Maximum entropy methods for generating simulated rainfall

Julia Piantadosi

Co-authors Phil Howlett, Jonathan Borwein and John Henstridge

Centre for Industrial and Applied Mathematics

University of South Australia

July 12, 2013

Abstract

- We shall find a multi-dimensional checkerboard copula of maximum entropy that matches an observed set of grade correlation coefficients. This problem is formulated as the maximization of a concave function on a convex polytope.
- Under mild constraint qualifications we show that a unique solution exists in the core of the feasible region.
- The theory of Fenchel duality is used to reformulate the problem as an unconstrained minimization which is well solved numerically using a Newton iteration.
- Finally, we discuss the numerical for some hypothetical examples and describe how this work can be applied to the modelling and simulation of monthly rainfall.

Modelling Accumulated Rainfall

- It has been usual to model both short-term and long-term rainfall accumulations at a specific location by a gamma distribution.
- We have observed that simulations in which monthly rainfall totals are modelled as mutually independent gamma random variables generate accumulated bi-monthly, quarterly and yearly totals having lower variance than the observed accumulations.
- We surmise that the variance of the generated totals will by increased if the model includes an appropriate level of positive correlation between individual months totals.

More generally . . .

The problem we address is:

> *how to construct a joint probability distribution which preserves the known marginal distributions and matches the observed grade correlations.*

More generally . . .

The problem we address is:

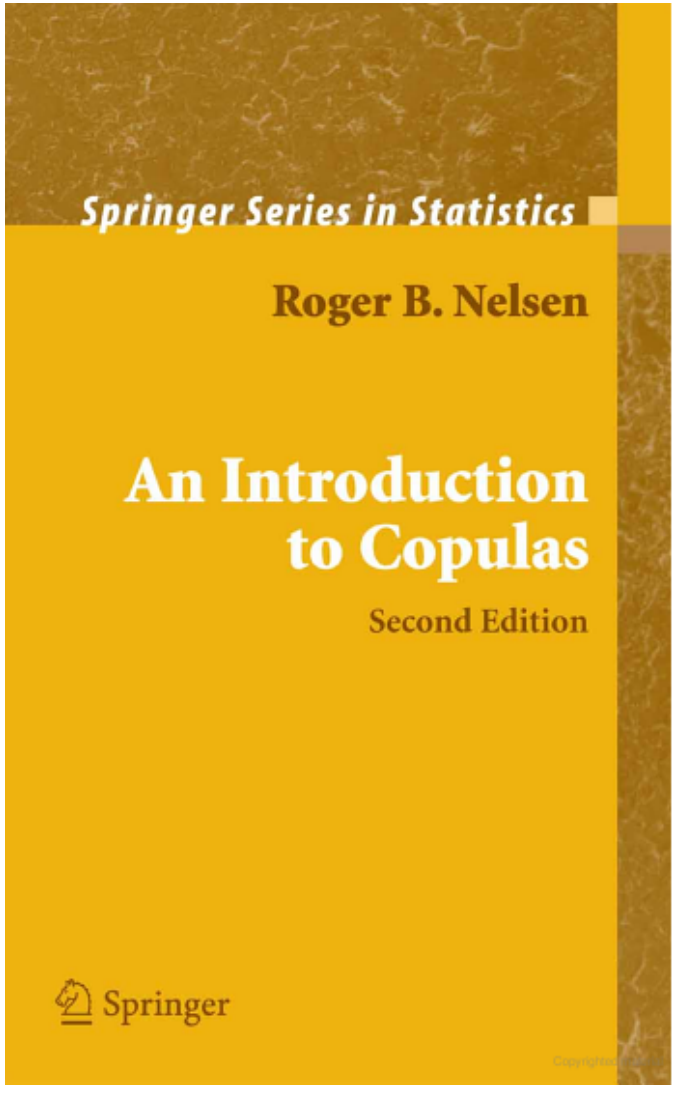
> *how to construct a joint probability distribution which preserves the known marginal distributions and matches the observed grade correlations.*

The method of copulas is one such possible solution method.

What exactly are copulas?

Copulas are functions that join or "couple" multivariate distribution functions to their one-dimensional marginal distributions.

