), and \( v = 1 \), we then get

\[
\det(I_n - w 1^T) = (1 - 1^T I_n^{-1} w) \det(I_n) = 1 - \sum_{i=1}^{n} w_i > 0,
\]

since \( \sum_{i=1}^{n} w_i = \frac{1}{\gamma} \sum_{i=1}^{n} \lambda_i y_i < 1 \) for all \( y \in \Delta \). Therefore, the system of equations A.1 indeed has a unique solution, so \( r \) is invertible and its image is \( \Delta \).

**Proof of Proposition 2.** First note that (2, Eq. 8.48)

\[
I(X; g(X) + \varepsilon) = h(g(X) + \varepsilon) - h(g(X) + \varepsilon | X) = h(g(X) + \varepsilon) - h(\varepsilon | X).
\]

It follows that maximizing \( I(X; g(X) + \varepsilon) \) is equivalent to maximizing \( h(g(X) + \varepsilon) \) (cf. 3, 4). For small noise, maximizing \( I(X; g(X) + \varepsilon) \) is approximately equivalent to maximizing \( h(g(X)) \). To see that this is an arbitrarily close approximation, note that by repeated application of the chain rule (2, Theorem 17.2.2 and Problem 2.4)\

\[
h(g(X) + \varepsilon) \leq h(g(X) + \varepsilon) + h(g(X), \varepsilon | g(X) + \varepsilon) = h(g(X), \varepsilon, g(X) + \varepsilon) = h(g(X), \varepsilon) = h(g(X)) + h(\varepsilon),
\]

where the last line uses independence of \( g(X) \) and \( \varepsilon \), which further implies (2, Lemma 17.2.1)

\[
h(g(X) + \varepsilon) \geq h(g(X)),
\]

so that

\[
h(g(X)) \leq h(g(X) + \varepsilon) \leq h(g(X)) + h(\varepsilon).
\]

By independence we have \( h(\varepsilon | X) = h(\varepsilon) \) (2, Theorem 17.2.1), so it follows from Equation A.3 that

\[
I(X; g(X) + \varepsilon) = h(g(X) + \varepsilon) - h(\varepsilon),
\]

so that

\[
h(g(X)) - h(\varepsilon) \leq I(X; g(X) + \varepsilon) \leq h(g(X)),
\]

and thus

\[
|I(X; g(X) + \varepsilon) - h(g(X))| \leq h(\varepsilon) < \delta,
\]

for any \( \delta > 0 \), as long as \( h(\varepsilon) < \delta \).

*This generalizes to any function \( f(g(X), \varepsilon) \) beyond \( f(g(X), \varepsilon) = g(X) + \varepsilon \).*
Proof of Proposition 3. We want to show that the pdf \( f_Y \) of Equation 3, which equivalently satisfies
\[
\log f_Y(y) = \lambda_0 - c(y) \quad \forall y \in \mathcal{C},
\]
with \( \lambda_0 = -\log \int_{\mathcal{C}} e^{-c(z)} dz \), uniquely maximizes
\[
h(Y) - \mathbb{E}[c(Y)] = -\int_{\mathcal{C}} g(y) \log g(y) dy - \int_{\mathcal{C}} g(y) c(y) dy
\]
among all pdfs \( g \) satisfying
\[
g(y) \geq 0 \quad \text{with equality for all } y \notin \mathcal{C},
\]
\[
\int_{\mathcal{C}} g(y) dy = 1.
\]
Our proof parallels Theorem 12.1.1. of Cover and Thomas (2). Recall that the Kullback-Leibler divergence satisfies (2, eq. 8.87)
\[
D(g||f_Y) = \int_{\mathcal{C}} g(y) \log \left( \frac{g(y)}{f_Y(y)} \right) dy \geq 0,
\]
and note that
\[
-\int_{\mathcal{C}} g(y) \log g(y) dy - \int_{\mathcal{C}} g(y) c(y) dy = -\int_{\mathcal{C}} g(y) \log \left( \frac{g(y)}{f_Y(y)} \right) dy - \int_{\mathcal{C}} g(y) c(y) dy
\]
\[
= -\int_{\mathcal{C}} g(y) \log \left( \frac{g(y)}{f_Y(y)} \right) dy - \int_{\mathcal{C}} g(y) \log (f_Y(y)) dy - \int_{\mathcal{C}} g(y) c(y) dy
\]
\[
\leq -\int_{\mathcal{C}} g(y) \log (f_Y(y)) dy + c(y) dy
\]
\[
= -\lambda_0 \int_{\mathcal{C}} g(y) dy
\]
\[
= -\lambda_0 \int_{\mathcal{C}} f_Y(y) dy
\]
\[
= -\int_{\mathcal{C}} f_Y(y) [\lambda_0 - c(y)] dy
\]
\[
= -\int_{\mathcal{C}} f_Y(y) [\lambda_0 - c(y)] dy
\]
where the inequality is strict unless \( f_Y = g \) almost everywhere (2, Theorem 8.6.1). We have shown that the pdf \( f_Y \) of Equation 3 attains strictly greater \( h(Y) - \mathbb{E}[c(Y)] \) than any other pdf with support \( \mathcal{C} \).

