Bayesian Multivariate Distributional Regression with Skewed Responses and Skewed Random Effects
Supplement

Patrick Michaelis*
Chair of Statistics, University of Goettingen, Germany
and
Nadja Klein
Melbourne Business School, University of Melbourne, Australia
and
Thomas Kneib
Chair of Statistics, University of Goettingen, Germany

July 13, 2017

*Financial support by the German Research Foundation (DFG), grant KN 922/4-2 is gratefully acknowledged.

1 Simulations

1.1 Univariate Simulations

Skew-Normal Distribution In the following, Figures for the simulation studys presented in the article are given. These are

- Histograms of the estimated effective number of parameters (Skew-normal, direct parameterization: 1, 2 and 3; Skew-normal, centered parameterization:5, 6 and 7)
- Coverage rates for all considered combinations (Skew-normal, direct parameterization: 4; Skew-normal, centered parameterization: 8)

The second aspect of the simulations was to analyze whether the connections between the parameters creates problems for the estimation. To do so, we look at the coverage rates of the nonlinear effects. Overall the average coverage rates are above the considered level of 0.95. The setup leads to lower and less stable coverage rates for the effects of x_2. With more observations, the coverage rates get more stable and higher overall, if they were low at certain points (typically extrema of the effects).

The coverage rates are a little more stable for the centered parameterization than for the direct parameterization when looking at the first two parameters. For the skewness parameter, the coverage rates are worse for small n but improve when n is larger.

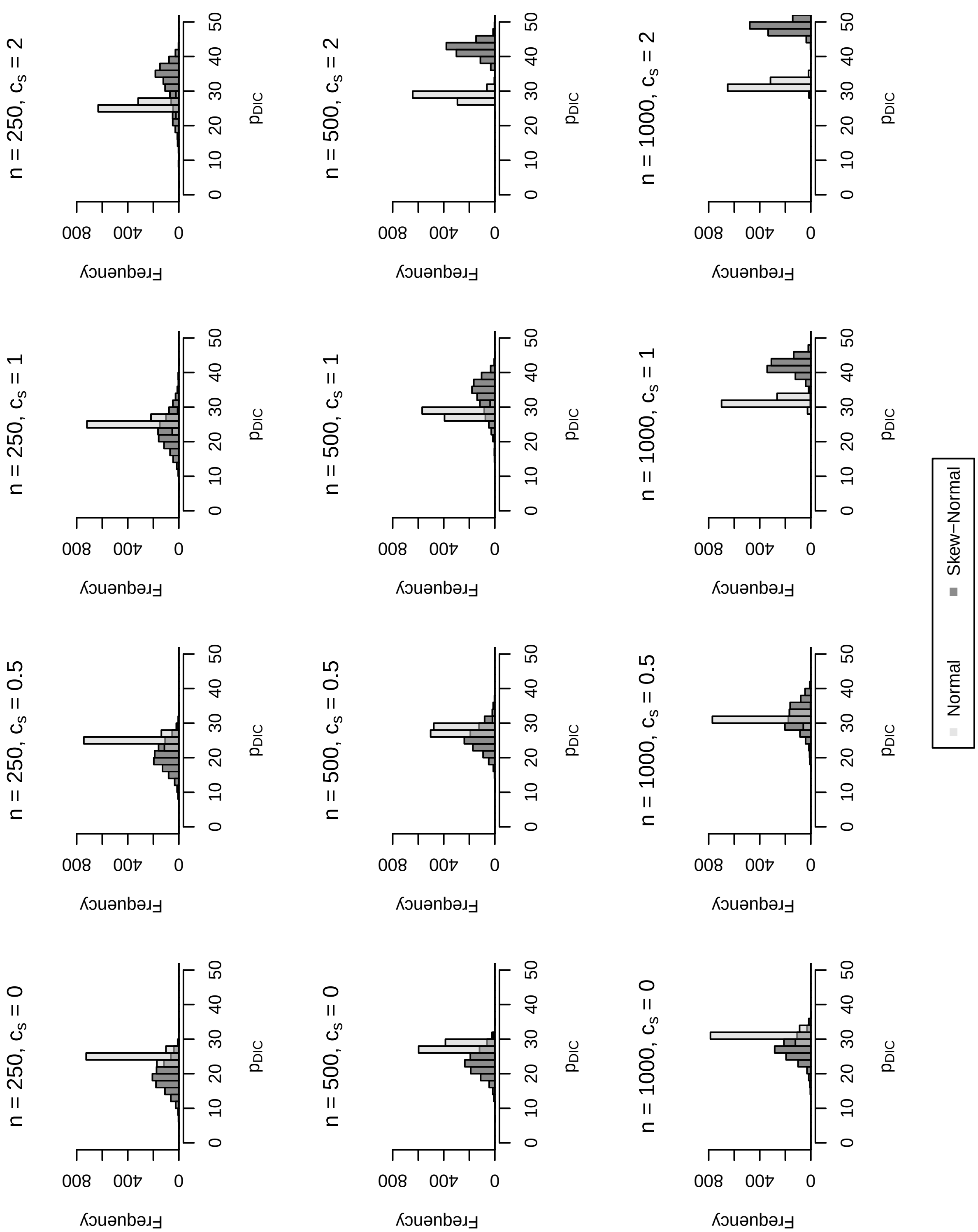

Figure 1: Histograms of the estimated effective number of parameters of the DIC for both the skew-normal and the normal distribution using the direct parameterization.

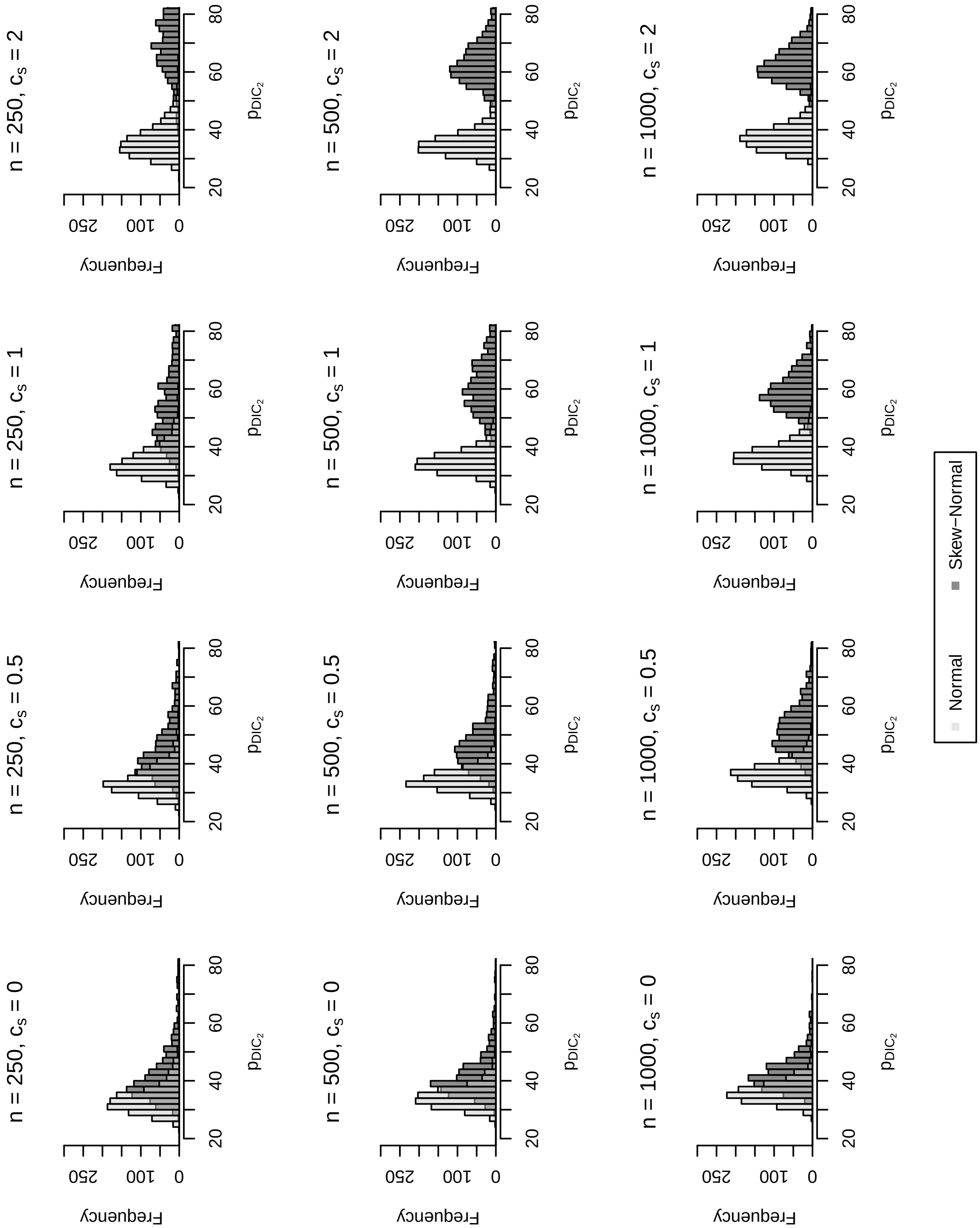

Figure 2: Histograms of the estimated effective number of parameters of the DIC_2 for both the skew-normal and the normal distribution using the direct parameterization.

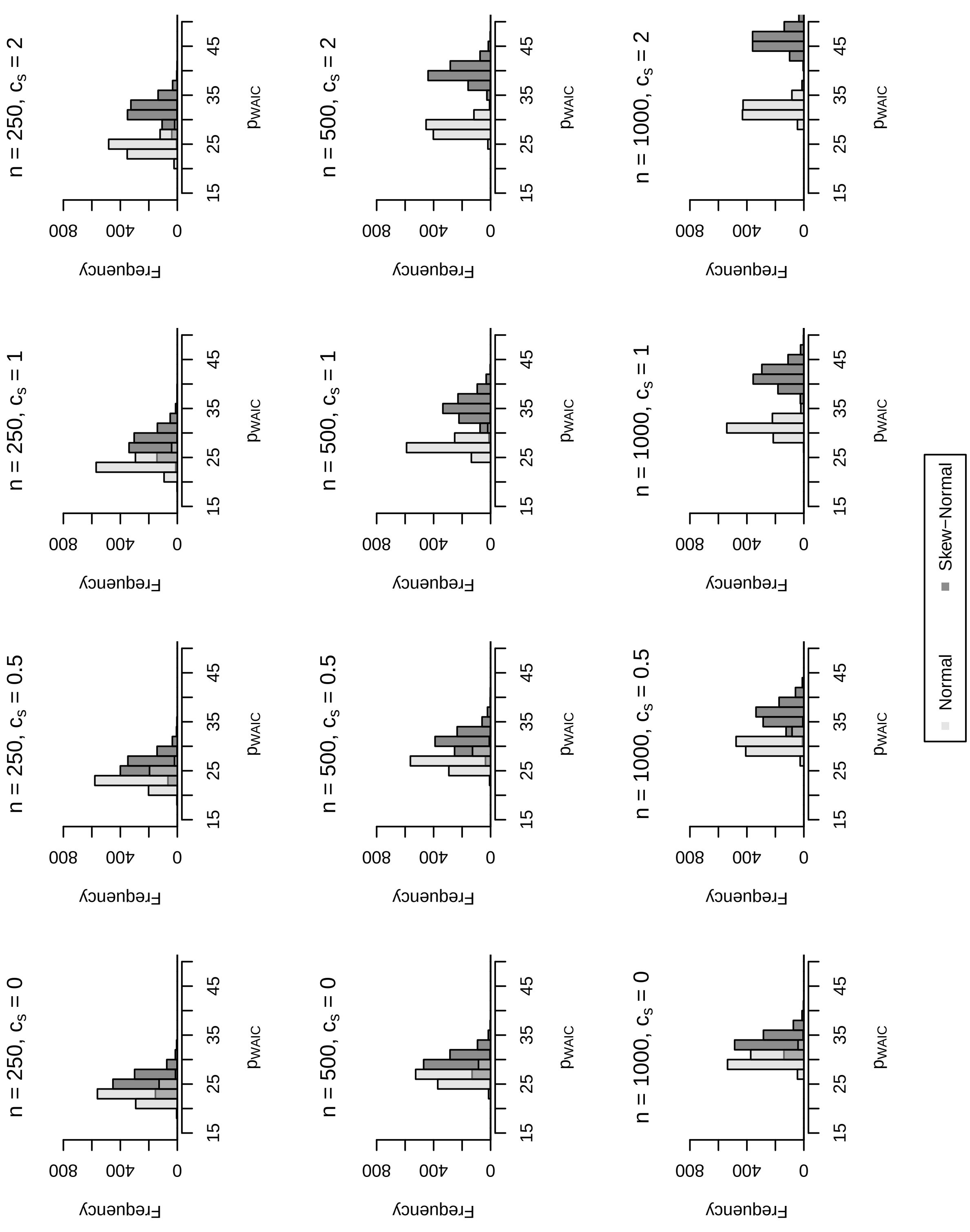

Figure 3: Histograms of the estimated effective number of parameters of the WAIC for both the skew-normal and the normal distribution using the direct parameterization.

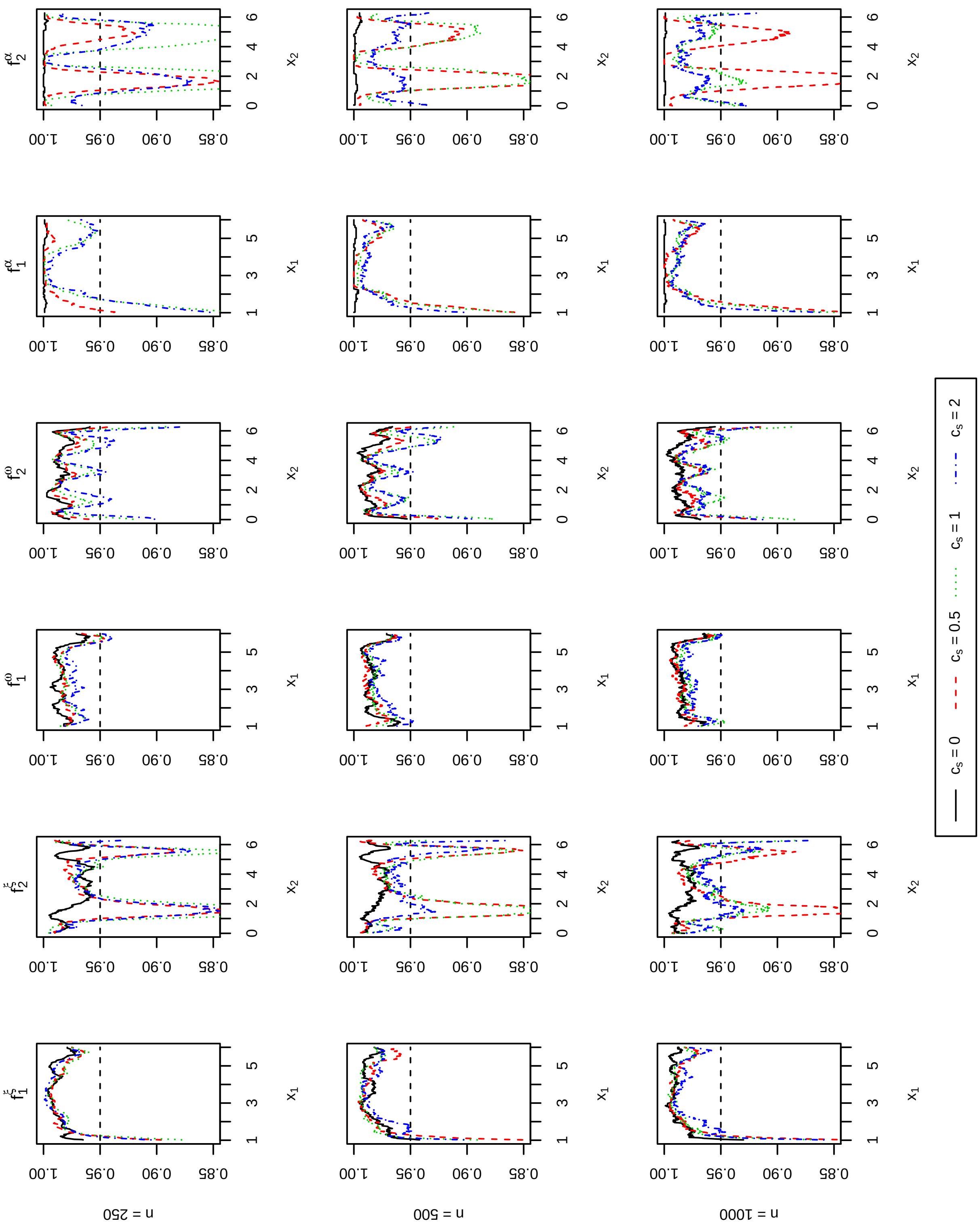

Figure 4: Coverage rates of the nonlinear effects for different values of c_s.