(Inverse Problems)

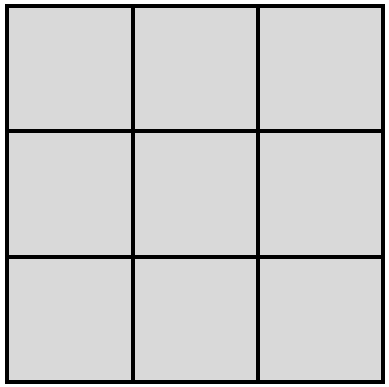


See also:
JMB, A.S. Lewis, and R. Nussbaum, "Entropy minimization, DAD problems and stochastic kernels," *J. Functional Analysis*, **123** (1994), 264–307.

Checkerboard copula

An m-dimensional *checkerboard* copula is a distribution with a corresponding density defined almost everywhere by a step function on an m-uniform subdivision of the hypercube $[0,1]^m$.

Example



Simple 2-D example
with 3 subdivisions

Of course we expect a better solution if we increase the number of subdivisions in the checkerboard.

An elementary form of the joint density

Let $\mathbf{h}$ be a non-negative m–dimensional hyper–matrix given by $\mathbf{h} = [h_{\mathbf{i}}] \in \mathbb{R}^{\ell}$ where $\ell = n^m$ and $\mathbf{i} \in \{1, \ldots, n\}^m$ with $h_{\mathbf{i}} \in [0, 1]$. Define the *marginal sums* $\sigma_r : \{1, \ldots, n\} \mapsto \mathbb{R}$ by

$$\sigma_r(i_r) = \sum_{j \neq r, i_j \in \{1,2,\ldots,n\}} h_{\mathbf{i}}$$

for each $r = 1, 2, \ldots, m$. If $\sigma_r(i_r) = 1$ for all $i_r \in \{1, 2, \ldots, n\}$ then $\mathbf{h}$ is *multiply stochastic*.
Define the partition $0 = a(1) < a(2) < \cdots < a(n) < a(n+1) = 1$ of the interval $[0, 1]$ by setting $a(k) = (k-1)/n$ for each $k = 1, \ldots, n+1$ and define a step function $c_{\mathbf{h}} : [0, 1]^m \mapsto \mathbb{R}$ by the formula

$$c_{\mathbf{h}}(\mathbf{u}) = n^{m-1} \cdot h_{\mathbf{i}} \quad \text{if} \quad \mathbf{u} \in I_{\mathbf{i}} = \times_{r=1}^{m} (a(i_r), a(i_r + 1))$$

for each $\mathbf{i} = (i_1, \ldots, i_m) \in \{1, 2, \ldots, n\}^m$.

The grade correlation coefficients

The *grade correlation coefficient* for continuous random variables X_r and X_s where $r < s$ is defined as the correlation for the grade random variables $U_r = F_r(X_r)$ and $U_s = F_s(X_s)$ by the formula

$$\begin{aligned} \rho_{r,s} &= \frac{E[(U_r - 1/2)(U_s - 1/2)]}{\sqrt{E[(U_r - 1/2)^2] \cdot E[(U_s - 1/2)^2]}} \\ &= 12(E[U_r U_s] - 1/4). \end{aligned}$$

A copula with fixed grade correlations

The grade correlation coefficient is

$$\rho_{r,s} = 12 \left(\frac{1}{n^3} \sum_{\mathbf{i} \in \{1,\ldots,n\}^m} h_{\mathbf{i}} (i_r - 1/2)(i_s - 1/2) \; - \; \frac{1}{4} \right)$$

where

$$-1 + \frac{1}{n^2} \leq \rho_{r,s} \leq 1 - \frac{1}{n^2}.$$

Copulas of Maximum Entropy

There are many copulas that could be used to construct a joint pdf and match the known grade correlation coefficients.

Copulas of Maximum Entropy

There are many copulas that could be used to construct a joint pdf and match the known grade correlation coefficients.

What we wish to do is avoid unnecessary assumptions.

Copulas of Maximum Entropy

There are many copulas that could be used to construct a joint pdf and match the known grade correlation coefficients.

What we wish to do is avoid unnecessary assumptions.
While producing a computationally efficient answer.

Copulas of Maximum Entropy

There are many copulas that could be used to construct a joint pdf and match the known grade correlation coefficients.

What we wish to do is avoid unnecessary assumptions.
While producing a computationally efficient answer.