Proof of Theorem 1. Since the function \( r \) is invertible and has continuous derivatives, a change of variables implies that
\[
f_X(x) = f_Y(r(x)) \cdot |\det (J_r(x))| \quad \forall x \in \mathbb{R}^n,
\]
whenever the determinant of the Jacobian \( J_r(x) \) of \( r \) is non-zero (5, Theorem 8.1.7).\(^\dagger\) We first compute the Jacobian of \( r(\hat{x}) = \gamma \hat{x} / d(\hat{x}) \) where \( d(\hat{x}) = b^\top + \lambda^T \hat{x} \):\(^\dagger\)
\[
J_r(\hat{x}) = \frac{\gamma}{d(\hat{x})^2} \cdot \begin{pmatrix}
d(\hat{x}) - \bar{x}_1 \lambda_1 & -\bar{x}_1 \lambda_2 & \cdots & -\bar{x}_1 \lambda_n \\
-\bar{x}_2 \lambda_1 & d(\hat{x}) - \bar{x}_2 \lambda_2 & \cdots & -\bar{x}_2 \lambda_n \\
\vdots & \vdots & \ddots & \vdots \\
-\bar{x}_n \lambda_1 & -\bar{x}_n \lambda_2 & \cdots & d(\hat{x}) - \bar{x}_n \lambda_n
\end{pmatrix} = \frac{\gamma}{d(\hat{x})^2} \cdot (d(\hat{x}) \cdot I_n - \hat{x} \lambda^T),
\]
\(^\dagger\)The relation extends to include the boundary of the domain \( \mathbb{R}^n \) (6, Corollary 15.9) and could alternatively be stated on \( \mathbb{R}^n \).
where \( I_n \) is the \( n \times n \) identity matrix. Using the matrix determinant lemma (Eq. A.2), with \( A = d(\bar{x}) \cdot I_n, u = -\bar{x}, \) and \( v = \lambda, \) we obtain

\[
\det(J_r(\bar{x})) = \left(\frac{\gamma}{d(\bar{x})^2}\right)^n \det(d(\bar{x}) \cdot I_n - \bar{x} \lambda^T)
\]

\[
= \left(\frac{\gamma}{d(\bar{x})^2}\right)^n \left(1 - \lambda^T (d(\bar{x}) \cdot I_n)^{-1} \bar{x}\right) \det(d(\bar{x}) \cdot I_n)
\]

\[
= \left(\frac{\gamma}{d(\bar{x})^2}\right)^n \left(1 - \frac{\lambda^T \bar{x}}{d(\bar{x})}\right) d(\bar{x})^n
\]

\[
= \left(\frac{\gamma}{d(\bar{x})^2}\right)^n \left(\frac{b^n - \lambda^T \bar{x}}{d(\bar{x})}\right) d(\bar{x})^n
\]

\[
= \gamma^n \frac{b^n}{(b^n + \lambda^T \bar{x})^{n+1}}.
\]