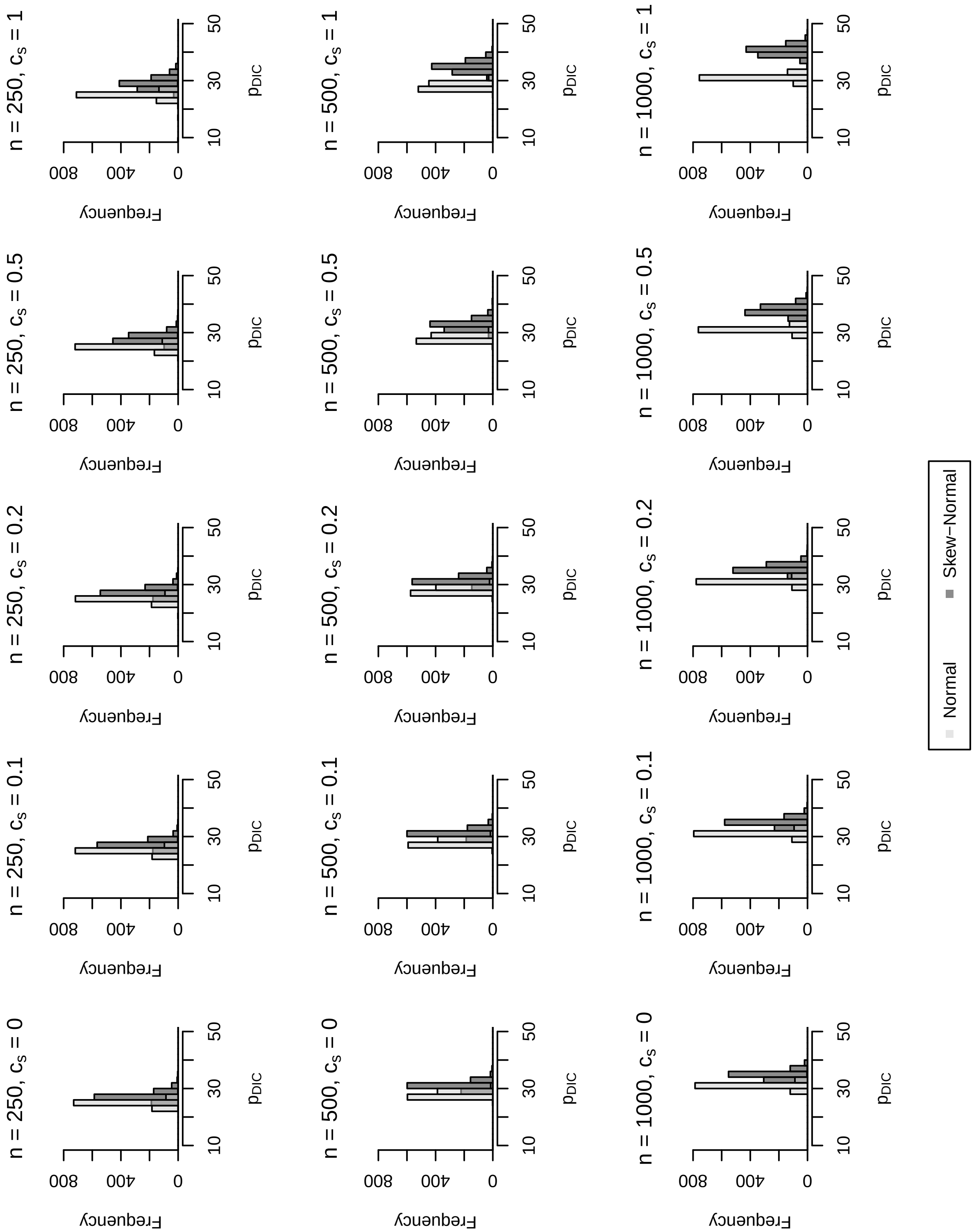

Figure 5: Histograms of the estimated effective number of parameters of the DIC for both the skew-normal and the normal distribution using the centered parameterization.

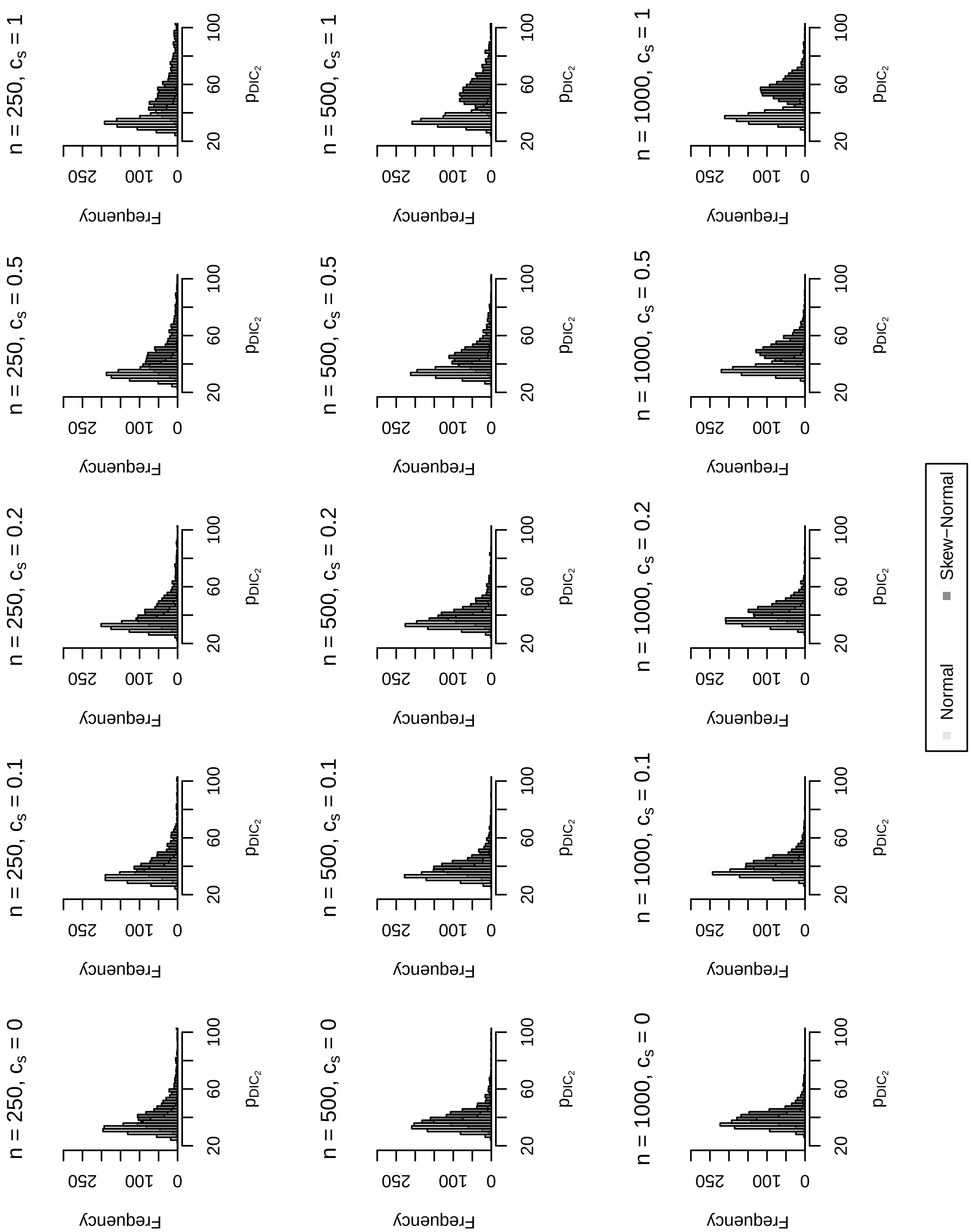

Figure 6: Histograms of the estimated effective number of parameters of the DIC_2 for both the skew-normal and the normal distribution using the centered parameterization.

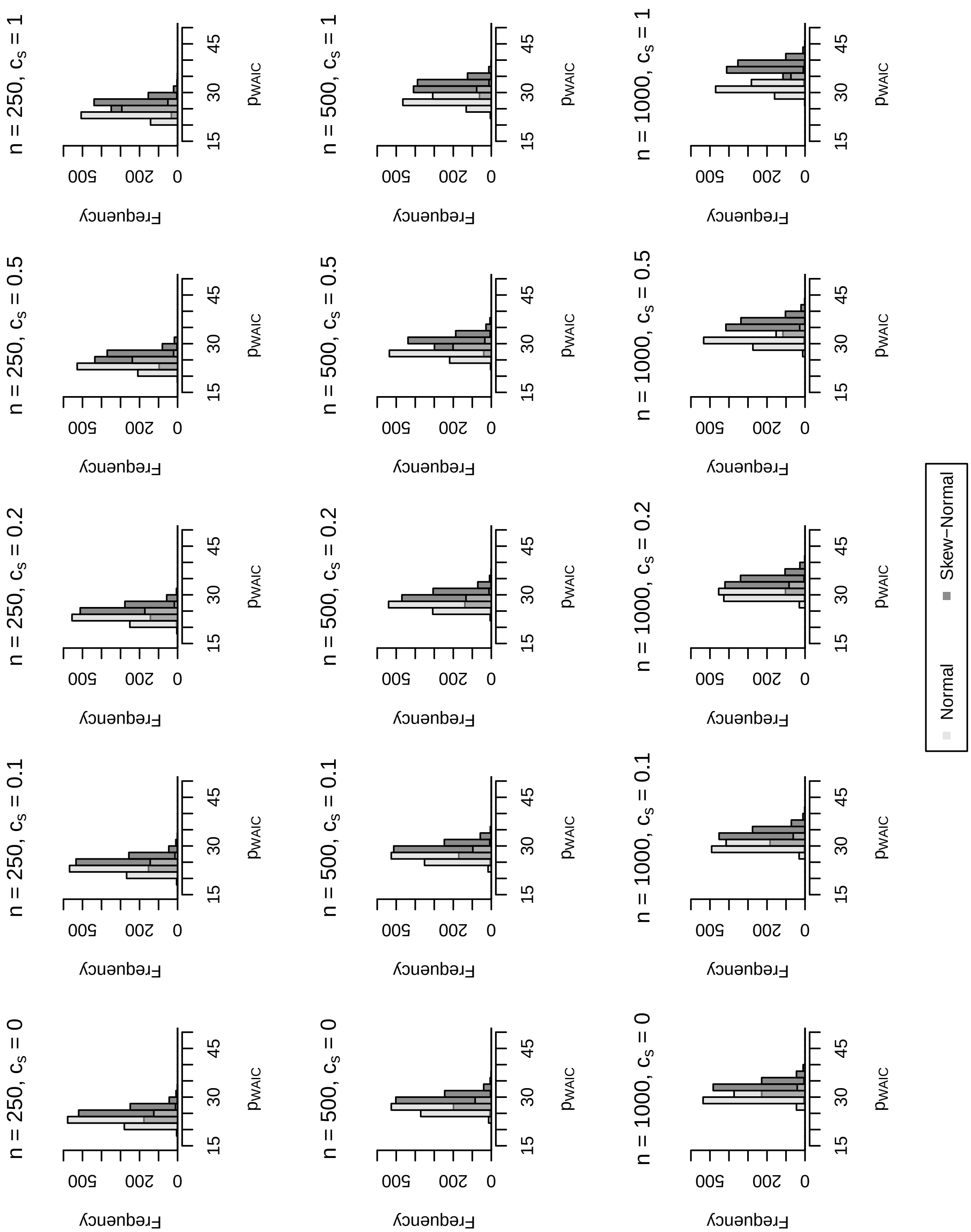

Figure 7: Histograms of the estimated effective number of parameters of the WAIC for both the skew-normal and the normal distribution using the centered parameterization.

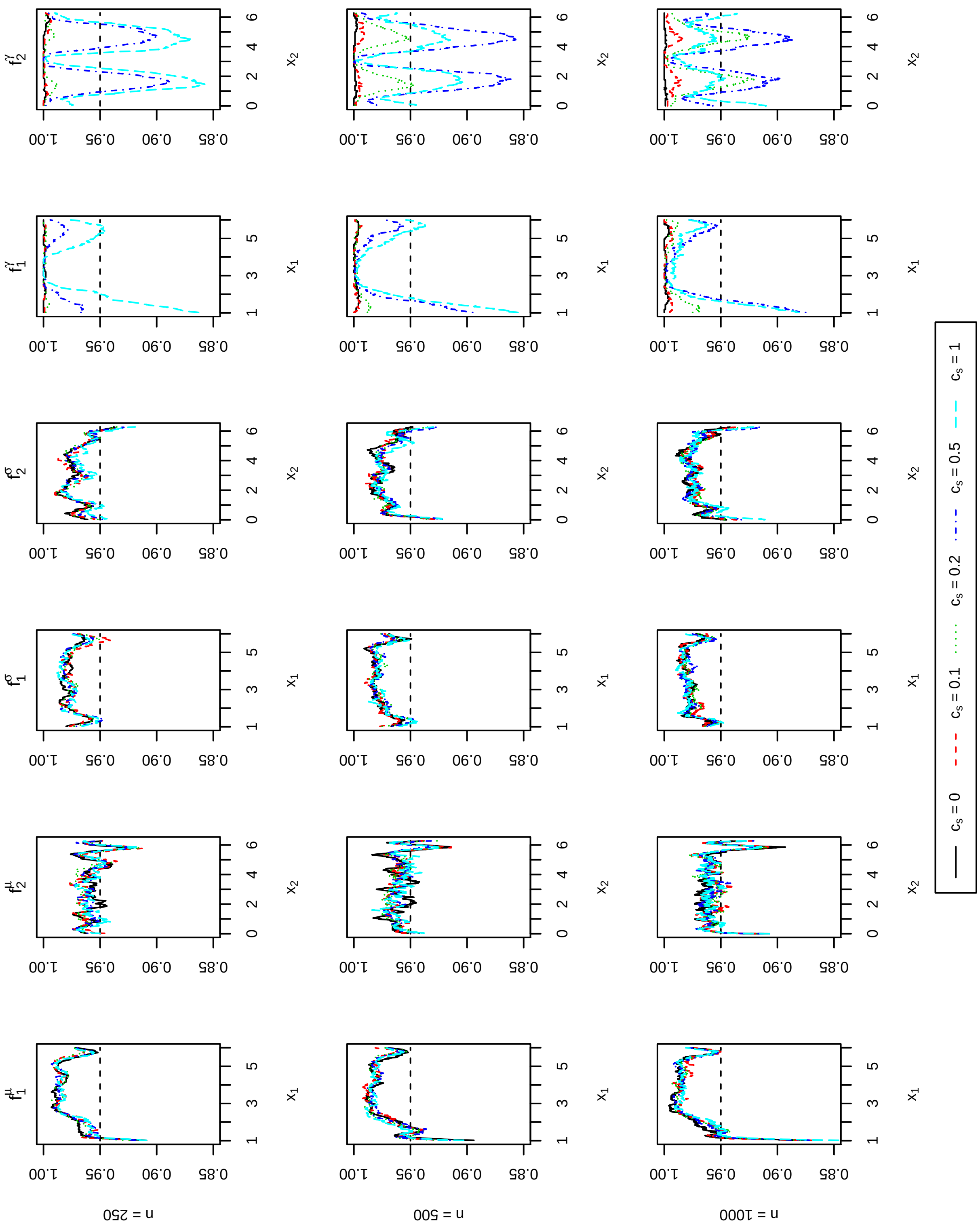

Figure 8: Coverage rates of the nonlinear effects for different values of c_s.

Skew-t Distribution The simulations for the skew-t distribution use the following nonlinear effects:

$$f_1^\xi = x_1^{-1}, \qquad f_2^\xi = \cos(2x_2)\sqrt{x_2},$$
$$f_1^\omega = 2\arctan(x_1), \qquad f_2^\omega = \cos(x_2),$$
$$f_1^\alpha = c_s \log(x_1), \qquad f_2^\alpha = c_s \sin(x_2),$$
$$f_1^\nu = -0.1x_1 + 0.025x_1^2, \qquad f_2^\nu = 0.4\cos(0.5x_2)$$

where $c_s = 0, 0.5, 1$ and 2. We used 250 repetitions and $n = 250, 500$ and 1000 observations. The results of the comparison of the t and the skew-t distribution are given in Table 1. Histograms of the estimated number of parameters are given in Figure 9 for the DIC, Figure 10 for the DCI_2 and Figure 11 for the WAIC. It is noteworthy that, even though the direct parameterization is used, the DIC performs reasonably well to decide whether the t or the skew-t distribution is more suitable if the number of observations is large. The DIC_2 tends towards the skew-t models, as before due to the high p_{DIC_2}. The WAIC is nevertheless more likely to choose the true model. Overall, as for the skew-normal distribution, the estimation is problematic for both DICs, especially for smaller numbers of observations.

Considering the coverage rates given in Figures 12 and 13, it is interesting to note, that the nonlinear effects of the degrees of freedom parameter are relatively stable, possibly due to the small effect. Otherwise, the same observations as for the skew-normal distributions can be made.

Skew-t Distribution

		DIC			DIC_2			WAIC		
	n	250	500	1000	250	500	1000	250	500	1000
c_s	0	13.2	39.2	63.9	90	90	72.7	72.8	82	83.5
	0.5	4	3.2	0	88.8	93.2	75.6	42.4	22	2.4
	1	0	0	0	92.3	79.6	8.8	6.9	0	0
	2	0	0	0	85.5	10.4	0	0	0	0

Table 1: Simulation results for comparison of the information criteria. The values given in the table are the share of the simulations for which the t distribution is chosen.

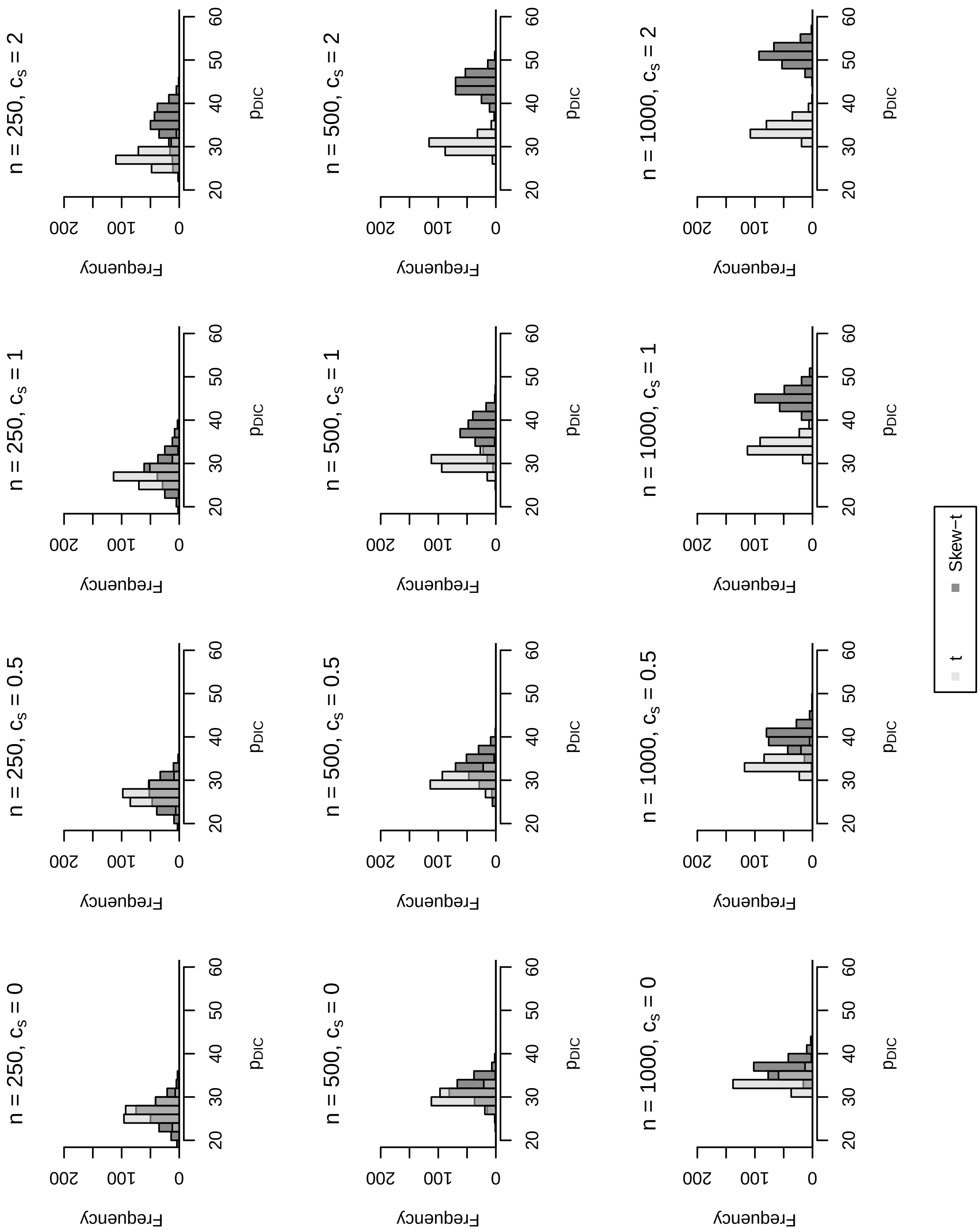

Figure 9: Histograms of the estimated effective number of parameters of the DIC for both the skew-t and the t distribution.