Copulas of Maximum Entropy

There are many copulas that could be used to construct a joint pdf and match the known grade correlation coefficients.

What we wish to do is avoid unnecessary assumptions.
While producing a computationally efficient answer.

- We will construct a copula of maximum entropy (or maximum disorder) that satisfies the grade correlation constraint.
- Hence by seeking to maximise the entropy we add as little information to the system as possible.

Cartoon by Sidney Harris

The checkerboard *copula of maximum entropy* is the checkerboard copula $C_{\mathbf{h}}$ defined by the hyper-matrix $\mathbf{h}$ that solves the following convex programming problem.

The primal problem

Find the hyper-matrix $\mathbf{h} \in \mathbb{R}^{\ell}$ to maximize the entropy

$$J(\mathbf{h}) = (-1)\left[\frac{1}{n} \sum_{\mathbf{i} \in \{1,\ldots,n\}^m} h_{\mathbf{i}} \log_e h_{\mathbf{i}} + (m-1)\log_e n\right]$$

subject to

$$\sum_{j \neq r, i_j \in \{1,\ldots,n\}} h_{\mathbf{i}} = 1, \quad \forall\ i_r \in \{1,\ldots,n\},\ r = 1,\ldots,m$$

and

$$h_{\mathbf{i}} \geq 0 \quad \forall\ \mathbf{i} \in \{1,\ldots,n\}^m$$

and the additional grade correlation coefficient constraints

$$12\left[\frac{1}{n^3} \cdot \sum_{\mathbf{i} \in \{1,\ldots,n\}^m} h_{\mathbf{i}}(i_r - 1/2)(i_s - 1/2)\right] - 3 = \rho_{r,s}$$

for $1 \leq r < s \leq m$ where $\rho_{r,s}$ is known for all $1 \leq r < s \leq m$.