In order to find the Jacobian of \( r(x) = \bar{r}(x^\alpha), \) note that by the multivariate chain rule:

\[
(J_r(x))_{ij} = \frac{\partial}{\partial x^i} \bar{r}_j(x^\alpha) = \sum_k \frac{\partial}{\partial x^k} \bar{r}_j(\bar{x}) \frac{\partial \bar{x}_k}{\partial x^j} x_k^\alpha = \frac{\partial \bar{r}_j(\bar{x})}{\partial x^j} x_j^\alpha = (J_r(\bar{x}))_{ij} x_j^\alpha^{-1} = (J_r(x))_{ij} x_j^\alpha^{-1},
\]

so that

\[
J_r(x) = \gamma \frac{d(x^\alpha)^2}{d(x)^2} \cdot 
\begin{pmatrix}
(d(x^\alpha) - x_1^\alpha \lambda_1) \cdot x_1^{\alpha-1} & (d(x^\alpha) - x_2^\alpha \lambda_2) \cdot x_2^{\alpha-1} & \cdots & (d(x^\alpha) - x_n^\alpha \lambda_n) \cdot x_n^{\alpha-1} \\
-x_1^\alpha \lambda_1 \cdot x_1^{\alpha-1} & -x_2^\alpha \lambda_2 \cdot x_2^{\alpha-1} & \cdots & -x_n^\alpha \lambda_n \cdot x_n^{\alpha-1} \\
\vdots & \vdots & \ddots & \vdots \\
-x_1^\alpha \lambda_1 \cdot x_1^{\alpha-1} & -x_2^\alpha \lambda_2 \cdot x_2^{\alpha-1} & \cdots & (d(x^\alpha) - x_n^\alpha \lambda_n) \cdot x_n^{\alpha-1}
\end{pmatrix}
\]

and, therefore,

\[
\det(J_r(x)) = \det(J_r(x^\alpha)) \cdot \alpha^n \prod_{i=1}^n x_i^{\alpha-1} = \gamma^n \alpha^n \frac{b^n \prod_{i=1}^n x_i^{\alpha-1}}{(b^n + \sum_{i=1}^n \lambda_i x_i^n)^{n+1}}.
\]

Since \( \det(J_r(x)) > 0 \) for all \( x \in \mathbb{R}^n_+ \) we then have

\[
f(x) = f(y) \cdot \gamma^n \alpha^n \frac{b^n \prod_{i=1}^n x_i^{\alpha-1}}{(b^n + \sum_{i=1}^n \lambda_i x_i^n)^{n+1}}, \quad \forall x \in \mathbb{R}^n_+.
\]

for any positive vector \( x \in \mathbb{R}^n_+. \)

**Proof of Theorem 2.** Note that the support of \( r(x) \) is its image which, by Proposition 1, is given by \( \Delta \). From Theorem 1 combined with Proposition 3 it follows that

\[
f(x) = \gamma^n \alpha^n \frac{b^n \prod_{i=1}^n x_i^{\alpha-1}}{(b^n + \sum_{i=1}^n \lambda_i x_i^n)^{n+1}} \exp(-c(r(x))) \quad \forall x \in \mathbb{R}^n_+,
\]

is a necessary and sufficient condition for \( r \) to maximize the entropy of the output distribution net of expected costs over all representations with support \( \Delta \). Setting \( \alpha \equiv \beta \) and \( b/\lambda_1^{1/\alpha} \equiv \beta_1 \) we get

\[
f(x) = \gamma^n \beta^n \frac{b^n \prod_{i=1}^n x_i^{\beta-1}}{(b^{n+1} + \sum_{i=1}^n (x_i/\beta)^{n+1})} \exp(-c(r(x))) \quad \forall x \in \mathbb{R}^n_+.
\]

The result follows from the fact that for any constant translation \( \mu \in \mathbb{R}^n \)

\[
fs(s) = f(x) \cdot \mu = \gamma^n \beta^n \frac{\prod_{i=1}^n (s_i - \mu_i)^{\beta-1} / b^\beta}{(1 + \sum_{i=1}^n (x_i/\beta)^{n+1})} \exp(-c(r(s - \mu))) \quad \forall s \geq \mu.
\]