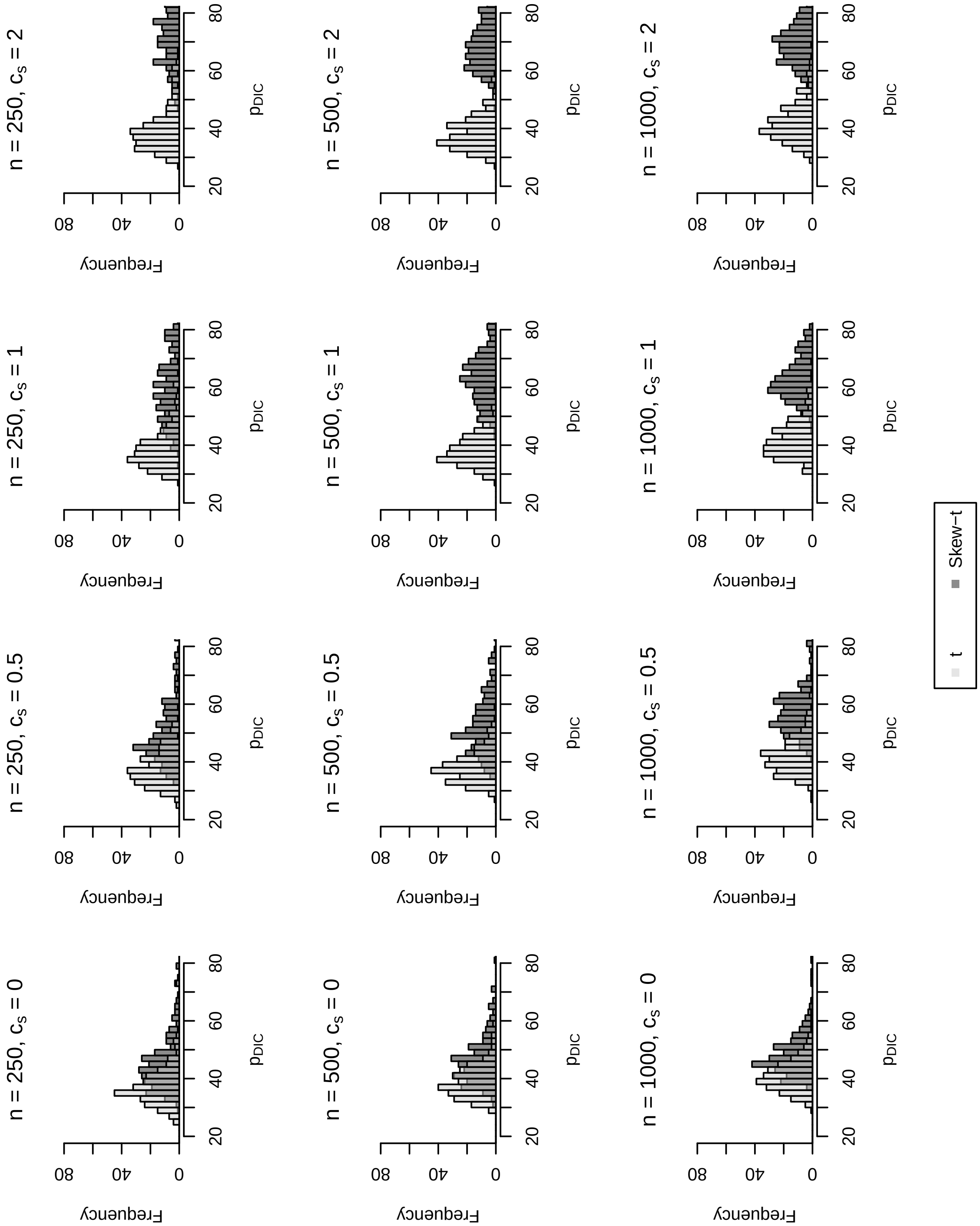

Figure 10: Histograms of the estimated effective number of parameters of the DIC_2 for both the skew-t and the t distribution.

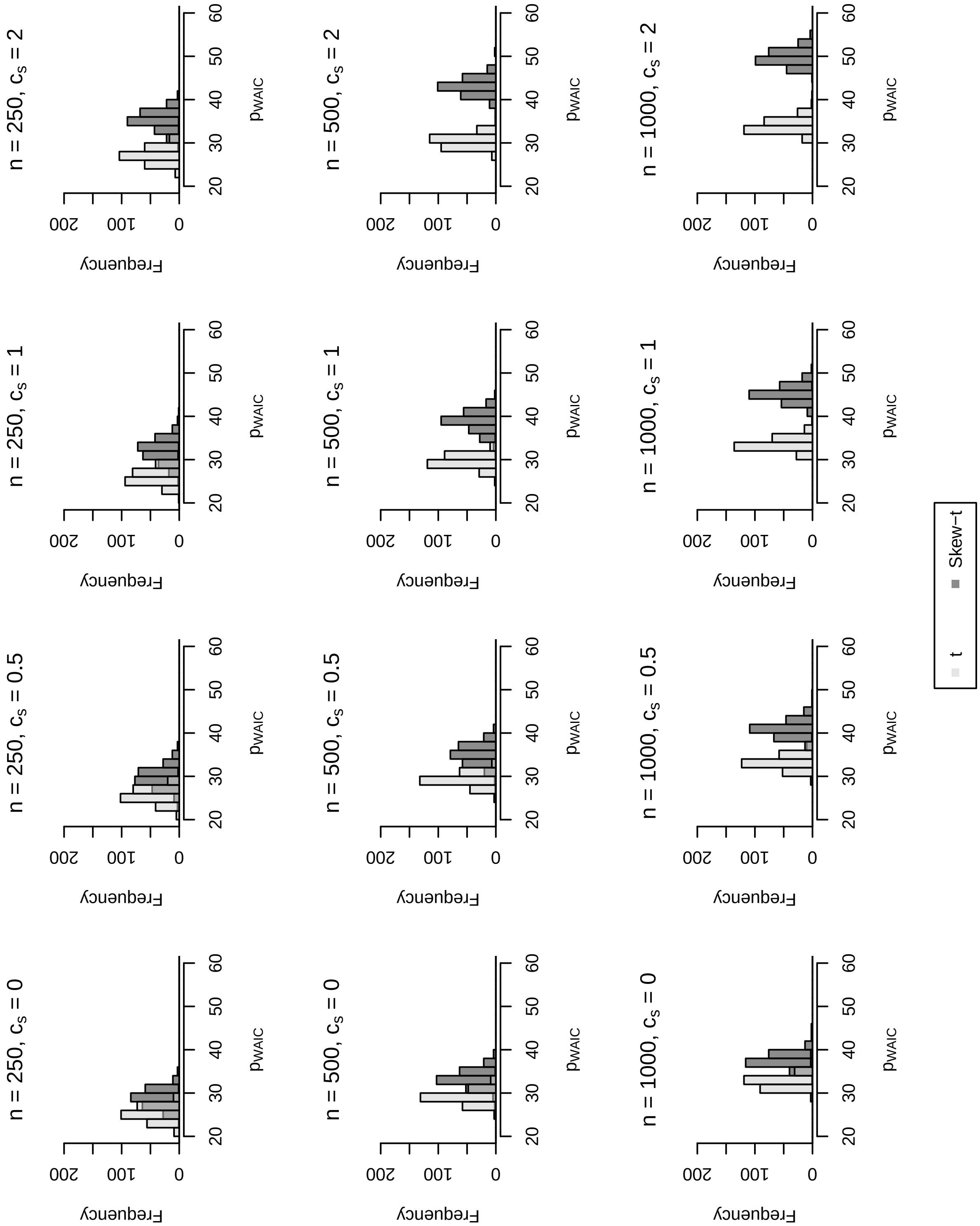

Figure 11: Histograms of the estimated effective number of parameters of the WAIC for both the skew-t and the t distribution.

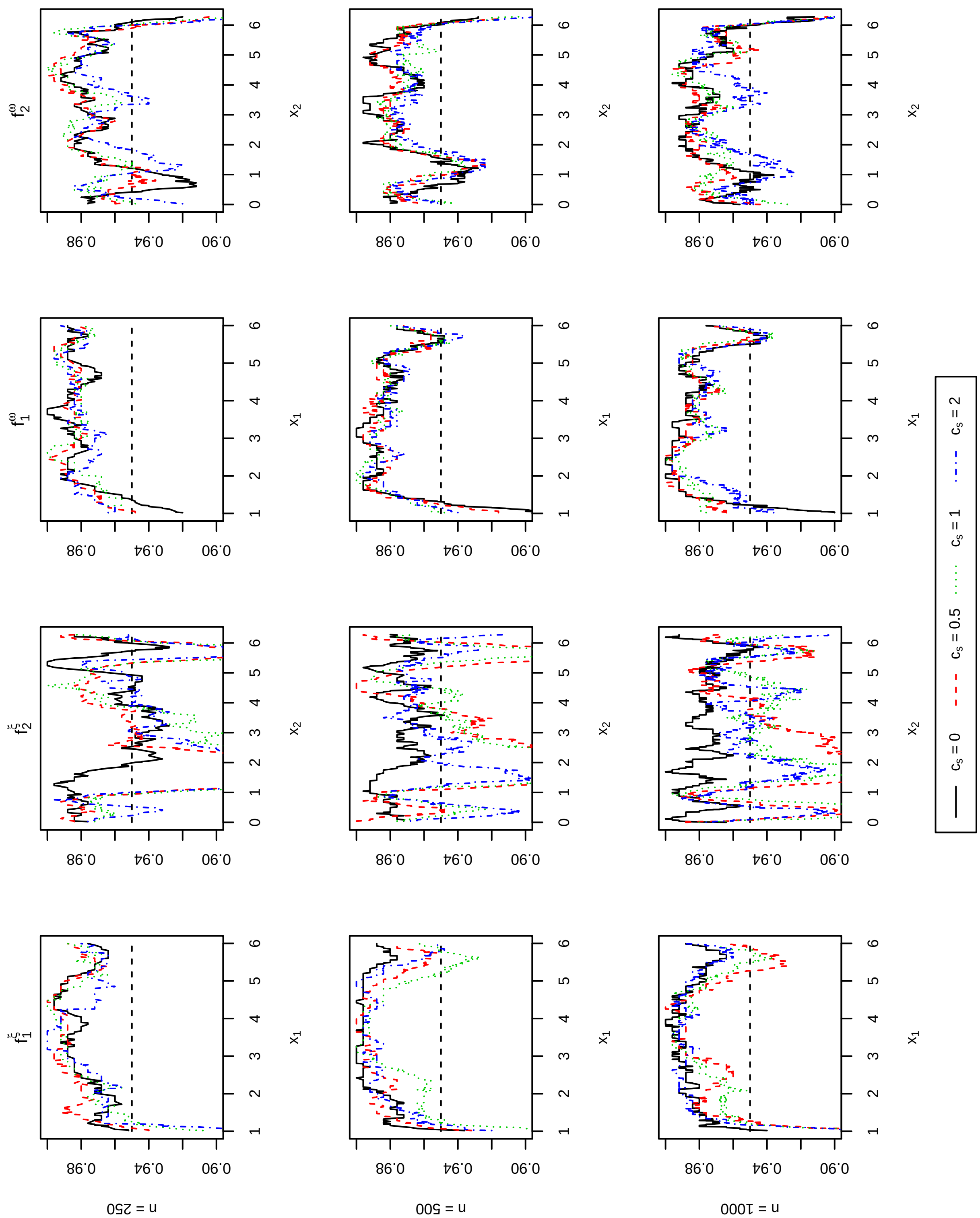

Figure 12: Coverage rates of the nonlinear effects for different values of c_s.

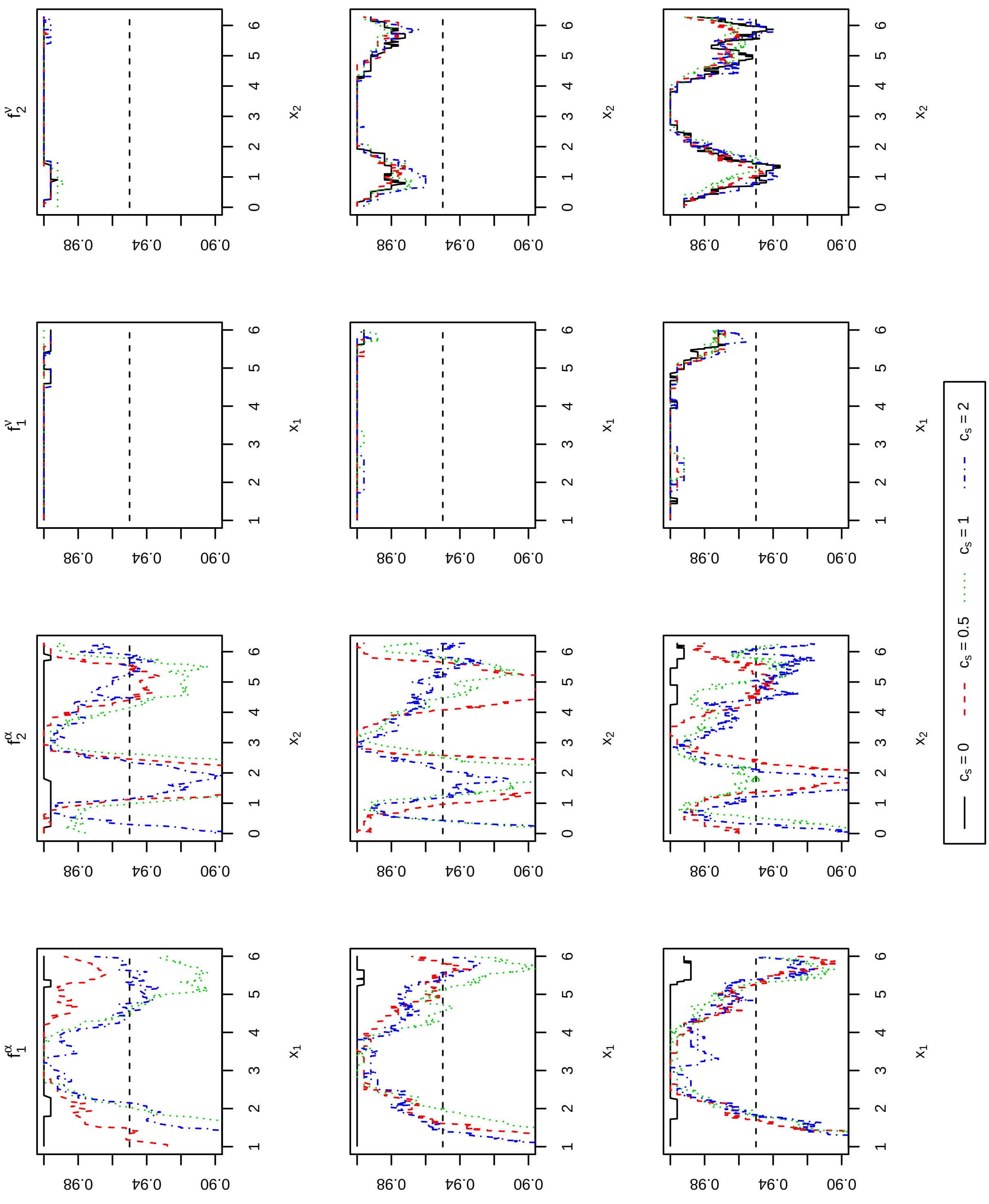

Figure 13: Coverage rates of the nonlinear effects for different values of c_s.

Model with Spatial Effect The model with a spatial effect is based on the model for the direct paramtereization of the skew-normal distribution. The nonlinear effects for the covariate x_1 stay the same as in the models used in the main article. The second covariate is now a spatial variable with effects (for $c_s = 1$) given in Figure 14. The values use for c_s are 0, 0.5, 1 and 2.

		DIC			DIC$_2$			WAIC		
	n	250	500	1000	250	500	1000	250	500	1000
c_s	0	0.6	1.2	1.8	67.5	67.4	71.6	46.4	51.3	63.5
	0.5	1.9	2.7	3	72.7	75	74.2	60	43.2	28.9
	1	1.4	0.6	0	77.3	87.6	96.3	52.9	20.4	3.4
	2	0.3	0	0	89.7	99.1	75.8	38.2	2.7	0

Table 2: Simulation results for the spatial model. The values given in the table are the percentages of the simulations for which the normal distribution is chosen.

In Table 2 the results of the comparison of the information criteria are given. The results are not very different from those for the nonlinear models. It is noteworthy that the DIC$_2$ increases for higher values of c_s. This might be a consequence of the higher difficulty to estimate the values in this model, especially for the parameter α. The coverage rates highlight this as well (Figures 18, 19, 20 and 21).