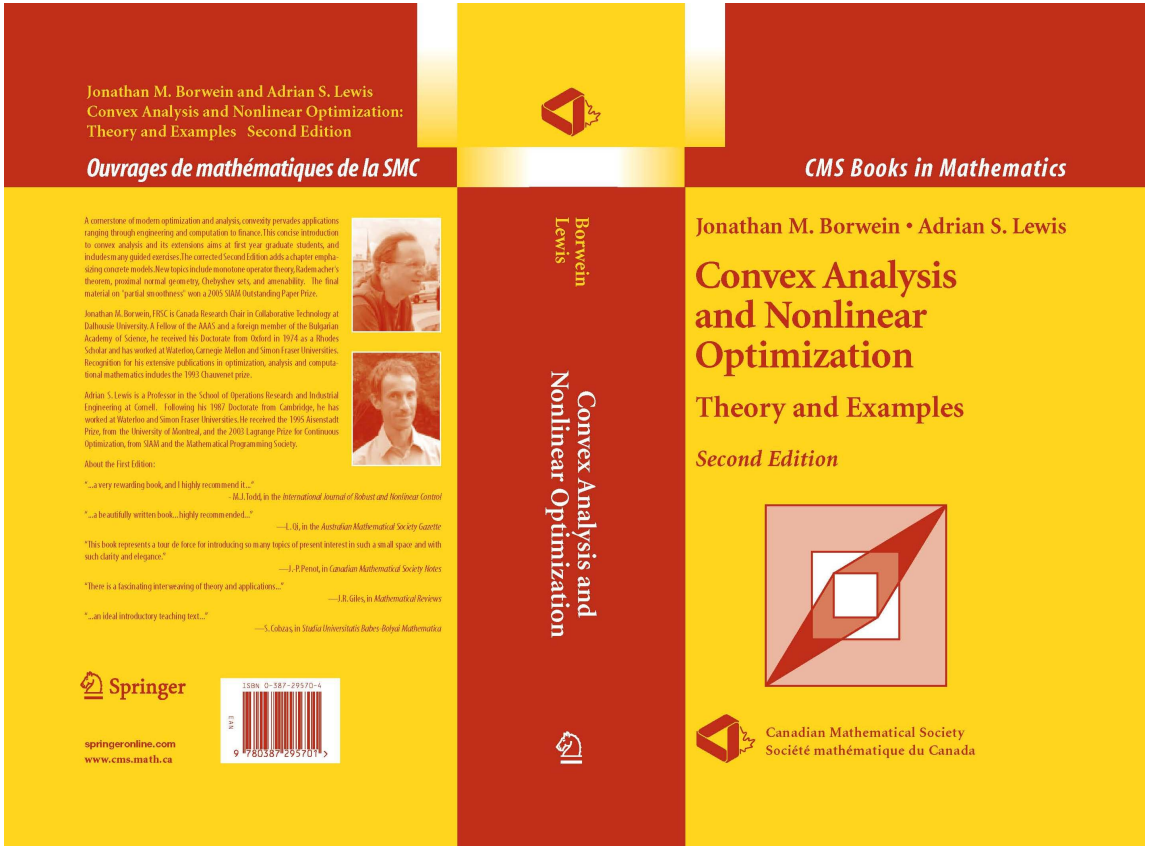
Solution procedure

We note that the problem is well posed. Nevertheless, it is not easy to compute a numerical solution directly.

Solution procedure

We note that the problem is well posed. Nevertheless, it is not easy to compute a numerical solution directly.

In fact it is much easier to solve the problem using the theory of Fenchel duality.



Borwein & Lewis (2006): Formulation and solution of the Fenchel dual problem

Let us define $g : \mathbb{R}^{\ell} \mapsto [0, \infty) \cup \{+\infty\}$ by setting

$$g(\mathbf{h}) := \begin{cases} (-1)J(\mathbf{h}) & \text{if } h_{\mathbf{j}} \geq 0 \text{ for all } \mathbf{j} \in \{1, 2, \ldots, m\}^n \\ +\infty & \text{otherwise} \end{cases}$$

where we have used the convention that $h \log_e h = 0$ when $h = 0$ and where we will allow functions to take values in the extended set of real numbers.

Mathematical statement of the primal problem

Find

$$\inf_{\mathbf{h} \in \mathbb{R}^{\ell}} \left\{ g(\mathbf{h}) \mid A\mathbf{h} = b \right\}. \tag{1}$$

If we assume that (1) has a unique solution $\mathbf{h} \in \mathcal{F}$ with $h_{\mathbf{j}} > 0$ for all $\mathbf{j} \in \{1, 2, \ldots, m\}^n$ then the Fenchel dual problem is an unconstrained maximization and the solution to the primal problem can be recovered from the solution to the dual problem.

The Fenchel conjugate

The *Fenchel conjugate* of the function g is the function $g^* : \mathbb{R}^\ell \mapsto \mathbb{R} \cup \{-\infty\}$ defined by

$$g^*(\mathbf{k}) := \sup_{\mathbf{h} \in \mathbb{R}^\ell} \left\{ \langle \mathbf{k}, \mathbf{h} \rangle - g(\mathbf{h}) \right\}.$$

For each fixed $\mathbf{k} \in \mathbb{R}^\ell$ we define

$$G(\mathbf{h}) = \sum_{\mathbf{i} \in \{1,\ldots,n\}^m} k_{\mathbf{i}} h_{\mathbf{i}} - \frac{1}{n} \sum_{\mathbf{i} \in \{1,\ldots,n\}^m} \left(h_{\mathbf{i}} \log_e h_{\mathbf{i}} - h_{\mathbf{i}} \right) - (m-1) \log_e n$$

where we note that $\sum_{\mathbf{i} \in \{1,2,\ldots,m\}^n} h_{\mathbf{i}} = n$:.

The Fenchel conjugate cont.

We can now use elementary calculus to show that $G(\mathbf{h})$ is maximized when $h_{\mathbf{i}} = \exp[nk_{\mathbf{i}}]$ and hence find that

$$g^*(\mathbf{k}) = \frac{1}{n} \sum_{\mathbf{i} \in \{1,\ldots,n\}^m} \exp[nk_{\mathbf{i}}] - (m-1)\log_e n.$$

Note that $A^* \in \mathbb{R}^{\ell \times k}$.

Using Corollary 3.3.11 from Borwein & Lewis (2006) we can now write a mathematical statement of the dual problem in standard form.