4 of 17  Stefan Bucher and Adam Brandenburger
Proof of Theorem 3. In the special case of constant costs \( c(y) = \bar{c} \) for all \( y \in \Delta \), Equation 5 of Theorem 2 reduces to

\[
fs(s) = \gamma^n \beta^n \frac{\prod_{i=1}^{n} (s_i - \mu_i)^{\beta - 1} / b^\beta}{\left(1 + \sum_{i=1}^{n} \left(\frac{s_i - \mu_i}{\sigma_i}\right)^\beta\right)^{n+1}} \exp\left(-\bar{c}\right) \frac{1}{s^\beta} \int_{\Delta} \exp\left(-\bar{c}\right) dz \quad s > \mu,
\]

where we have used \( b^\beta = \lambda_i \sigma_i^\beta \) and where

\[
\int_{\Delta} dz = \frac{\gamma^n}{\lambda_1 \lambda_2 \ldots \lambda_n \cdot n!},
\]

so that

\[
fs(s) = \gamma^n \beta^n \frac{\prod_{i=1}^{n} (s_i - \mu_i)^{\beta - 1} / b^\beta}{\left(1 + \sum_{i=1}^{n} \left(\frac{s_i - \mu_i}{\sigma_i}\right)^\beta\right)^{n+1}} \gamma^n \prod_{i=1}^{n} \lambda_i = \beta^n \frac{n! \prod_{i=1}^{n} \lambda_i = \beta^n \frac{n! \prod_{i=1}^{n} \lambda_i}{\left(1 + \sum_{i=1}^{n} \left(\frac{s_i - \mu_i}{\sigma_i}\right)^\beta\right)^{n+1}} \quad s > \mu. \tag{7}
\]

It remains to show that Equation 7 is the pdf of a multivariate Pareto type III distribution with joint survival function (7, Eq. 6.1.17 with \( \gamma_i = 1/\beta \) for \( i = 1, \ldots, n \),

\[
\hat{F}_s(s; \mu, \sigma, \beta) = \left[1 + \sum_{i=1}^{n} \left(\frac{s_i - \mu_i}{\sigma_i}\right)^\beta\right]^{-1} \quad s > \mu. \tag{6}
\]

To see this, define

\[
\hat{F}_s^{(k)}(s_1, \ldots, s_n; \mu, \sigma, \beta) \equiv \frac{\partial^k}{\partial s_1 \cdots \partial s_k} \hat{F}_s(s_1, \ldots, s_n; \mu, \sigma, \beta) = \frac{\partial}{\partial s_k} \hat{F}_s^{(k-1)}(s_1, \ldots, s_n; \mu, \sigma, \beta),
\]

and note that, for \( k = 0, 1, 2, \ldots, n \),

\[
\hat{F}_s^{(k)}(s_1, \ldots, s_n; \mu, \sigma, \beta) = (-1)^k \left(\prod_{i=1}^{k} \frac{\beta \cdot i}{\sigma_i} \left(\frac{s_i - \mu_i}{\sigma_i}\right)^{\beta - 1}\right) \left[1 + \sum_{i=1}^{n} \left(\frac{s_i - \mu_i}{\sigma_i}\right)^\beta\right]^{-(k+1)} \tag{A.4}
\]

This can be shown by induction. Note that \( \hat{F}_s^{(0)}(s_1, \ldots, s_n; \mu, \sigma, \beta) = \hat{F}_s(s_1, \ldots, s_n; \mu, \sigma, \beta) \) so that Equation A.4 holds for \( k = 0 \), and observe that if Equation A.4 holds for \( k - 1 \), then

\[
\hat{F}_s^{(k)}(s_1, \ldots, s_n; \mu, \sigma, \beta) = \left(-1\right)^{k-1} \left(\prod_{i=1}^{k-1} \frac{\beta \cdot i}{\sigma_i} \left(\frac{s_i - \mu_i}{\sigma_i}\right)^{\beta - 1}\right) \left[1 + \sum_{i=1}^{n} \left(\frac{s_i - \mu_i}{\sigma_i}\right)^\beta\right]^{-(k+1)} \left[\frac{\partial}{\partial s_k} \hat{F}_s^{(k-1)}(s_1, \ldots, s_n; \mu, \sigma, \beta) \right]
\]

so that Equation A.4 also holds for \( k \).