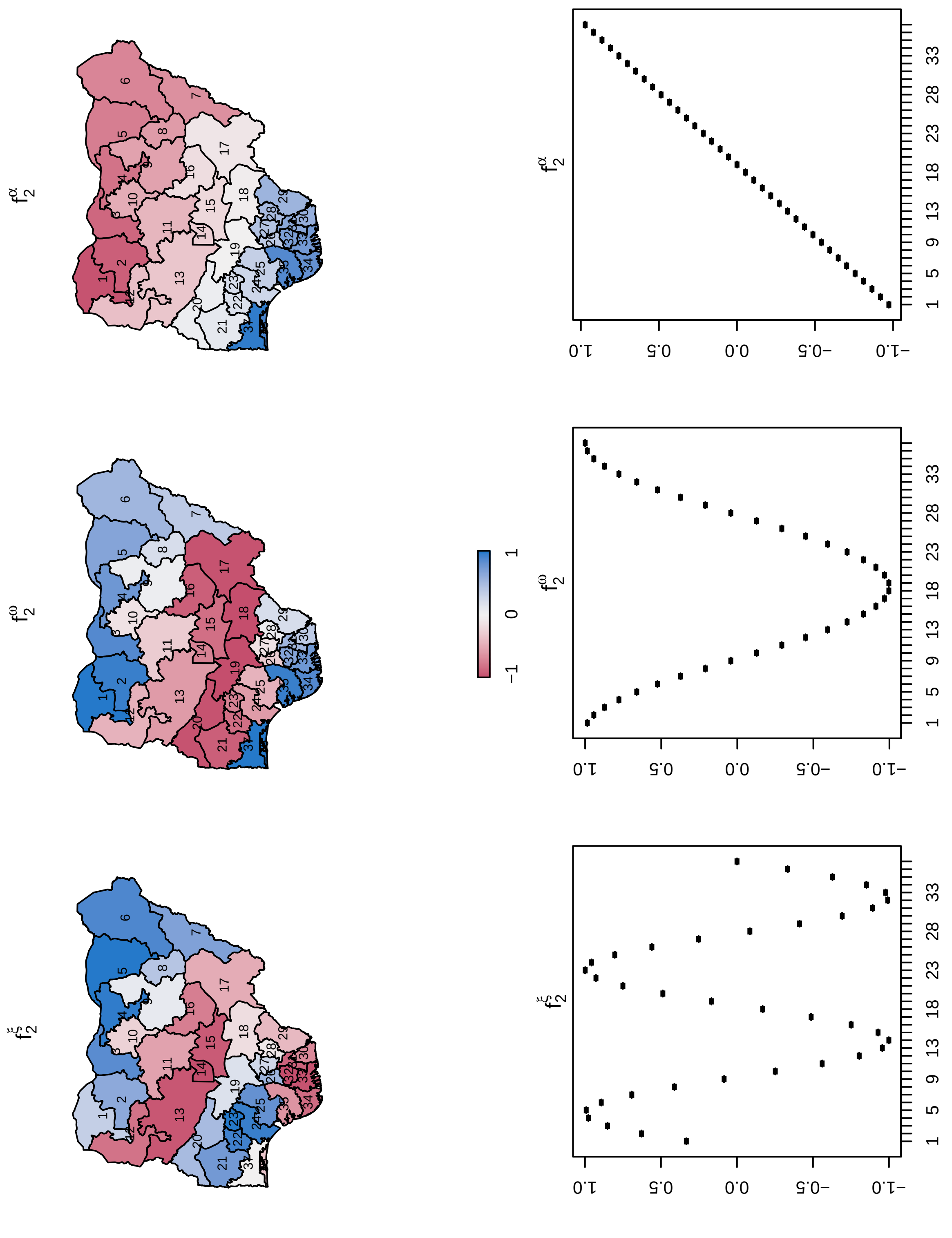

Figure 14: Visualisations of the true effect on x_2 for the model with a spatial effect.

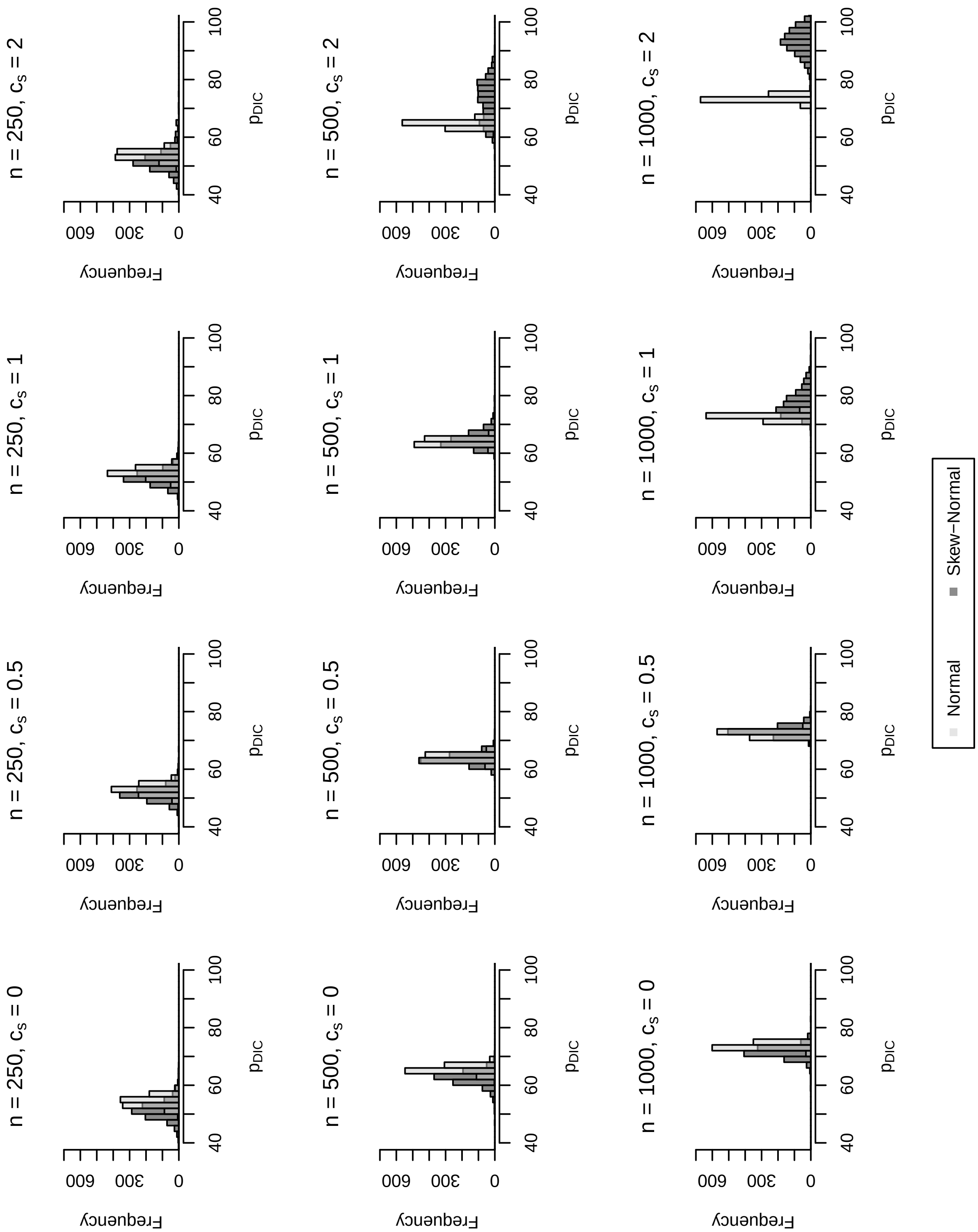

Figure 15: Histograms of the estimated effective number of parameters of the DIC for both the skew-normal and the normal distribution using the direct parameterization.

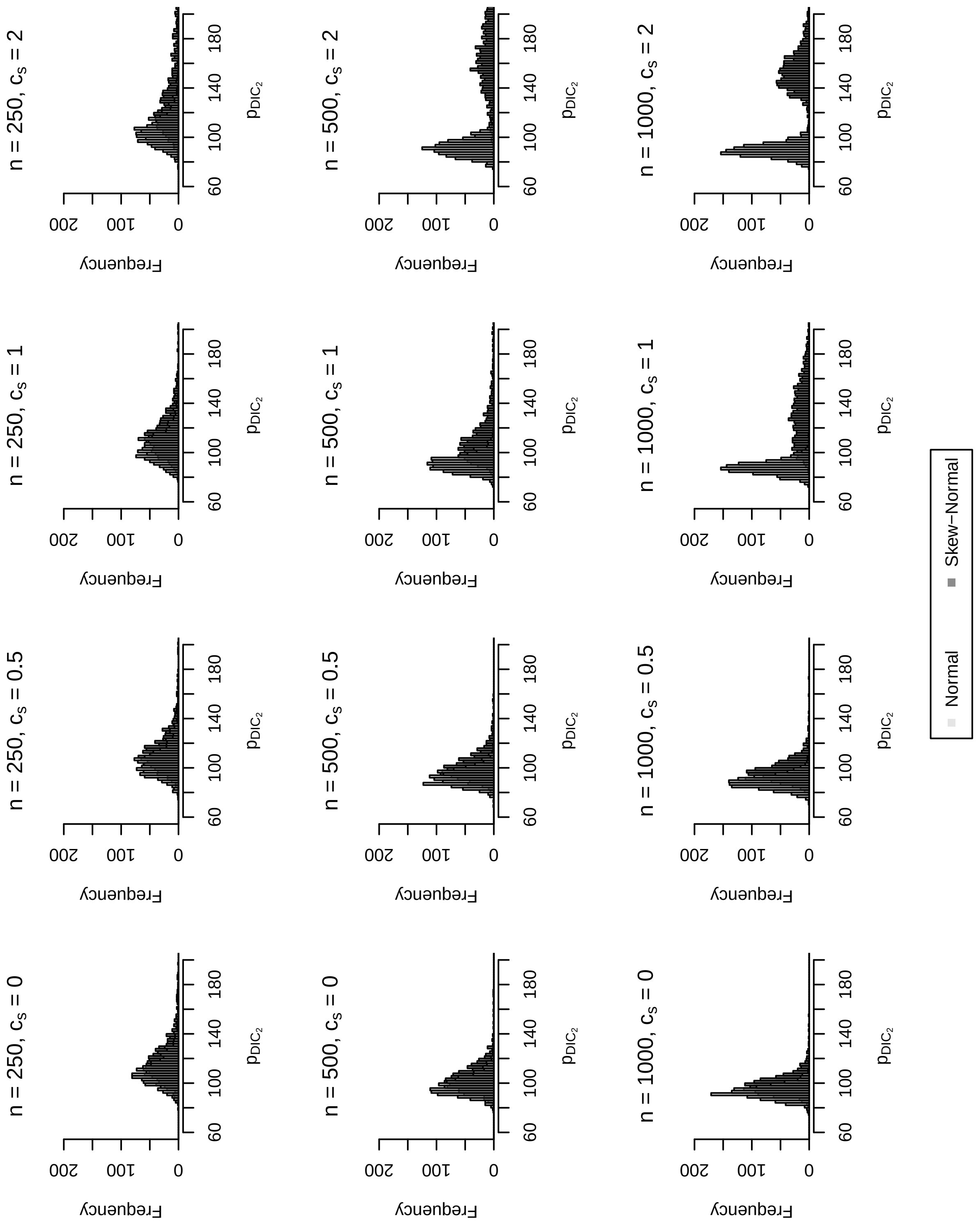

Figure 16: Histograms of the estimated effective number of parameters of the DIC_2 for both the skew-normal and the normal distribution using the direct parameterization.

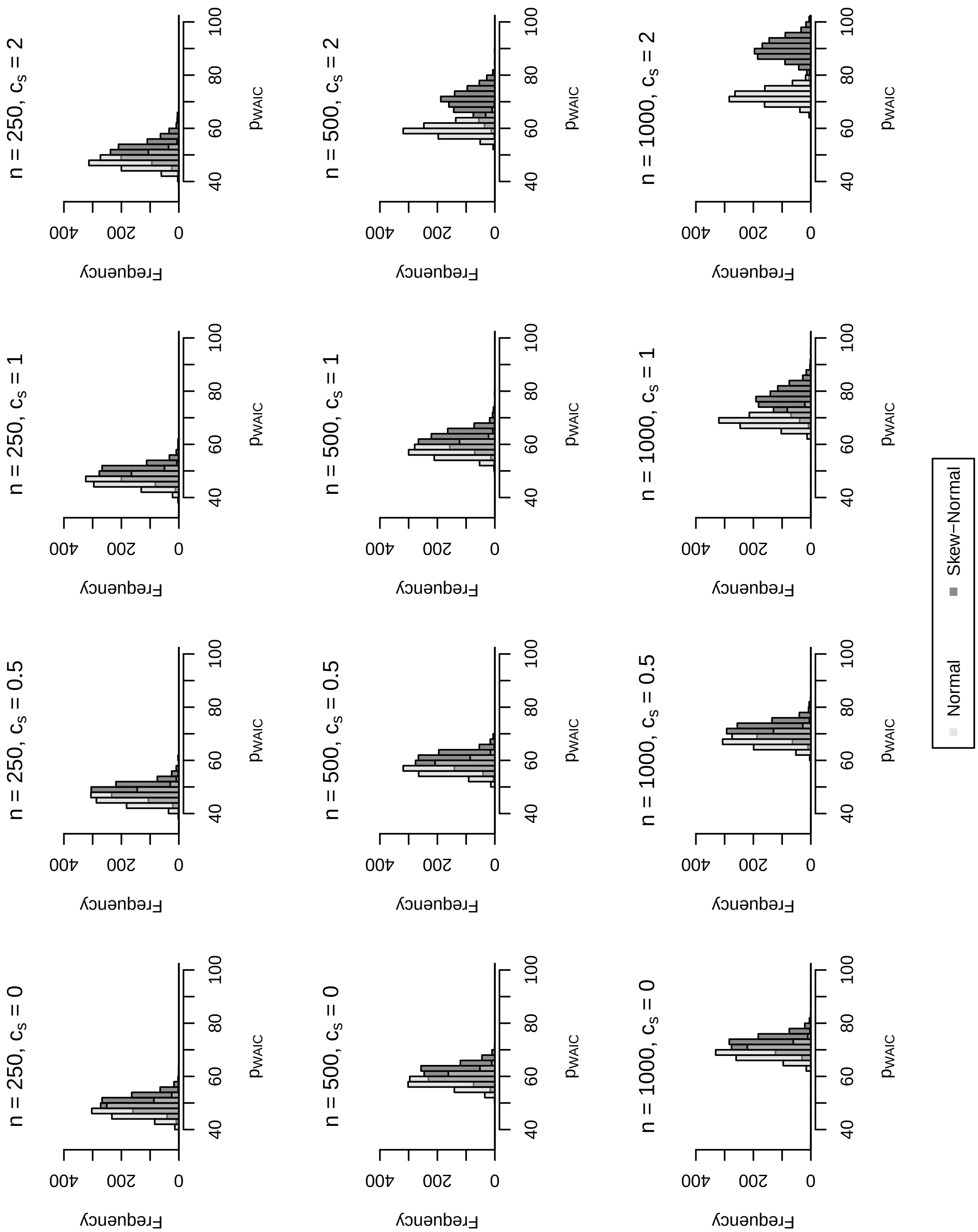

Figure 17: Histograms of the estimated effective number of parameters of the WAIC for both the skew-normal and the normal distribution using the direct parameterization.

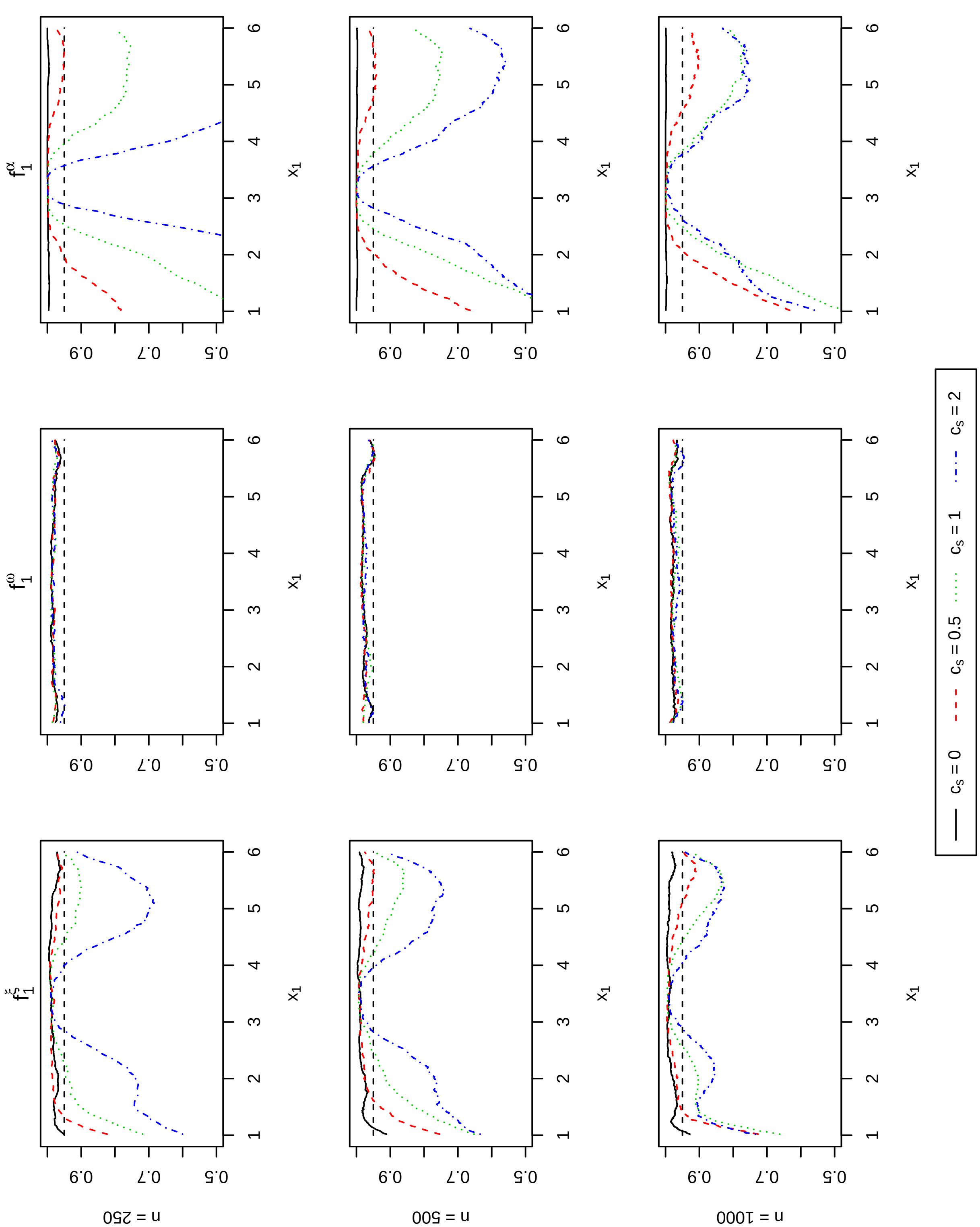

Figure 18: Coverage rates of the nonlinear effects for different values of c_s.