The Fenchel dual problem

Find

$$\sup_{\varphi \in \mathbb{R}^k} \left\{ \langle \mathbf{b}, \varphi \rangle - g^*(A^*\varphi) \right\}.$$

Let

$$H(\varphi) := \sum_{j=1}^{k} b_j \varphi_j - \frac{1}{n} \sum_{i=1}^{\ell} \exp\left[n \cdot \sum_{j=1}^{k} a_{ij}^* \varphi_j \right] + (m-1) \log_e n$$

and use elementary calculus once again to show that if the maximum of $H(\varphi)$ occurs when $\varphi = \overline{\varphi}$ then

$$\sum_{i=1}^{\ell} a_{ir}^* \exp\left[n \cdot \sum_{j=1}^{k} a_{ij}^* \overline{\varphi}_j \right] = b_r$$

for all $r = 1, 2, \ldots, k$.

A solution scheme for the dual problem

The key equations are written in the form:

$$\mathbf{q}(\varphi) = \mathbf{0}$$

where

$$q_r(\varphi) := \sum_{i=1}^{\ell} a_{ir}^* \exp\left[n \cdot \sum_{j=1}^{k} a_{ij}^* \overline{\varphi}_j \right] - b_r$$

for each $r = 1, 2, \ldots, k$. Now the Newton iteration can be written as

$$\varphi^{(j+1)} = \varphi^{(j)} - J^{-1}[\varphi^{(j)}]\mathbf{q}[\varphi^{(j)}]$$

using the MATLAB inverse of the Jacobian matrix $J \in \mathbb{R}^{k \times k}$.

Recovering the primal solution

In general there is a closed form for the primal solution $\overline{\mathbf{h}}$. Let $\overline{\mathbf{k}} = A^* \overline{\varphi}$ and suppose $\overline{k}_{\mathbf{j}} > 0$ for all $\mathbf{j} \in \{1, 2, \ldots, m\}^n$. Then the unique solution to the primal problem is given by

$$\overline{\mathbf{h}} = \nabla g^*(A^* \overline{\varphi}).$$

Reference: Theorem 3.3.5, Borwein, J. and Lewis A., Convex Analysis and Nonlinear Optimization, Theory and Examples, Second Edition. CMS Books in Maths, Springer-Verlag (2006).

Numerical example for the dual problem

In the case where $m = 2$ and $n = 3$ the objective function is given by

$$g^*(\mathbf{k}) = \frac{1}{3}\sum_{r=1}^{3}\sum_{s=1}^{3} \exp\left[3k_{rs}\right] - \log_e 3$$

and the constraints can be written in the form $A\mathbf{h} = \mathbf{b}$ where

$$A = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ \frac{1}{9} & \frac{1}{3} & \frac{5}{9} & \frac{1}{3} & 1 & \frac{5}{3} & \frac{5}{9} & \frac{5}{3} & \frac{25}{9} \end{bmatrix} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ \rho + 3 \end{bmatrix}.$$

Numerical example for the dual problem cont.

If we set $\rho = 0.7$ and let $\boldsymbol{\varphi}^{(0)} = \mathbf{0} \in \mathbb{R}^6$ then after 8 iterations the solution, shown to four decimal place accuracy, is given by

$$\overline{\boldsymbol{\varphi}} \approx \begin{bmatrix} -0.8682 \\ -2.4632 \\ -0.2825 \\ -1.1506 \\ -2.7457 \\ 1.8474 \end{bmatrix} \quad \text{and} \quad \overline{\mathbf{h}} \approx \begin{bmatrix} 0.7933 & 0.2010 & 0.0058 \\ 0.2010 & 0.5980 & 0.2010 \\ 0.0058 & 0.2010 & 0.7933 \end{bmatrix}$$

where we have written $\overline{\mathbf{h}} = [\overline{h}_{ij}]$ for convenience. The MATLAB calculations show that

$$\|A\overline{\mathbf{h}} - \mathbf{b}\| < 8 \times 10^{-15}$$

and the value of the objective function is given by $g^*(\overline{\mathbf{k}}) \approx -0.5761$. The *duality gap* satisfies the inequality

$$J(\overline{\mathbf{h}}) - g^*(\overline{\mathbf{k}}) < 3 \times 10^{-16}.$$

Case study

We have available 150 years of official monthly rainfall records supplied by the Australian Bureau of Meteorology for Sydney, in NSW, Australia, during the period 1859-2008.