Next, note that, given \( s_1, \ldots, s_n \) and denoting, for any \( i \), the event \( S_i > s_i \) by \( A_i \) and its complement by \( A_i^c \), the cdf is obtained from the survival function as

\[
Fs(s_1, \ldots, s_n) = \mathbb{P}(S_1 \leq s_1, \ldots, S_n \leq s_n) = \mathbb{P}\left(\bigcap_{i=1}^{n} A_i^c\right) = 1 - \mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = 1 - \sum_{k=1}^{n} (-1)^{k-1} \sum_{I \subseteq \{1, \ldots, n\} \setminus \{i\} = k} \mathbb{P}\left(\bigcap_{i \in I} A_i\right) = 1 + \sum_{k=1}^{\infty} (-1)^{k} \sum_{I \subseteq \{1, \ldots, n\} \setminus \{i\} = k} \lim_{s_i \to -\infty \forall i \notin I} \hat{F}_s(s_1, \ldots, s_n),
\]

\footnote{The minus sign in (7, Eq. 6.1.17) appears to be a typo.}
where the fourth equality follows from the probabilistic version of the inclusion-exclusion principle. Therefore

\[
\frac{\partial^n}{\partial s_1 \cdots \partial s_n} \tilde{F}_s(s_1, \ldots, s_n) = \frac{\partial^n}{\partial s_1 \cdots \partial s_n} \left[ 1 + \sum_{k=1}^{n} (-1)^k \sum_{I \subseteq \{1, \ldots, n\}} \left| I \right| = k \lim_{t \to -\infty} \tilde{F}_s(s_1, \ldots, s_n) \right]
\]

\[
= (-1)^n \frac{\partial^n}{\partial s_1 \cdots \partial s_n} \tilde{F}_s(s_1, \ldots, s_n),
\]

so, using Equation A.4 with \( k = n \), we find that the pdf associated with the survival function of Equation 6 is given by

\[
f_s(s_1, \ldots, s_n; \mu, \sigma, \beta) = \frac{\partial^n}{\partial s_1 \cdots \partial s_n} \tilde{F}_s(s_1, \ldots, s_n; \mu, \sigma, \beta)
\]

\[
= (-1)^n \tilde{F}_s^{(n)}(s_1, \ldots, s_n; \mu, \sigma, \beta)
\]

\[
= \left( \prod_{i=1}^{n} \frac{\beta - i}{\sigma_i} \left( \frac{s_i - \mu_i}{\sigma_i} \right)^{\beta - 1} \right) \left[ 1 + \sum_{i=1}^{n} \left( \frac{s_i - \mu_i}{\sigma_i} \right)^{\beta} \right]^{-(n+1)}
\]

\[
= \beta^n \prod_{i=1}^{n} \frac{1}{\sigma_i} \left( \frac{s_i - \mu_i}{\sigma_i} \right)^{\beta - 1} \left[ 1 + \sum_{i=1}^{n} \left( \frac{s_i - \mu_i}{\sigma_i} \right)^{\beta} \right]^{n+1},
\]

\( s > \mu \),

which indeed coincides with Equation 7.

**Derivation of the Marginal Distribution (Equation 8).** The survival function of the marginal distribution is

\[
\tilde{F}_{s_i}(s_i; \mu_i, \sigma_i, \beta) = \tilde{F}_s(s_1 = \mu_1, \ldots, s_{i-1} = \mu_{i-1}, s_i, s_{i+1} = \mu_{i+1}, \ldots, s_n = \mu_n; \mu, \sigma, \beta)
\]

\[
= \left[ 1 + \left( \frac{s_i - \mu_i}{\sigma_i} \right)^{\beta} \right]^{-1}, \quad s_i > \mu_i,
\]

so its cdf is obtained as

\[
F_{s_i}(s_i; \mu_i, \sigma_i, \beta) = 1 - \tilde{F}_{s_i}(s_i; \mu_i, \sigma_i, \beta) = 1 - \frac{1}{1 + \left( \frac{s_i - \mu_i}{\sigma_i} \right)^{\beta}} = \frac{\left( \frac{s_i - \mu_i}{\sigma_i} \right)^{\beta}}{1 + \left( \frac{s_i - \mu_i}{\sigma_i} \right)^{\beta}},
\]

\[8\]

and its pdf is

\[
f_{s_i}(s_i; \mu_i, \sigma_i, \beta) = \beta \frac{1}{\sigma_i} \left( \frac{s_i - \mu_i}{\sigma_i} \right)^{\beta - 1} \left[ 1 + \left( \frac{s_i - \mu_i}{\sigma_i} \right)^{\beta} \right]^{-2},
\]

which is a univariate Pareto type III distribution. Its mode, for \( \beta > 1 \), is \( (7, \text{Eq. 3.3.4}) \)

\[
\mu_i + \sigma_i \left( \frac{\beta - 1}{\beta + 1} \right)^{1/\beta}.
\]