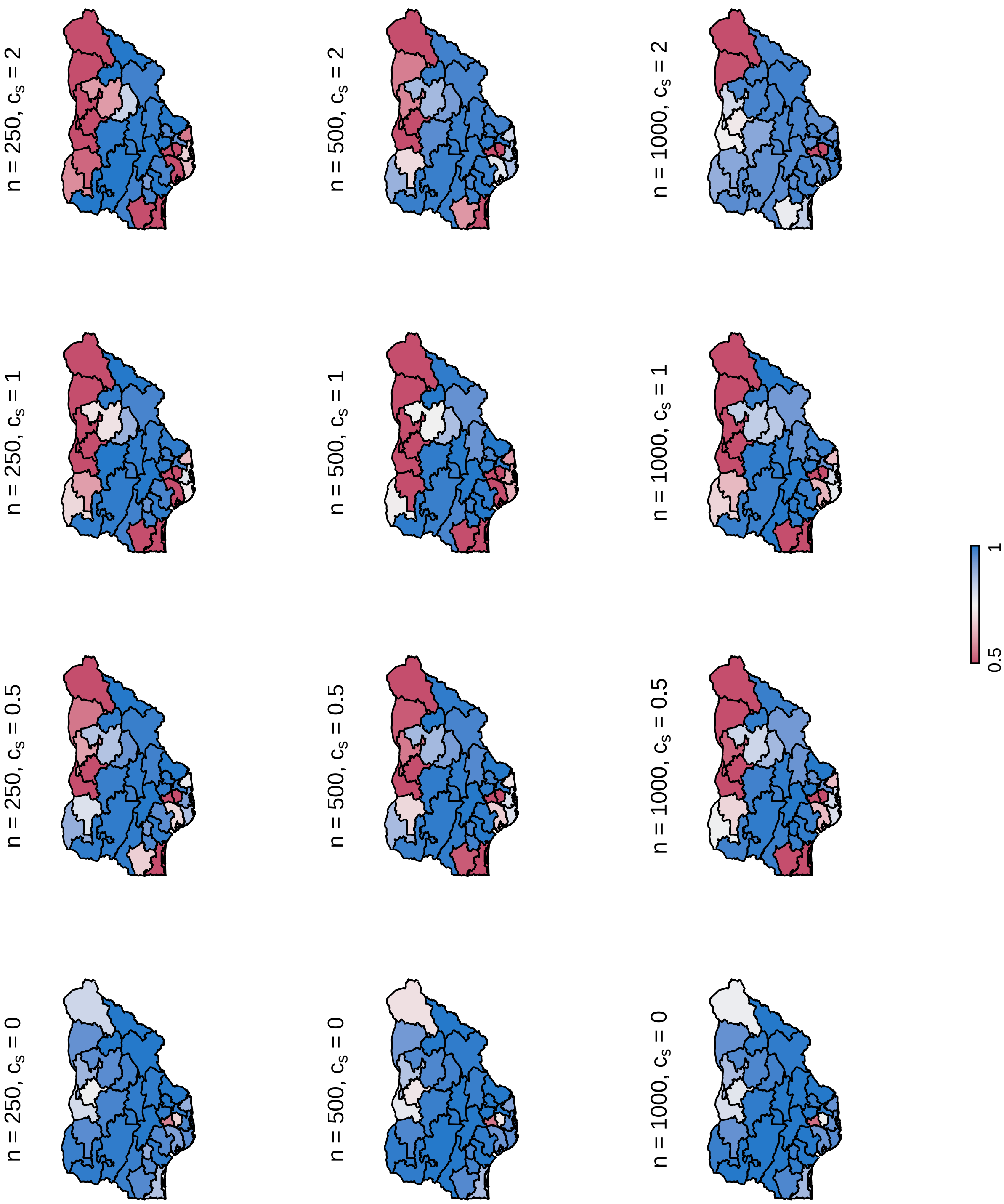

Figure 19: Coverage rates of the spatial effects for different values of c_s.

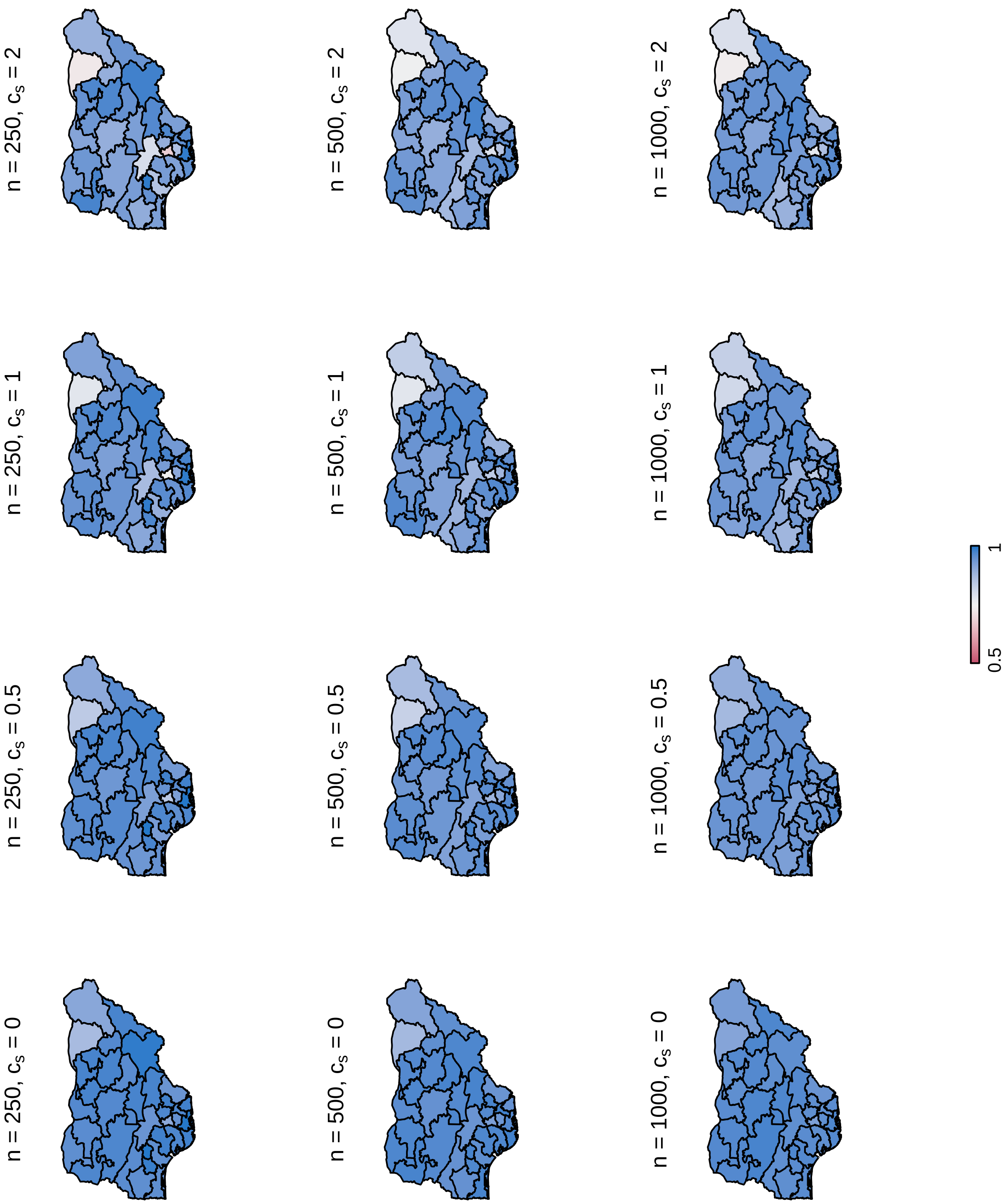

Figure 20: Coverage rates of the spatial effects for different values of c_s.

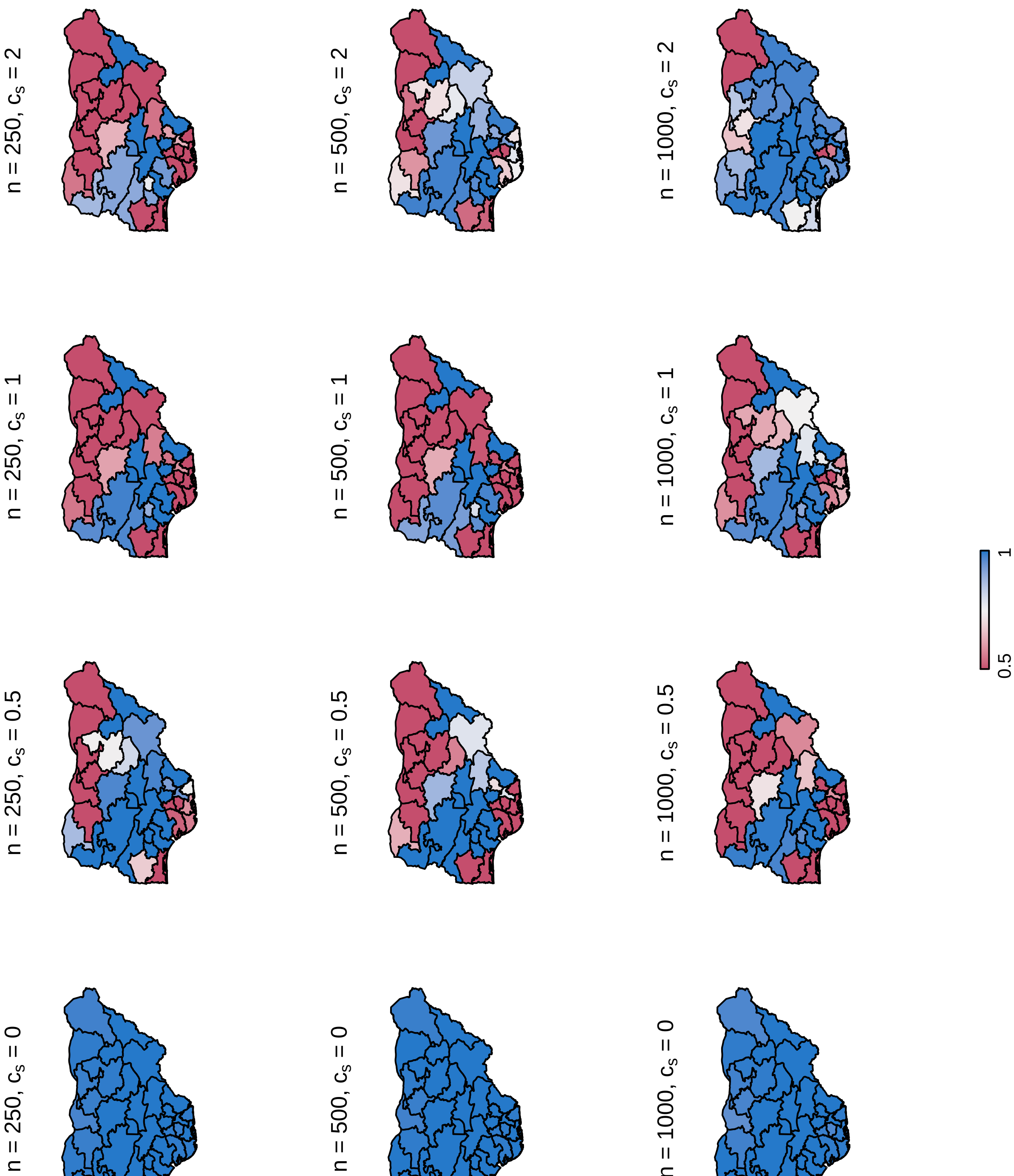

Figure 21: Coverage rates of the spatial effects for different values of c_s.

1.2 Bivariate Simulations

Skew-Normal Distribution The predictors for the parameters of the bivariate skew-normal distribution are given by

$$
\begin{aligned}
f_1^{\xi_1} &= x_1^{-1}, & f_2^{\xi_1} &= \cos(2x_2)\sqrt{x_2},\\
f_1^{\xi_2} &= 2x_1 - 3\log(x_1)^2, & f_2^{\xi_2} &= 0.1\sin(x_2)\exp(0.5x_2),\\
f_1^{\omega_1} &= 0.7x_1 - 0.1x_1^2, & f_2^{\omega_2} &= \cos(x_2),\\
f_1^{\omega_2} &= 0.025x_1^3 - 0.15x_1^2, & f_2^{\omega_2} &= 0.1\cos(x_2)x_2,\\
f_1^{\rho} &= 0.6\log(x_1), & f_2^{\rho} &= 0.5\cos(x_2),\\
f_1^{\alpha_1} &= c_s\log(x_1 - 0.1x_1^2), & f_2^{\alpha_1} &= c_s 0.5\sin(x_2),\\
f_1^{\alpha_2} &= c_s 1.8x_1^{-0.5}, & f_2^{\alpha_2} &= c_s\tanh\left(\frac{x_2}{2\pi}\right).
\end{aligned}
$$

As for the univariate case, we used $c_s = 0, 0.5, 1, 2$, but due to the high number of paramters we omit simulations for $n = 250$ observations and only use $n = 500$ and $n = 1000$. The results for the comparison of the information criteria are given in Table 3. While the DIC favours the skew-normal distribution independend of the choice of c_s, the DIC_2 favours the normal distribution as long as c_s and n are small. The $WAIC$ is best suited to choose the adequate distribution. For $n = 1000$ the results improve considerably. As before this is due to the estimation of the number of effective parameters (c.f. Figures 22, 23 and 24).

In Figures 25, 26 and 27 the coverage rates for the nonlinear effects are given. Overall the coverage rates are above the desired level. In some cases the coverage rates fall on the upper and lower ends of the range of the covariates. This could be due to the smoothing of the effects.

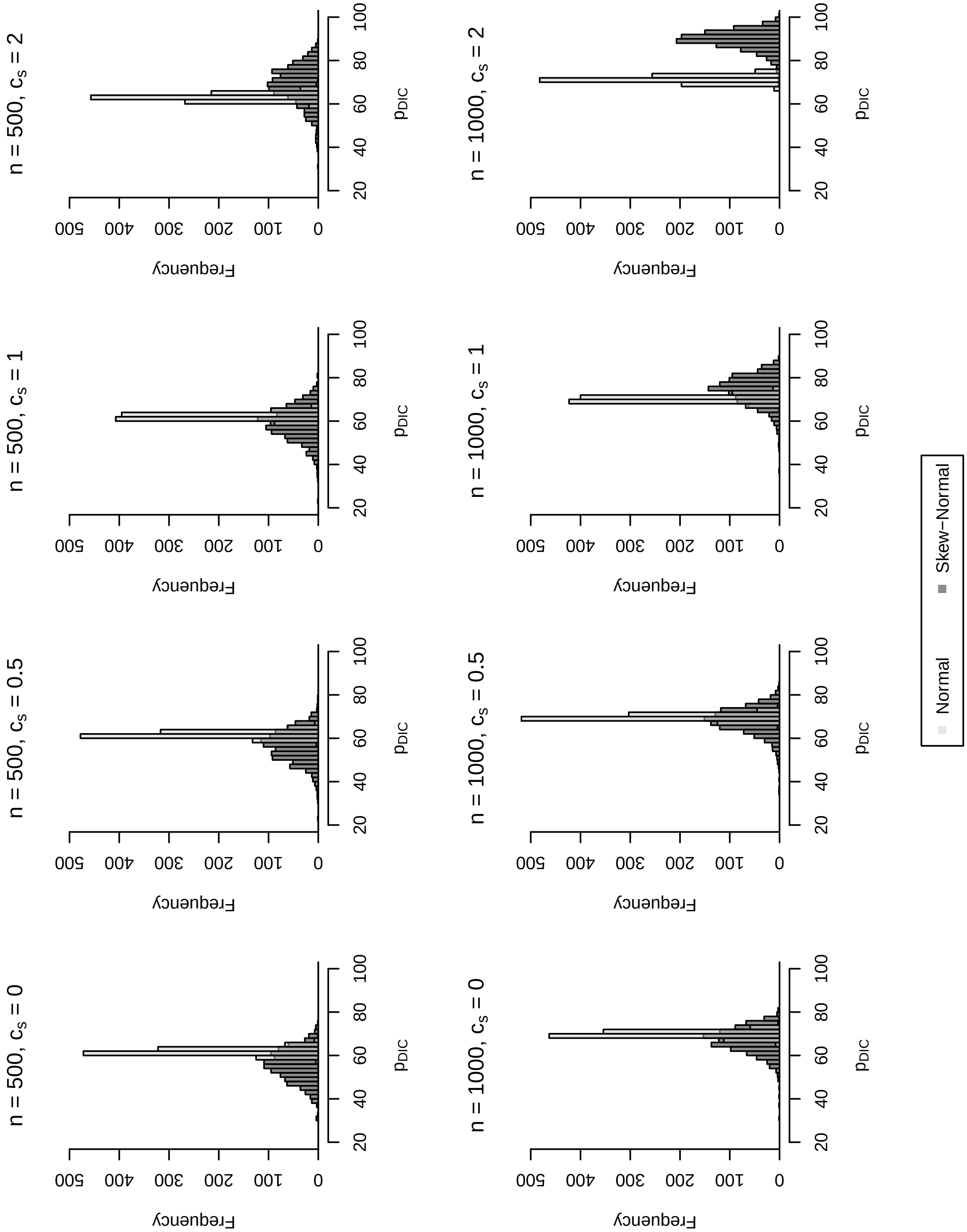

Figure 22: Histograms of the estimated effective number of parameters of the DIC for both the bivariate skew-normal and the bivariate normal distribution.