Table : Monthly means (m) and standard deviations (s) for Sydney

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
m	103	118	130	126	103	131	98	82	70	77	84	78
s	76	110	103	112	111	116	82	84	60	66	76	63

Grade correlation coefficients for all monthly pairs

	Ja	Fe	Mr	Ap	Ma	Jn	Jl	Au	Se	Oc	No	De
Ja		**.18**	-.06	**-.19**	-.01	-.02	-.02	.13	.09	**-.16**	.05	-.04
Fe	**.18**		-.03	-.08	-.09	.05	-.01	.10	.09	-.05	.08	-.07
Mr	-.06	-.03		.11	.04	**.19**	-.14	-.15	-.12	.15	-.05	-.01
Ap	**-.19**	-.08	.11		**.18**	.05	.13	.12	-.08	.11	.09	-.03
Ma	-.01	-.09	.04	**.18**		.05	-.02	-.05	-.08	-.07	.05	-.06
Jn	-.02	.05	**.19**	.05	.05		-.04	-.07	**-.17**	.02	.05	-.05
Jl	-.02	-.01	-.14	.13	-.02	-.04		.11	.12	.08	-.08	-.02
Au	.13	.10	-.15	.12	-.05	-.07	.11		.13	.13	.12	-.09
Se	.09	.09	-.12	-.08	-.08	**-.17**	.12	.13		.04	.07	-.01
Oc	**-.16**	-.05	.15	.11	-.07	.02	.08	.13	.04		**.22**	-.03
No	.05	.08	-.05	.09	.05	.05	-.08	.12	.07	**.22**		.08
De	-.04	-.07	-.01	-.03	-.06	-.05	-.02	-.09	-.01	-.03	.08	

Example $m = 2$ and $n = 3$

Consider the months Oct-Nov (spring for us). If we set $\rho = 0.22$ (observed correlation) and let $\boldsymbol{\varphi}^{(0)} = \mathbf{0} \in \mathbb{R}^6$ then after 7 iterations the solution, shown to four decimal place accuracy, is given by

$$\overline{\boldsymbol{\varphi}} \approx \begin{bmatrix} -0.1746 \\ -0.3805 \\ -0.2920 \\ -0.4666 \\ -0.6726 \\ 0.2854 \end{bmatrix} \quad \text{and} \quad \overline{\mathbf{h}} \approx \begin{bmatrix} 0.4580 & 0.3281 & 0.2139 \\ 0.3281 & 0.3439 & 0.3281 \\ 0.2139 & 0.3281 & 0.4580 \end{bmatrix}$$