**Derivation of Moments (Including Equations 9 and 11).** From \( (7, \text{Eq. 6.1.27}) \) it follows,\(^6\) with \( \alpha = 1 \) and \( \gamma = 1/\beta \) and using \( \Gamma(1) = 1 \), that for \( \beta > 1 \) the mean is given by

\[
E[S_i] = \mu_i + \sigma_i \Gamma \left( \frac{\beta - 1}{\beta} \right) \Gamma \left( \frac{\beta + 1}{\beta} \right) = \mu_i + \sigma_i \frac{\pi/\beta}{\sin(\pi/\beta)},
\]

\[9\]

where \( \Gamma \) is the Gamma function and the second equality follows from its recursive expression and Euler’s reflection formula, which implies that

\[
\Gamma(1 + z)\Gamma(1 - z) = \pi \Gamma(z) \Gamma(1 - z) = \pi \frac{\Gamma(z)}{\sin(\pi z)},
\]

for all \( z \in \mathbb{Z} \). From \( (7, \text{Eq. 6.1.22}) \) it follows that the conditional mean, for \( \beta > 1/n \), is given by

\[
E[S_i | \{ S_j = s_j \}_{j \neq i}] = \mu_i + \sigma_i \left[ 1 + \sum_{j \neq i} \left( \frac{s_j - \mu_j}{\sigma_j} \right)^{\beta} \right]^{1/\beta} \frac{\Gamma(n - 1/\beta) \Gamma \left( \frac{\beta + 1}{\beta} \right)}{\Gamma(n)}.
\]

\(^6\)We write \( \alpha \) for the \( \alpha \) in \( (7) \), to differentiate it from the exponent in the divisive normalization transform.
From (7, Eq. 3.3.11) it follows that, for $\beta > 2$, the variance of the Pareto type III distribution is
\[
\text{Var}(S_i) = \sigma_i^2 \left[ \frac{1}{\beta} \right] = \sigma_i^2 \left[ \frac{2\pi/\beta}{\sin(2\pi/\beta)} - \left( \frac{\pi/\beta}{\sin(\pi/\beta)} \right)^2 \right].
\]
The covariance (Equation 11) is obtained (7, Eq. 6.1.29), for $\beta > 2$, as
\[
\text{Cov}(S_i, S_j) = \sigma_i \sigma_j \left( \frac{1}{\beta} \right)^2 \left( \frac{2\pi/\beta}{\sin(2\pi/\beta)} - \left( \frac{\pi/\beta}{\sin(\pi/\beta)} \right)^2 \right), \quad i \neq j,
\]
so that the correlation coefficient is
\[
\text{Corr}(S_i, S_j) = \frac{\text{Cov}(S_i, S_j)}{\sqrt{\text{Var}(S_i) \text{Var}(S_j)}} = \frac{\Gamma \left( \frac{\beta+2}{\beta} \right) - \left( \Gamma \left( \frac{\beta+1}{\beta} \right) \right)^2}{\Gamma \left( \frac{\beta+2}{\beta} \right) - \left( \Gamma \left( \frac{\beta+1}{\beta} \right) \right)^2}, \quad i \neq j.
\]