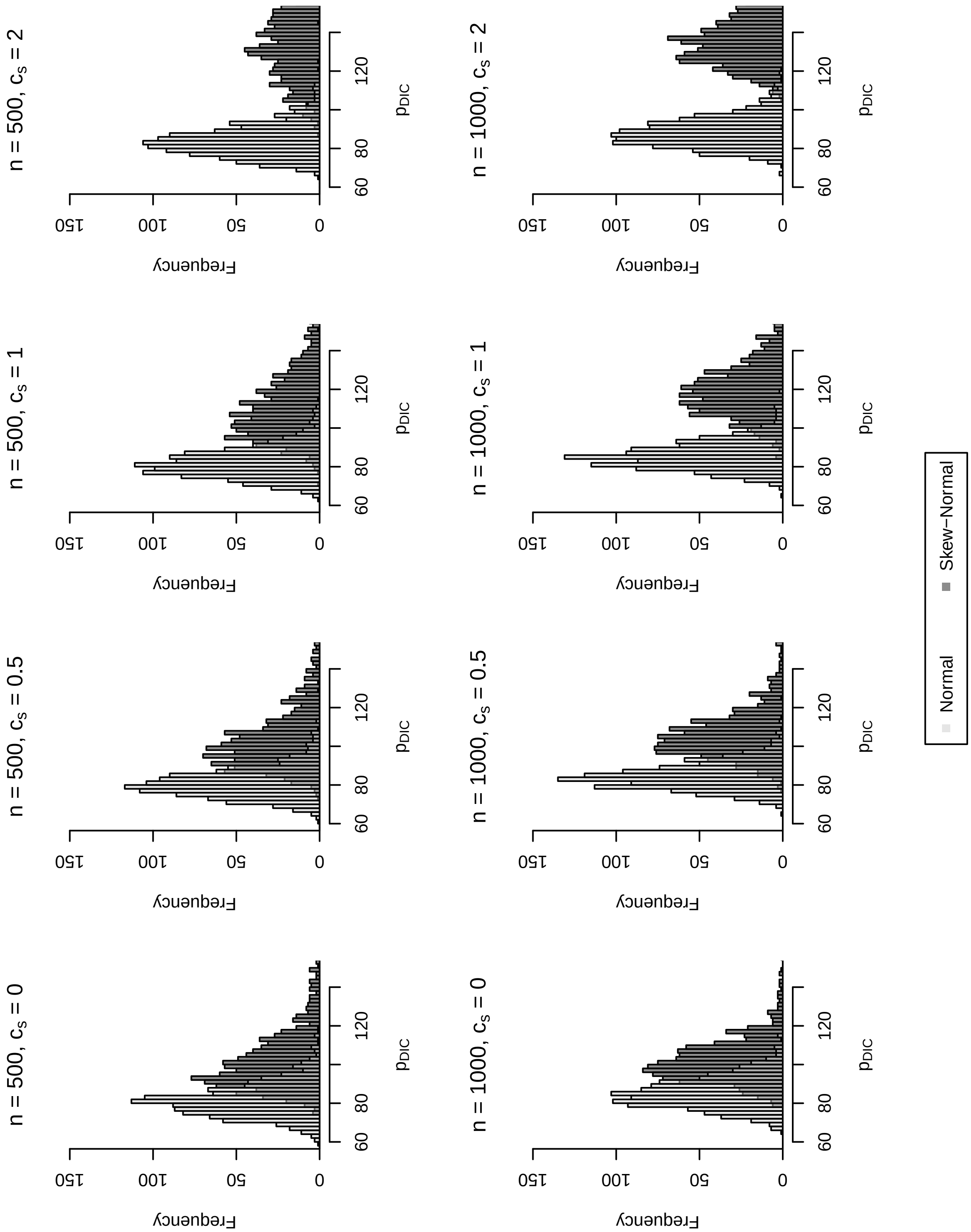

Figure 23: Histograms of the estimated effective number of parameters of the DIC_2 for both the bivariate skew-normal and the bivariate normal distribution.

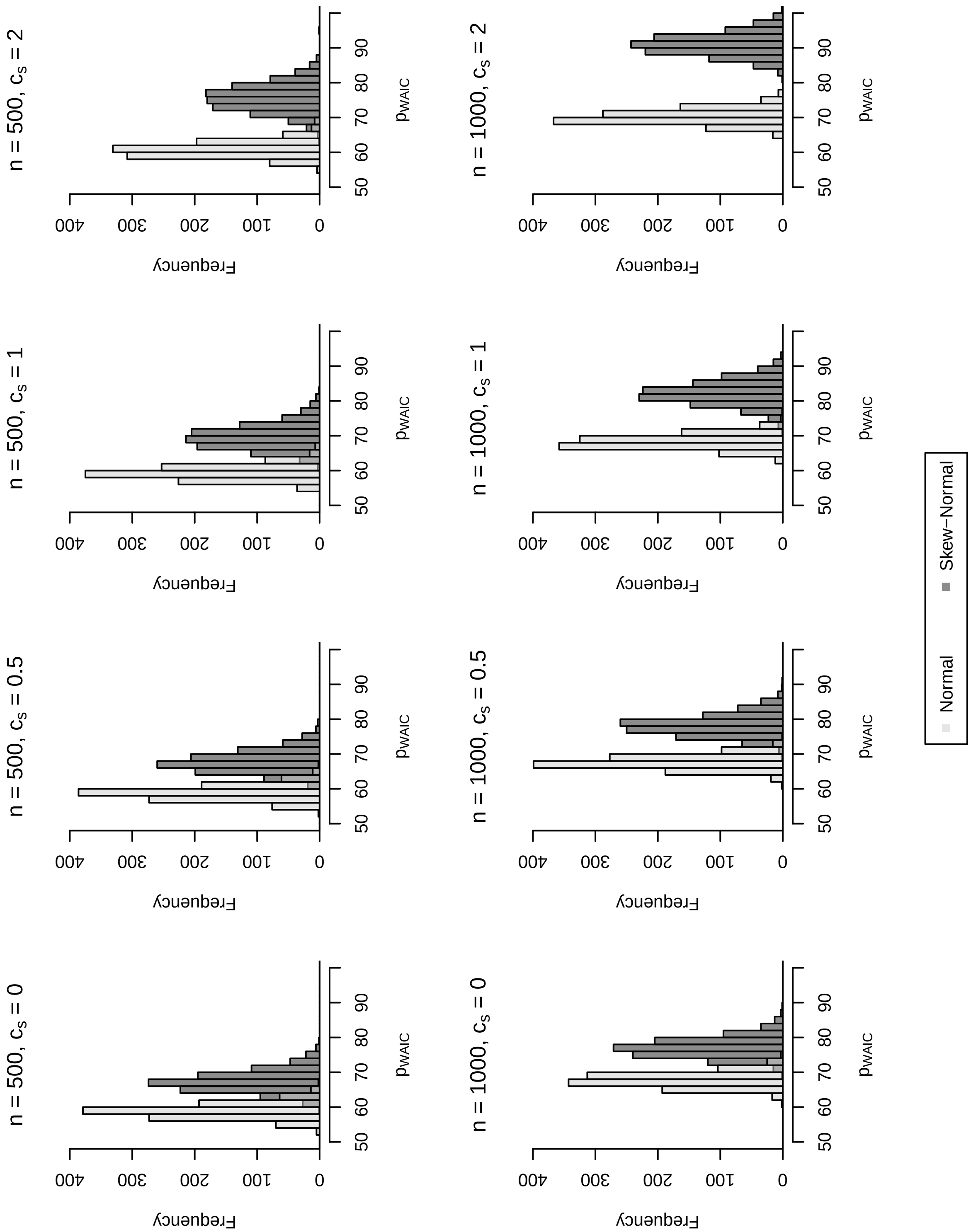

Figure 24: Histograms of the estimated effective number of parameters of the WAIC for both the bivariate skew-normal and the bivariate normal distribution.

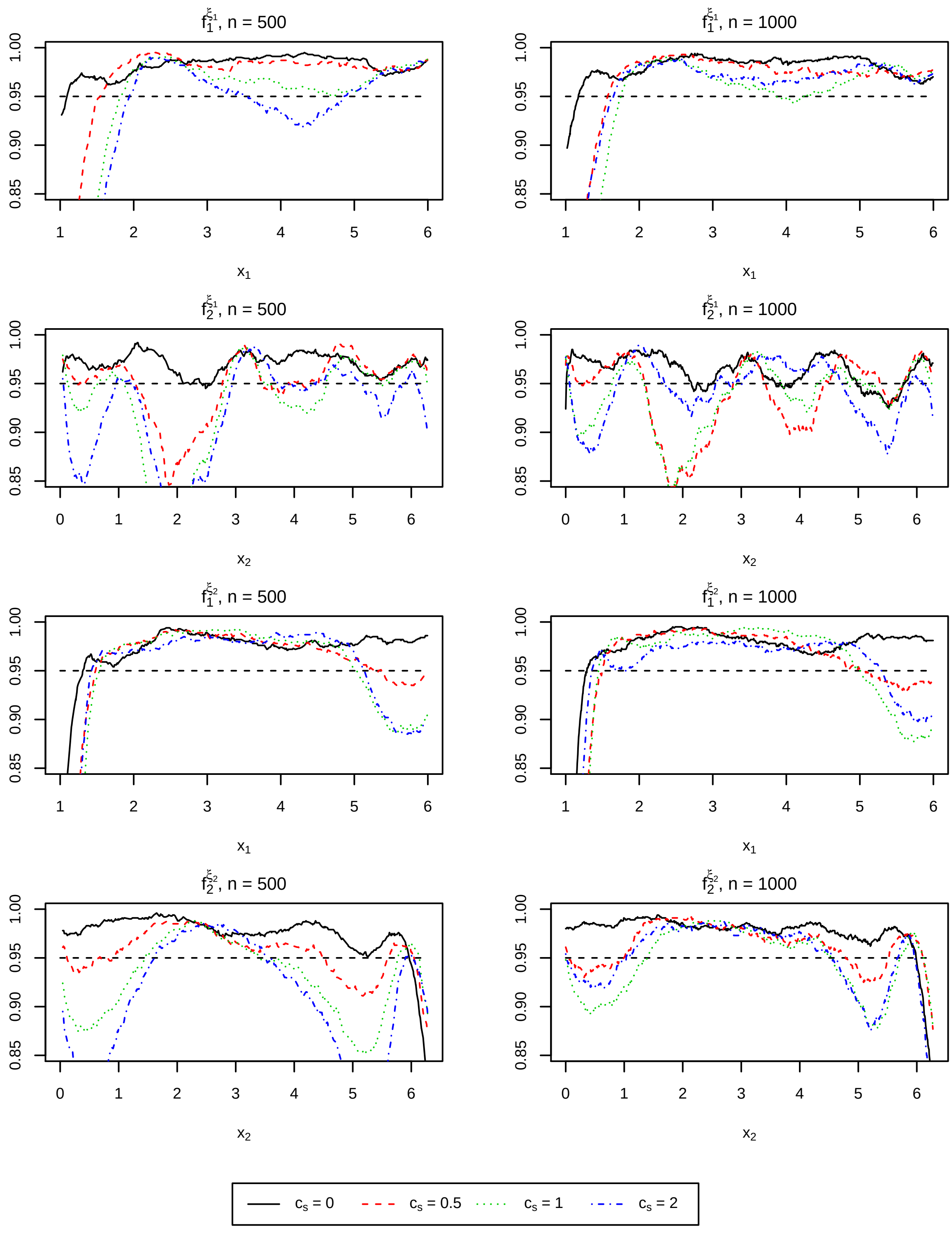

Figure 25: Coverage rates of the nonlinear effects for different values of c_s for both the bivariate skew-normal distribution. (1)

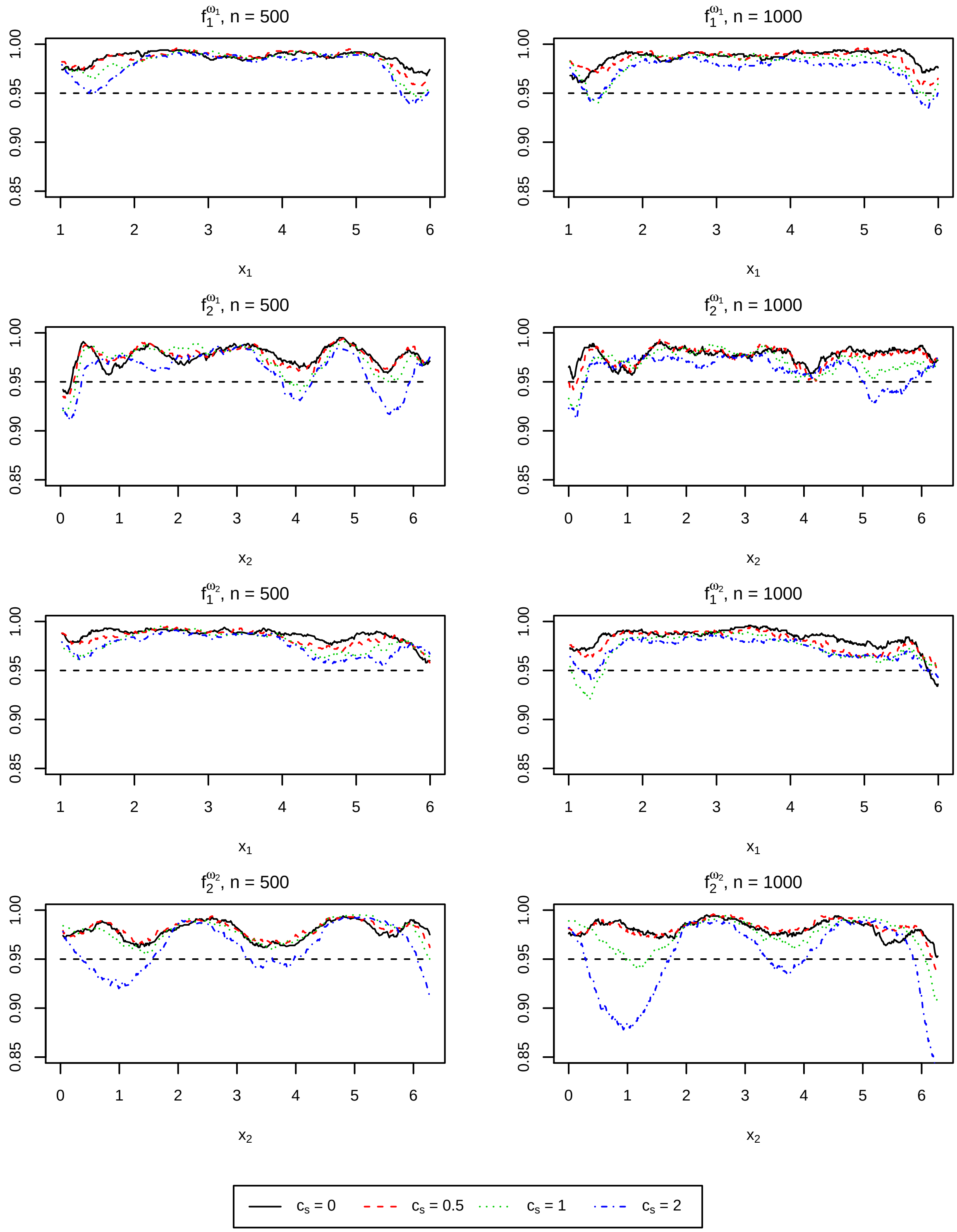

Figure 26: Coverage rates of the nonlinear effects for different values of c_s for both the bivariate skew-normal distribution. (2)

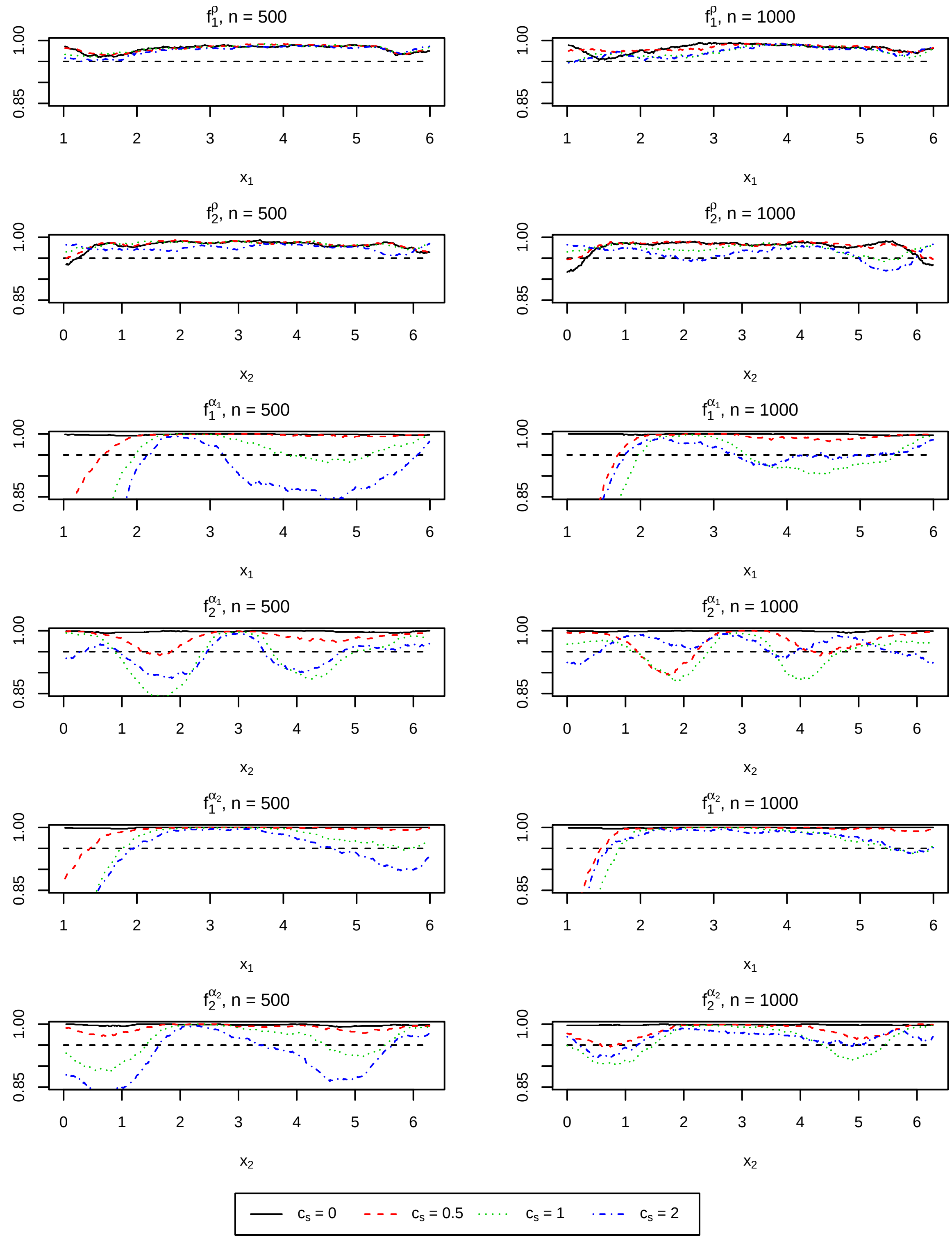

Figure 27: Coverage rates of the nonlinear effects for different values of c_s for both the bivariate skew-normal distribution. (3)

		DIC		DIC_2		WAIC	
	n	500	1000	500	1000	500	1000
c_s	0	0.6	2.3	94.8	93.2	75.1	84.6
	0.5	0.8	0.6	98.5	97.5	67.8	61.4
	1	0.2	0	98.2	97.2	41.9	12.6
	2	0	0	98.5	57	5.3	0

Table 3: Simulation results for the bivariate skew-normal model. The values given in the table are the percentage of the simulations for which the bivariate normal distribution is chosen.