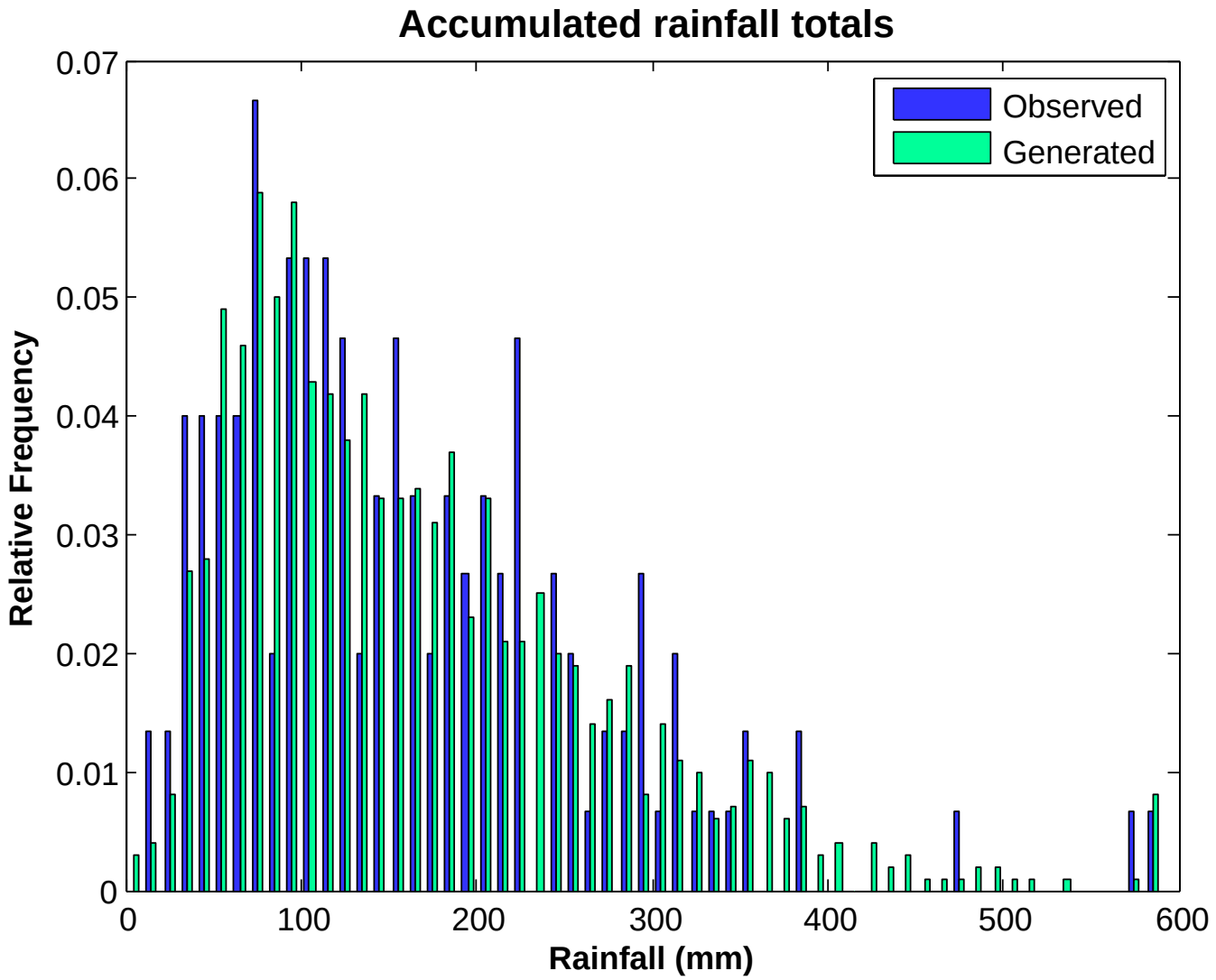
where we have written $\overline{\mathbf{h}} = [\overline{h}_{ij}]$ for convenience. The MATLAB calculations show that

$$\|A\overline{\mathbf{h}} - \mathbf{b}\| < 8 \times 10^{-15}$$

and the value of the objective is given by $g^*(\overline{\mathbf{k}}) \approx -0.9696$. The *duality gap* satisfies the inequality

$$J(\overline{\mathbf{h}}) - g^*(\overline{\mathbf{k}}) < 5 \times 10^{-16}.$$

Accumulated rainfall totals over months Oct-Nov



Results

Table : Comparison of mean and variance for the observed accumulated totals; generated accumulated totals using independent random variables (Independent Model) and generated accumulated totals using a copula of maximum entropy (Correlated Model)

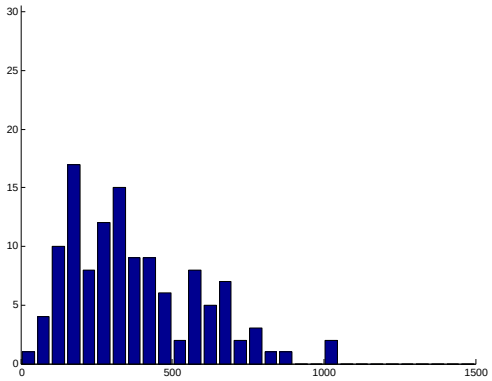
	Mean (mm)	**Variance**
Observed Data	160.488	10830.299
Independent Model	161.705	8732.117
Correlated Model (Max Ent)	160.451	10769.729

Table : P-values for Kolmogorov-Smirnov goodness of fit test

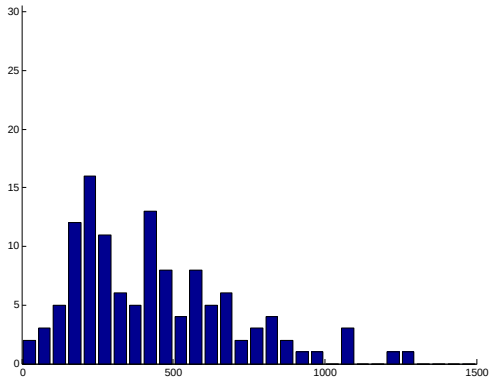
	Kolmogorov-Smirnov test
Observed versus generated	0.7637