### Derivation of the Conditional Variance for $\beta = 1$ (Equation 12)

The variance of a univariate Pareto type II distribution with cumulative distribution function (7, Eq. 3.2.2)
\[
P^{(1)}(x; \mu, \sigma, \alpha) = 1 - \left[ 1 + \frac{x - \mu}{\sigma} \right]^{-\alpha}, \quad x > \mu
\]
is, for any $\alpha > 2$, given by (7, Eq. 3.3.13)
\[
\text{Var}(X) = \sigma^2 \left[ \frac{\Gamma(\hat{\alpha} - 2) \Gamma(3)}{\Gamma(\hat{\alpha})} - \left( \frac{\Gamma(\hat{\alpha} - 1) \Gamma(2)}{\Gamma(\hat{\alpha})} \right) \right]^2
\]
\[
= \sigma^2 \left[ \frac{\Gamma(3)}{(\hat{\alpha} - 1)(\hat{\alpha} - 2)} - \left( \frac{\Gamma(2)}{(\hat{\alpha} - 1)^2} \right) \right] = \sigma^2 \frac{\hat{\alpha}}{(\hat{\alpha} - 1)^2(\hat{\alpha} - 2)}.
\]
From Equation 13 (see below) we know that the conditional distribution of a multivariate Pareto III distribution with $\beta = 1$ is a univariate Pareto type II distribution with location $\mu = \mu_i$, scale $\sigma = \sigma_i \left[ 1 + \sum_{j \neq i} \left( \frac{s_j - \mu_j}{\sigma_j} \right) \right]$, and shape parameter $\alpha = n$. For any $n > 2$ we thus have
\[
\text{Var}(S_i | \{ S_j \}_{j \neq i}) = \sigma_i^2 \left[ \frac{1}{n - 1} \sum_{j \neq i} \left( \frac{s_j - \mu_j}{\sigma_j} \right) \right]^2
\]
\[
= \sigma_i^2 \left[ \frac{1}{n - 1} \sum_{j \neq i} \left( \frac{s_j - \mu_j}{\sigma_j} \right) \right]^2 \left[ 1 + \sum_{j \neq i} \left( \frac{s_j - \mu_j}{\sigma_j} \right) \right]^2,
\]
where we have used
\[
\left( 1 + \sum_{j=1}^{n-1} a_j \right)^2 = 1 + 2 \sum_{j=1}^{n-1} a_j + \left( \sum_{j=1}^{n-1} a_j \right)^2 = 1 + 2 \sum_{j=1}^{n-1} a_j + \sum_{j=1}^{n-1} \sum_{k=1, k \neq j}^{n-1} a_j a_k + \sum_{j=1}^{n-1} a_j^2.
\]

### Derivation of the Conditional Distribution (Equation 13)

From (7, Eq. 6.1.20), with $\alpha = 1$ and $\gamma_i = 1/\beta$, it follows that the conditional distribution of the multivariate Pareto type III distribution of Equation 6 is a univariate Pareto type IV distribution, $S_i | \{ S_j \}_{j \neq i} \sim P^{(IV)} \left( s_i; \mu_i, \sigma, \gamma_i \right) = \sigma_i \left[ 1 + \sum_{j \neq i} \left( \frac{s_j - \mu_j}{\sigma_j} \right) \right]^{\beta/\gamma}, \gamma = 1/\beta, \alpha = 1 + n - 1,
\]
where the cdf of a univariate Pareto type IV distribution is (7, Eq. 3.2.8)
\[
P^{(IV)}(x; \mu, \sigma, \gamma, \alpha) = 1 - \left[ 1 + \left( \frac{x - \mu}{\sigma} \right)^{1/\gamma} \right]^{-\alpha}, \quad x > \mu.
\]
This implies that the conditional distribution has cdf, for $s_i > \mu_i$,
\[
F_{S_i | \{ S_j \}_{j \neq i}} (s_i; \{ s_j \}_{j \neq i}, \mu_i, \sigma, \beta) = 1 - \left[ 1 + \frac{s_i - \mu_i}{\sigma_i \left[ 1 + \sum_{j \neq i} \left( \frac{s_j - \mu_j}{\sigma_j} \right) \right]^{\beta/\gamma}} \right]^{-\beta/\gamma} = 1 - \left[ 1 + \frac{\left( \frac{s_i - \mu_i}{\sigma_i} \right)^\beta}{1 + \sum_{j \neq i} \left( \frac{s_j - \mu_j}{\sigma_j} \right)^\beta} \right]^{-n}.
\]
\[\text{A minus sign appears to be missing in (7, Eq. 6.1.28).}\]
Proof of Proposition 4. We have $U_i \sim \text{Exp}(\lambda = 1)$ with c.d.f. $P(U_i \leq u_i) = 1 - \exp(-u_i)$ and, independently of all $U_i$, $Z \sim \text{Exp}(\lambda = 1)$ with p.d.f. $f_Z(z) = \exp(-z)$. This implies