Skew-t Distribution The predictors are given by

$$
\begin{aligned}
f_1^{\xi_1} &= x_1^{-1}, & f_2^{\xi_1} &= \cos(2x_2)\sqrt{x_2}, \\
f_1^{\xi_2} &= 2x_1 - 3\log(x_1)^2, & f_2^{\xi_2} &= 0.1\sin(x_2)\exp(0.5x_2), \\
f_1^{\omega_1} &= 0.7x_1 - 0.1x_1^2, & f_2^{\omega_2} &= \cos(x_2), \\
f_1^{\omega_2} &= 0.025x_1^3 - 0.15x_1^2, & f_2^{\omega_2} &= 0.1\cos(x_2)x_2, \\
f_1^{\rho} &= 0.6\log(x_1), & f_2^{\rho} &= 0.5\cos(x_2), \\
f_1^{\alpha_1} &= c_s\log(x_1 - 0.1x_1^2), & f_2^{\alpha_1} &= c_s 0.5\sin(x_2), \\
f_1^{\alpha_2} &= c_s 1.8x_1^{-0.5}, & f_2^{\alpha_2} &= c_s\tanh\left(\frac{x_2}{2\pi}\right), \\
f_1^{\nu} &= 0.5x_1, & f_2^{\nu} &= \frac{1}{1+x_2}.
\end{aligned}
$$

As before we use $c_s = 0, 0.5, 1, 2$ and $n = 500, 1000$. For this distribution we restrict ourselves to 250 repetitions due to the high computation times.

In Table 4 the results for the comparison of the information criteria are given and in Figures 28, 29 and 30 histograms of the estimated effective number of parameters are given. The general observations are the same as for the bivariate skew-normal distribution, but the criteria perform slightly better. This is in line with the results for the univariate distributions.The same holds for the coverage rates given in Figures 31, 32, 33 and 34.

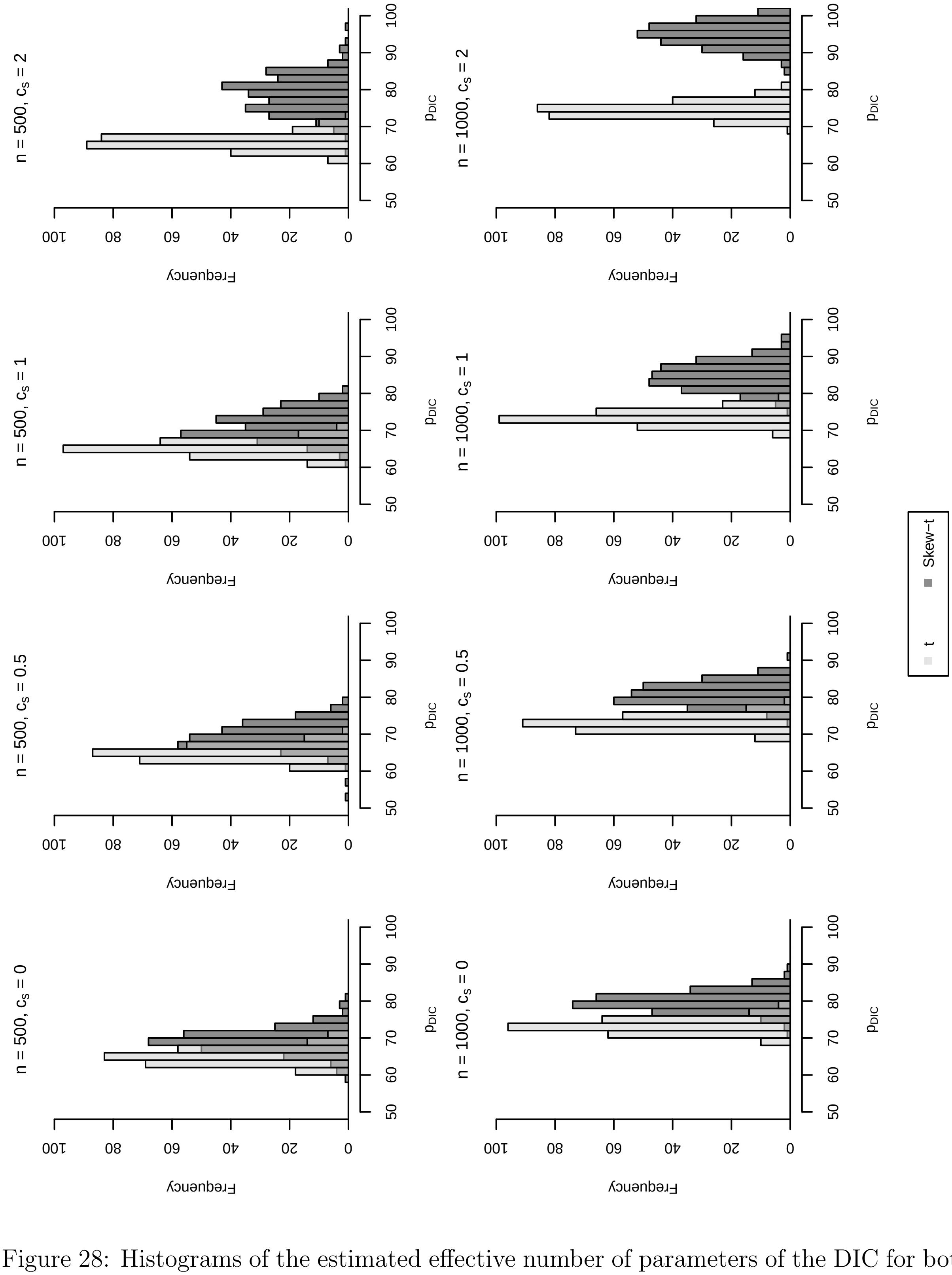

Figure 28: Histograms of the estimated effective number of parameters of the DIC for both the bivariate skew-t and the bivariate t distribution.

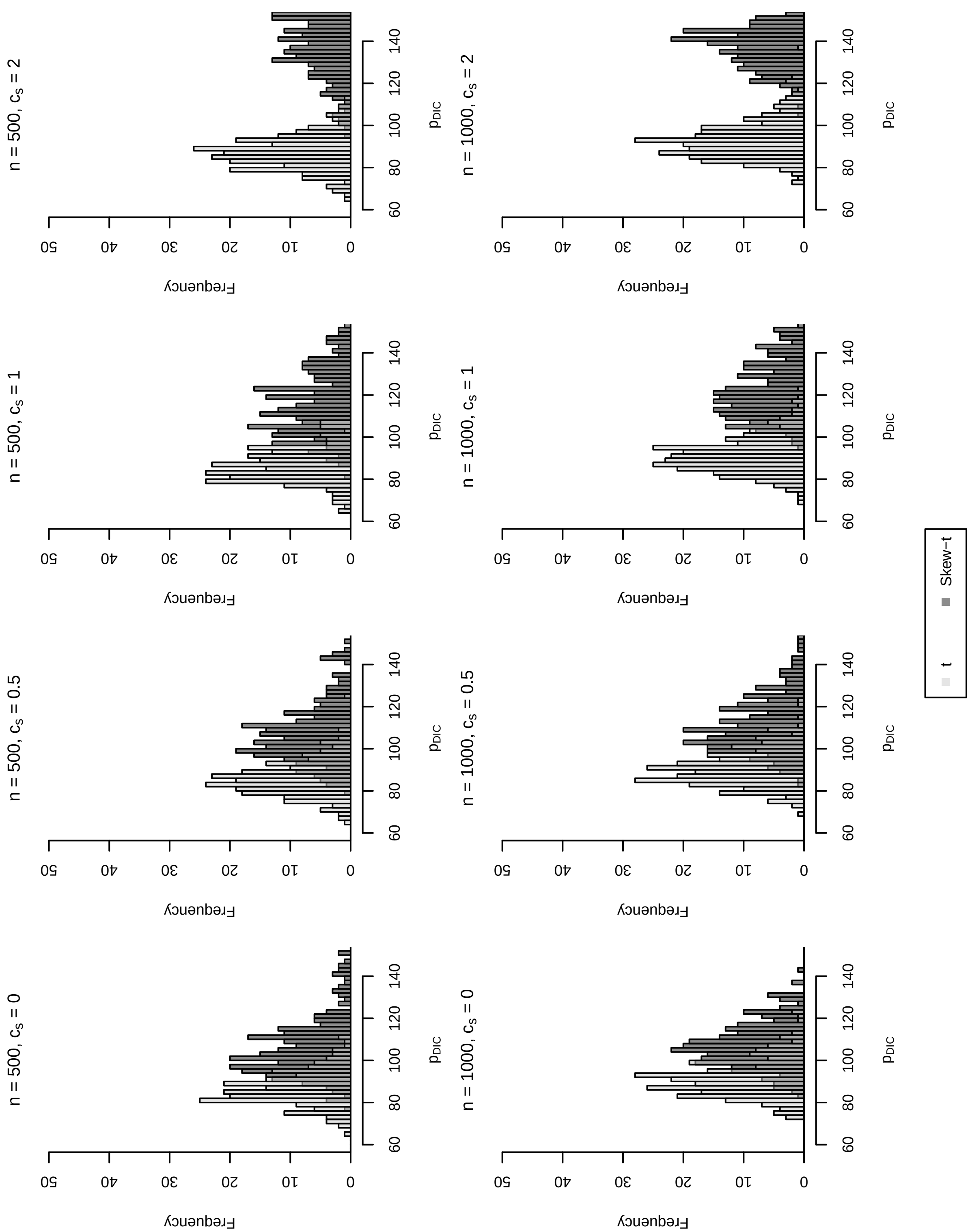

Figure 29: Histograms of the estimated effective number of parameters of the DIC_2 for both the bivariate skew-t and the bivariate t distribution.

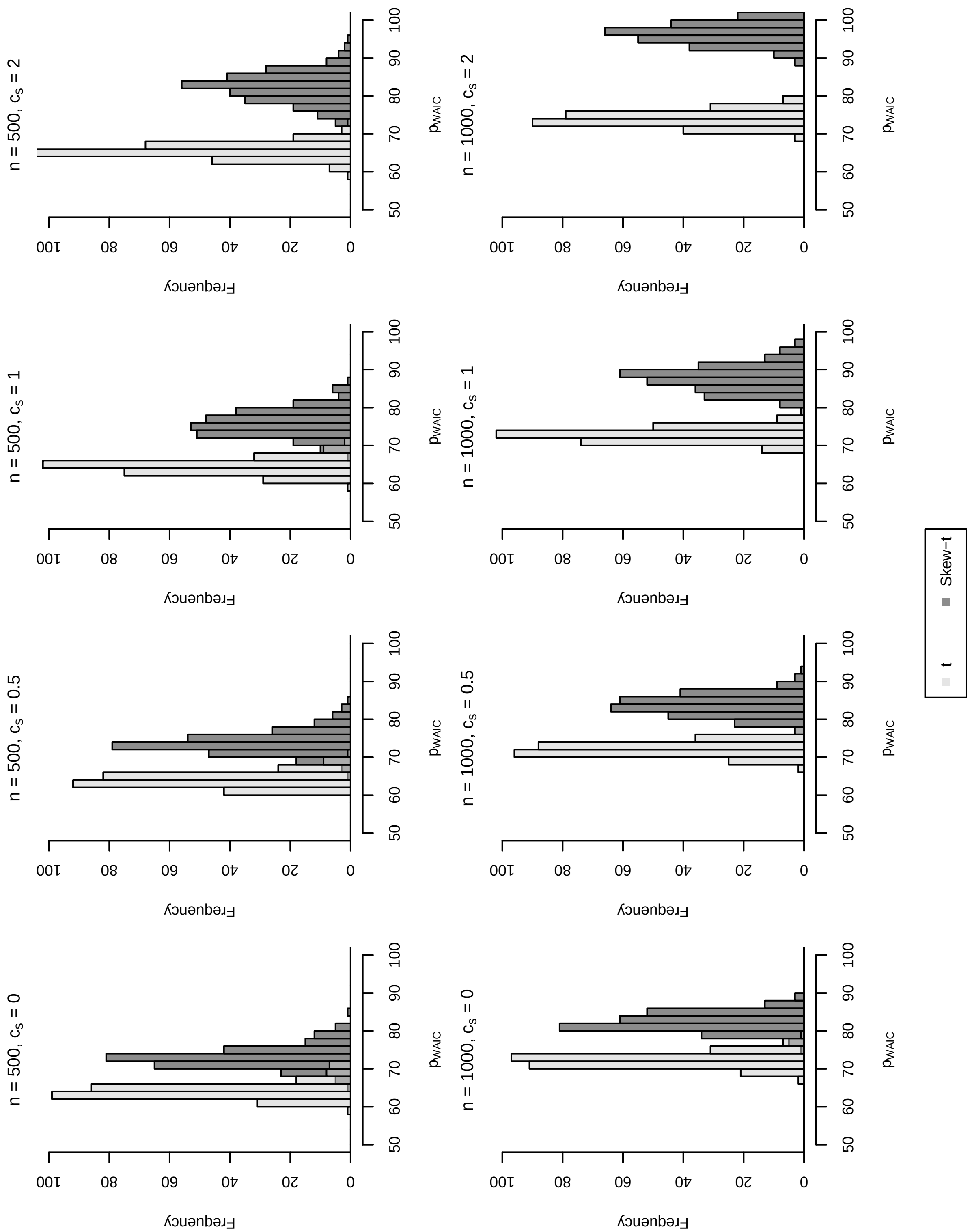

Figure 30: Histograms of the estimated effective number of parameters of the WAIC for both the bivariate skew-t and the bivariate t distribution.

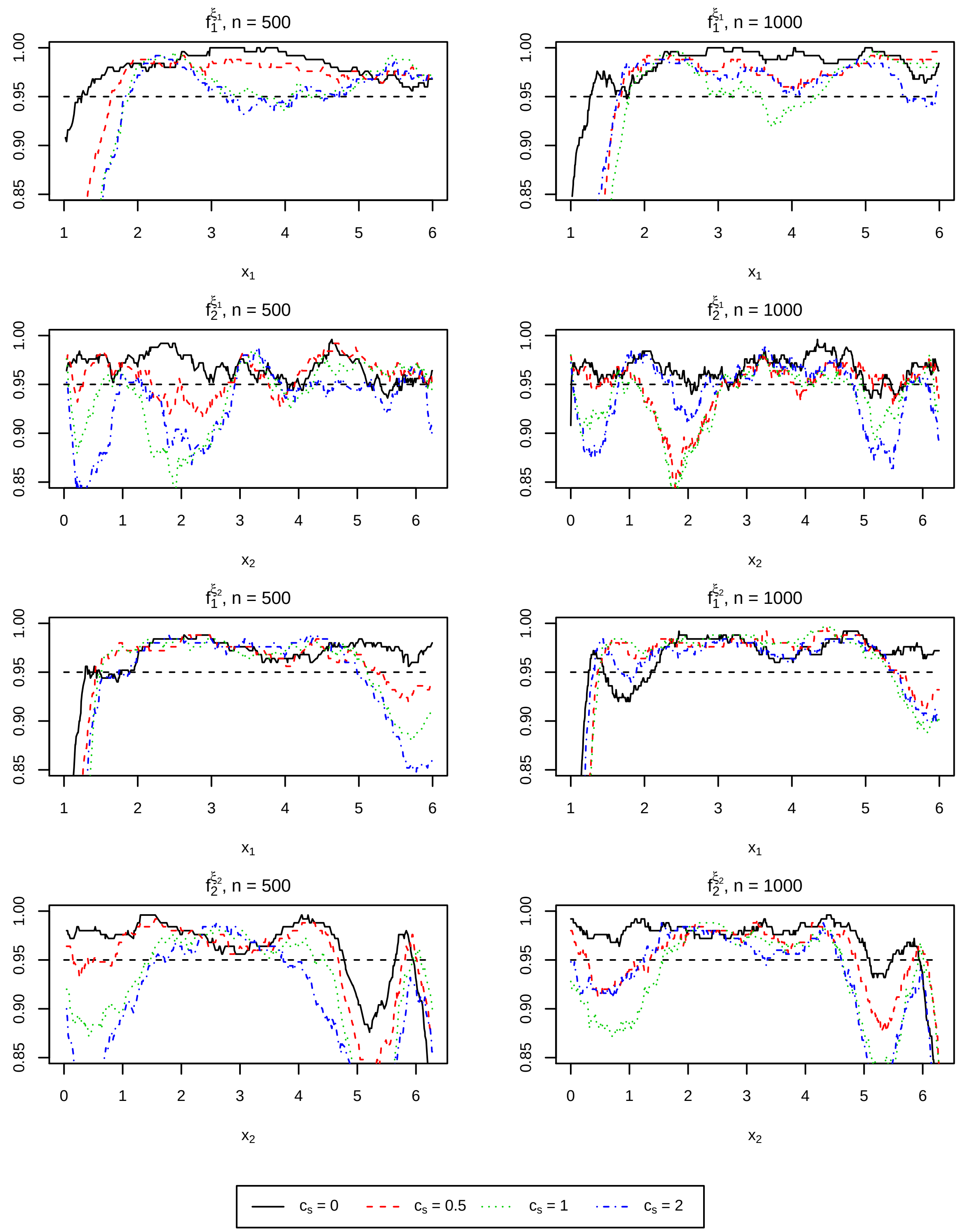

Figure 31: Coverage rates of the nonlinear effects for different values of c_s for both the bivariate skew-t distribution. (1)