Simulations: total rainfall (February through April) Kempsey in northern New South Wales

Observed statistics: $\mu = 427$, $\sigma^2 = 53762$ (μ, σ^2)

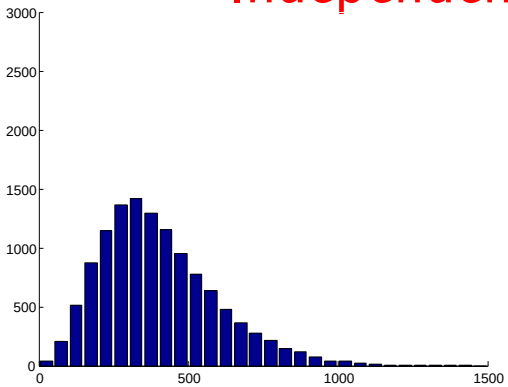


a) Time:122 years (401,45384)



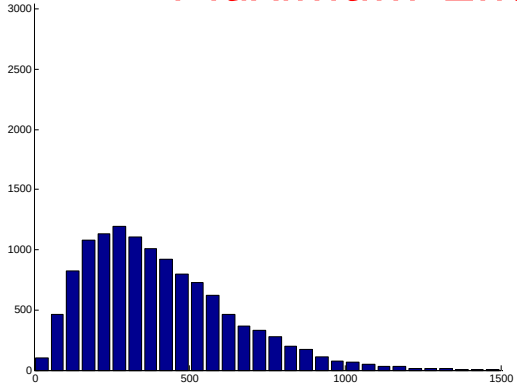
b) Time:122 years (457,69175)

Independence



c) Time:12200 years (427,38208)

Maximum Entropy



d) Time:12200 years (427,54448)

Comments

- By reformulating the problem as an *unconstrained concave optimization problem* using the theory of Fenchel duality we were able to show that solution of the dual problem and subsequent recovery of the primal solution is a much more tractable procedure.

Comments

- By reformulating the problem as an *unconstrained concave optimization problem* using the theory of Fenchel duality we were able to show that solution of the dual problem and subsequent recovery of the primal solution is a much more tractable procedure.

- The underlying entropy model assures one that the dual problem has many attractive features both theoretically and numerically.

Comments

- By reformulating the problem as an *unconstrained concave optimization problem* using the theory of Fenchel duality we were able to show that solution of the dual problem and subsequent recovery of the primal solution is a much more tractable procedure.

- The underlying entropy model assures one that the dual problem has many attractive features both theoretically and numerically.

- Lastly, our theoretical ideas can be applied to more realistic modelling and simulation of monthly rainfall (using Sydney and Kempsey data, see *J. Hydrology* paper) and comparing Gaussian copulas.

References

1. Borwein, J .M., Lewis, A. S., Convex Analysis and Nonlinear Optimization, Ed. 2. CMS Books, Springer-Verlag, 2006.
2. Joe, H., Multivariate models and dependence concepts, Monographs on Statistics and Applied Probability, **73** Chapman & Hall, 1997.
3. Nelsen, R.B., An introduction to copulas, Vol. 139 of Lecture Notes in Statistics, Springer, New York, 1999.
4. Piantadosi, J., Howlett, P.G. and Borwein, J. (2012) Copulas with maximum entropy. *Optimization Letters*, **6** (1), 99-125.
5. Piantadosi, J., Howlett, P.G., Borwein, J., Henstridge, J. (2012) Maximum entropy methods for generating simulated rainfall, *Numerical Algebra, Control and Optimization*, 2(2), 233-256.
6. Piantadosi, J., Howlett, P.G., Borwein, J., Henstridge, J. (2011) Generation of simulated rainfall data at different time-scales, `www.mssanz.org.au/modsim2011/D10/wongsosaputro.pdf`.
7. Piantadosi, J., Boland, J., Howlett, P.G. (2009) Simulation of rainfall totals on various time scales – daily, monthly and yearly. *Environmental Modeling and Assessment*, 14(4), 431–438.
8. Piantadosi J., Howlett P. and Borwein J., "Modelling and simulation of seasonal rainfall." In revision *Journal of Hydrology*, Feb 2013.