\[
P(S_1 > s_1, \ldots, S_n > s_n) = P\left(U_1 > Z \left(\frac{s_1 - \mu_1}{\sigma_1}\right)\beta, \ldots, U_n > Z \left(\frac{s_n - \mu_n}{\sigma_n}\right)\beta \gada {\text{if} \ Z = z} \exp(-z) dz\right)
\]

\[
= \int_0^\infty \exp(-z) \prod_{i=1}^n P\left(U_i > Z \left(\frac{s_i - \mu_i}{\sigma_i}\right)\beta \gada {\text{if} \ Z = z} \exp(-z) dz\right) dz
\]

\[
= \int_0^\infty \exp(-z) \prod_{i=1}^n \exp\left(-z \left(\frac{s_i - \mu_i}{\sigma_i}\right)\beta\right) dz
\]

\[
= \int_0^\infty \exp\left(-z \left[1 + \sum_{i=1}^n \left(\frac{s_i - \mu_i}{\sigma_i}\right)\beta\right]\right) dz
\]

\[
= \left[1 + \sum_{i=1}^n \left(\frac{s_i - \mu_i}{\sigma_i}\right)\beta\right]^{-1}
\]

where the third equality follows from the independence of $Z$ and all $U_i$.

References

1. DA Harville, Matrix algebra from a statistician’s perspective. (Springer-Verlag, New York), (2008).
2. TM Cover, JA Thomas, Elements of information theory. (Wiley-Interscience, Hoboken, NJ), (2006).
3. JP Nadal, N Parga, Nonlinear neurons in the low-noise limit: a factorial code maximizes information transfer. Network: Comput. Neural Syst. 5, 565–581 (1994).
4. AJ Bell, TJ Sejnowski, An Information-Maximization Approach to Blind Separation and Blind Deconvolution. Neural Comput. 7, 1129–1159 (1995).
5. JK Blitzstein, J Hwang, Introduction to Probability. (Chapman and Hall/CRC), 2nd edition, (2019).
6. RL Schilling, Measures, Integrals and Martingales. (Cambridge University Press), (2005).
7. BC Arnold, Pareto Distributions. (Chapman and Hall/CRC, Boca Raton), 2nd edition, (2015).
Fig. S1. Histogram of estimates of the Pareto distribution's parameter $\beta$ across images, from the empirical analysis of filter responses to naturalistic images.
Fig. S2. Histograms showing the negative log-likelihood of the best-fitting Pareto (top) and multivariate-t (bottom) models to the statistics of naturalistic images. The counts on the y-axis correspond to the number of images, error bars show the standard error of the mean. The figure is analogous to Figure 3A in the main text but shows negative log-likelihoods, so that the goodness of fit is decreasing along the x-axis.
Fig. S3. Joint density of the Pareto distribution for $n = 2$ and a range of values of $\beta$. Row-wise from left to right: $\beta = 0.5, 0.75, 1, 2, 3, 5, 10, 20.$
Fig. S4. Joint histogram of 100000 draws of the Pareto distribution with \( n = 2 \) for a range of values of \( \beta \). Row-wise from left to right: \( \beta = 0.5, 0.75, 1, 2, 3, 5, 10, 20 \).
Fig. S5. Conditional density where $s$ follows a bivariate Pareto distribution with $\mu = 0$, $\sigma = 1$, and a range of values of $\beta$. Row-wise from left to right: $\beta = 0.5, 0.75, 1, 2, 3, 5, 10, 20$. 

Stefan Bucher and Adam Brandenburger 13 of 17
Fig. S6. Conditional histogram where \( s \) follows a bivariate Pareto distribution with \( \mu = 0, \sigma = 1 \), and a range of values of \( \beta \). Row-wise from left to right: \( \beta = 0.5, 0.75, 1, 2, 3, 5, 10, 20 \).
Fig. S7. Conditional density of a bivariate Pareto distribution extended to $\mathbb{R}^2$, with $\mu = 0$, $\sigma = 1$, and $\beta = 1$. 
Fig. S8. Conditional histogram of a bivariate Pareto distribution extended to $\mathbb{R}^2$, with $\mu = 0$, $\sigma = 1$, and $\beta = 1$. 
Fig. S9. Divisive normalization transform for $n = 2$ and a range of values of $\alpha$. Row-wise from left to right: $\alpha = 0.5, 0.75, 1, 2, 3, 5, 10, 20$. 