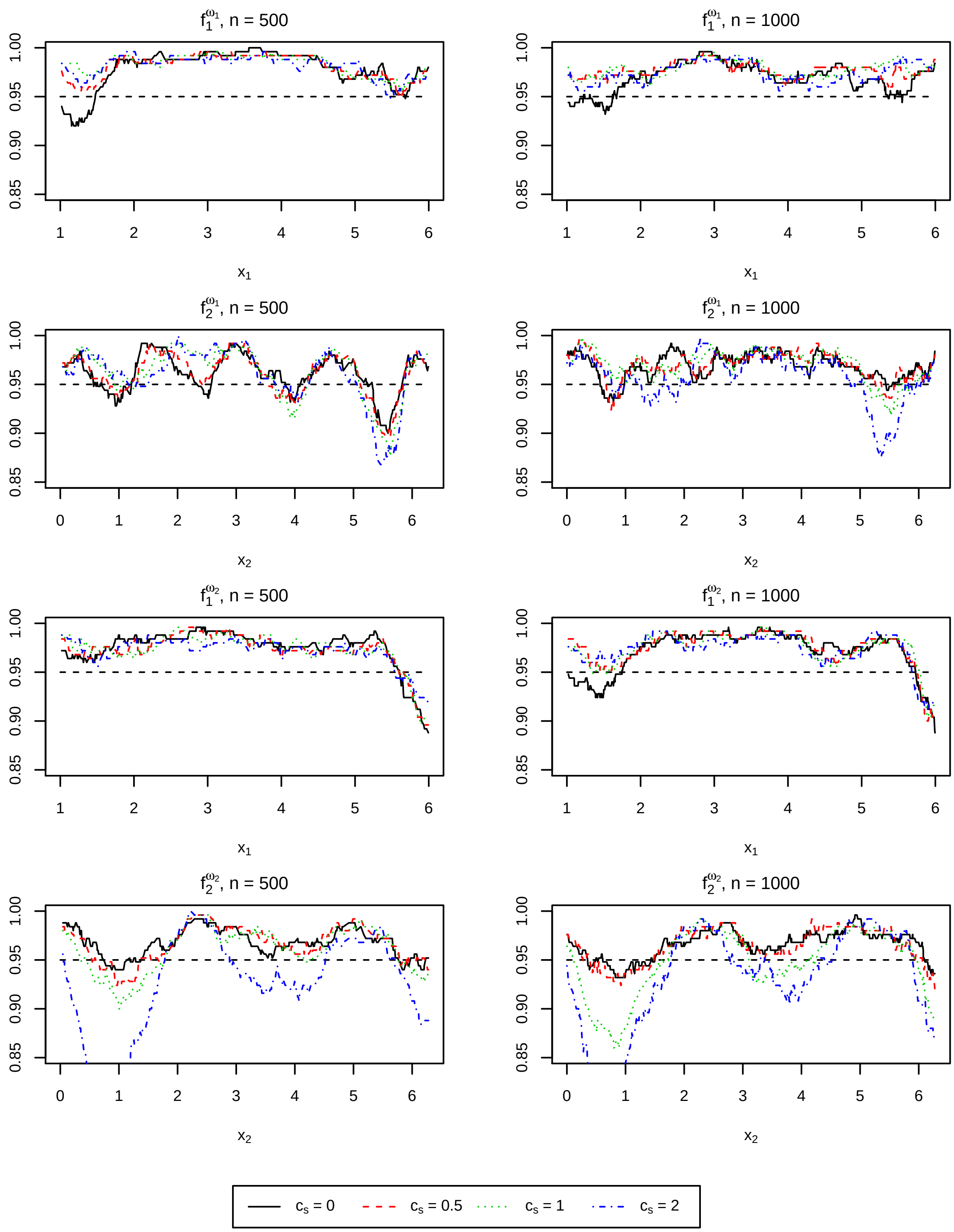

Figure 32: Coverage rates of the nonlinear effects for different values of c_s for both the bivariate skew-t distribution. (2)

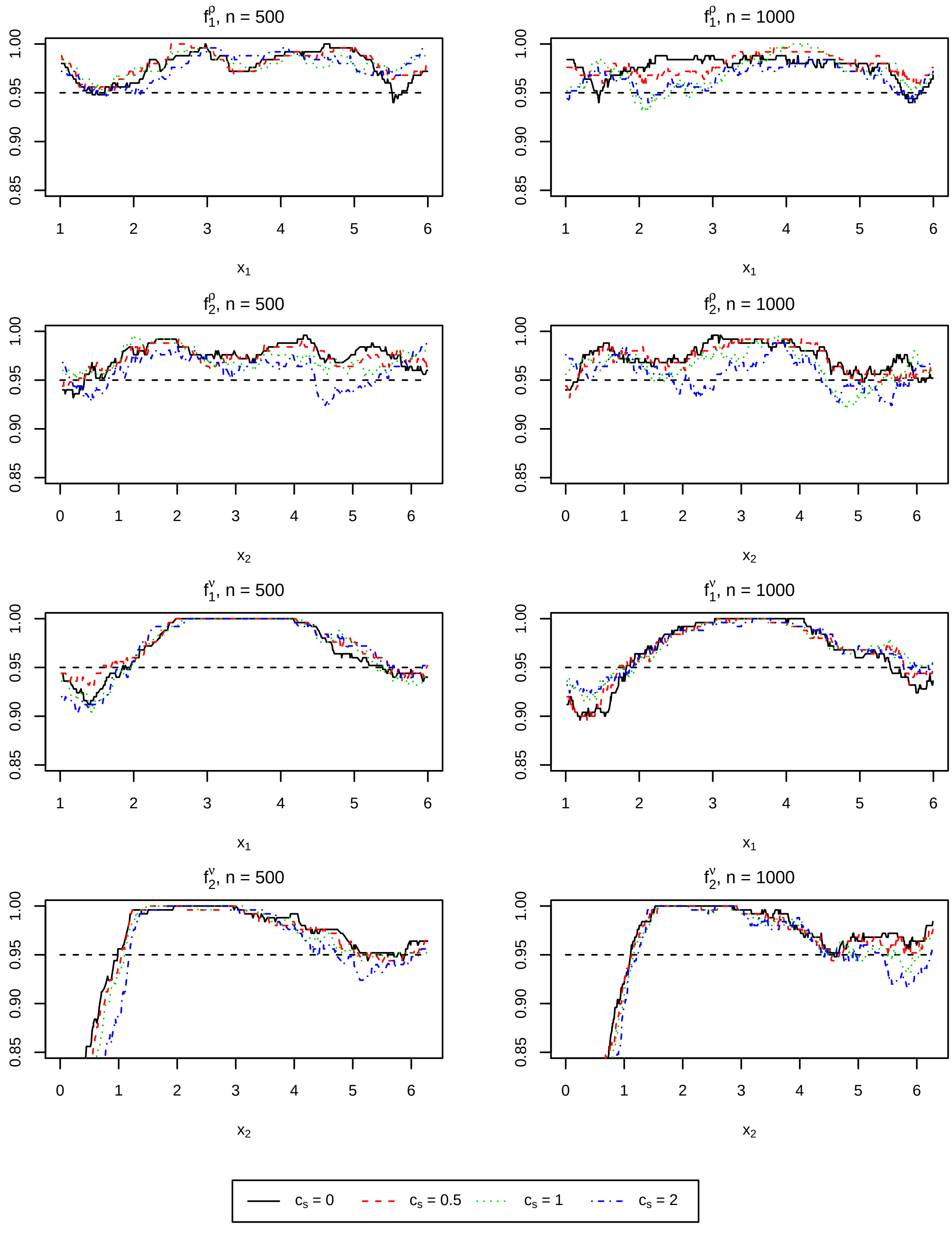

Figure 33: Coverage rates of the nonlinear effects for different values of c_s for both the bivariate skew-t distribution. (3)

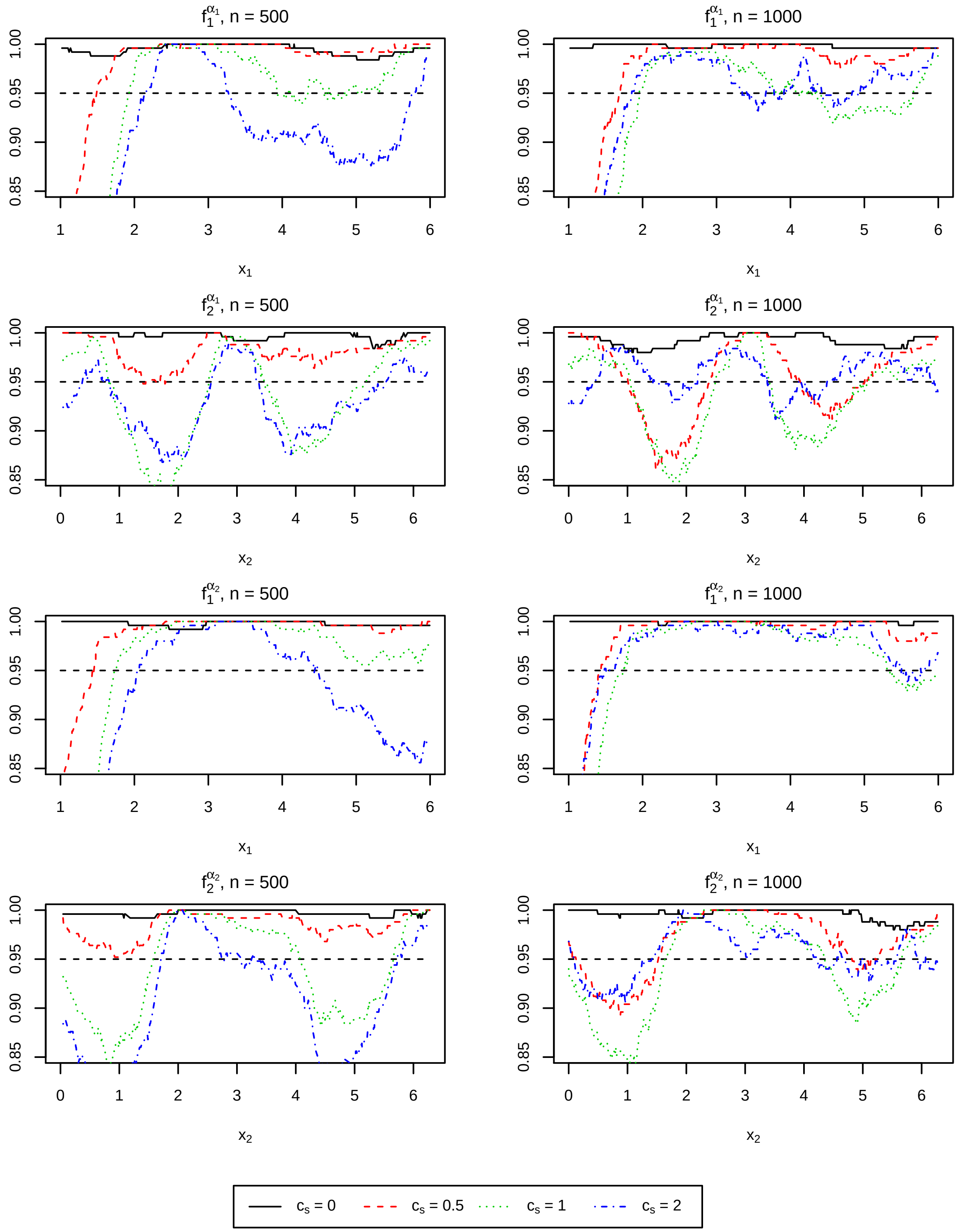

Figure 34: Coverage rates of the nonlinear effects for different values of c_s for both the bivariate skew-t distribution. (4)

		DIC		DIC_2		WAIC	
	n	500	1000	500	1000	500	1000
c_s	0	35.8	58.5	95	96.8	80.1	88.3
	0.5	24.9	22	95.8	93.5	66.2	49.6
	1	5	1.2	68.3	94.8	34.7	6.4
	2	0	0	77.8	57.5	1.8	0

Table 4: Simulation results for the bivariate skew-t model. The values given in the table are the percentage of the simulations for which the bivariate t distribution is chosen.

1.3 Random Effect Simulation

In this exemplary simulation for the estimation of the random effect we used the normal distribution as the response distribution. The equations for the predictors consist of two nonlinear effects and a random effect on μ. The nonlinear effects are given by:

$$
\begin{aligned}
f_1^{\mu} &= \cos(2x_1)\sqrt{x_1}, \\
f_1^{\sigma} &= \sin(x_1).
\end{aligned}
$$

The random effect is based on $n_{GR} = 100$ groups and the effects are skew-normally distributed with $\alpha = 0, 2, 4, 8$ and ω is chosen such that the variance is equal to one. Overall we have $n = 1000$ observations and we used a burnin of 10,000 iterations and 50,000 iterations for the estimation.

α	WAIC	DIC_1
0	0.20	0.21
2	0.20	0.18
4	0.09	0.08
8	0.05	0.04

The results of the model comparison are given in Table 1.3. It is noteworthy, that there is a preference for the more flexible model in both criteria and that the results are very

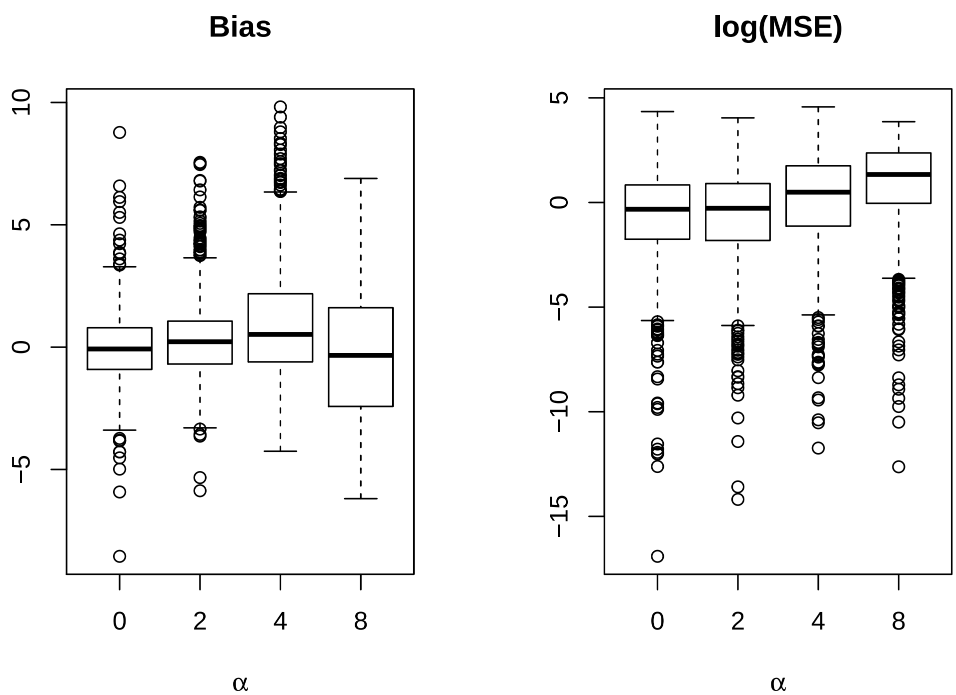

Figure 35: Estimation results for the α parameter of the random effects.

close. As a consequence the usability of the criteria for model choice in this framework are very limited. In Figure 35 the results of the estimation of the parameter α are given. It is worth noting, that this works reasonably well given the number of groups and observations.

2 Application: Childhood Undernutrition in Nigeria

In the following additional plots for the application on undernutrition among children in Nigeria are given:

- Plots of the spatial effects in Figures 36, 37, 38, and 39.
- A visualization of the linear effects in Figure 40.
- Plots of the nonlinear effects in Figures 41 and 42.

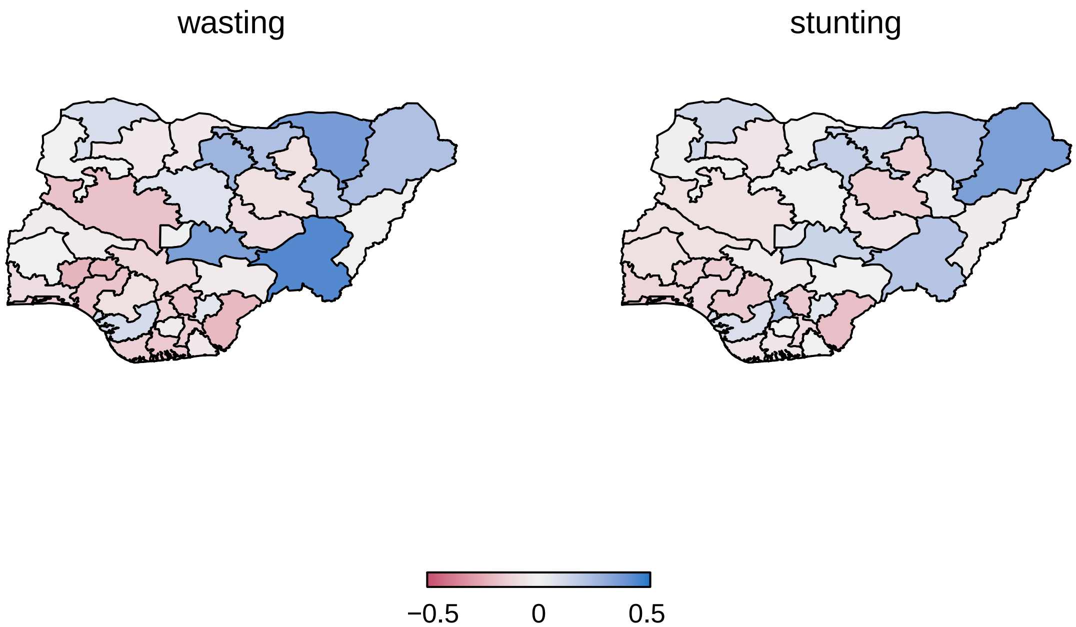

Figure 36: Estimated effects for the parameter ω for the bivariate skew-t distribution in the application to childhood undernutrition.

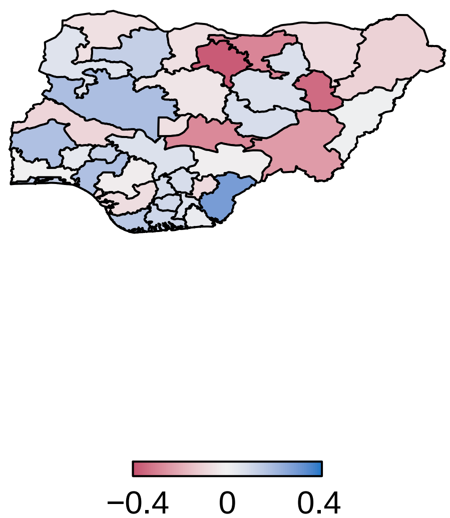

Figure 37: Estimated effects for the parameter ρ for the bivariate skew-t distribution in the application to childhood undernutrition.

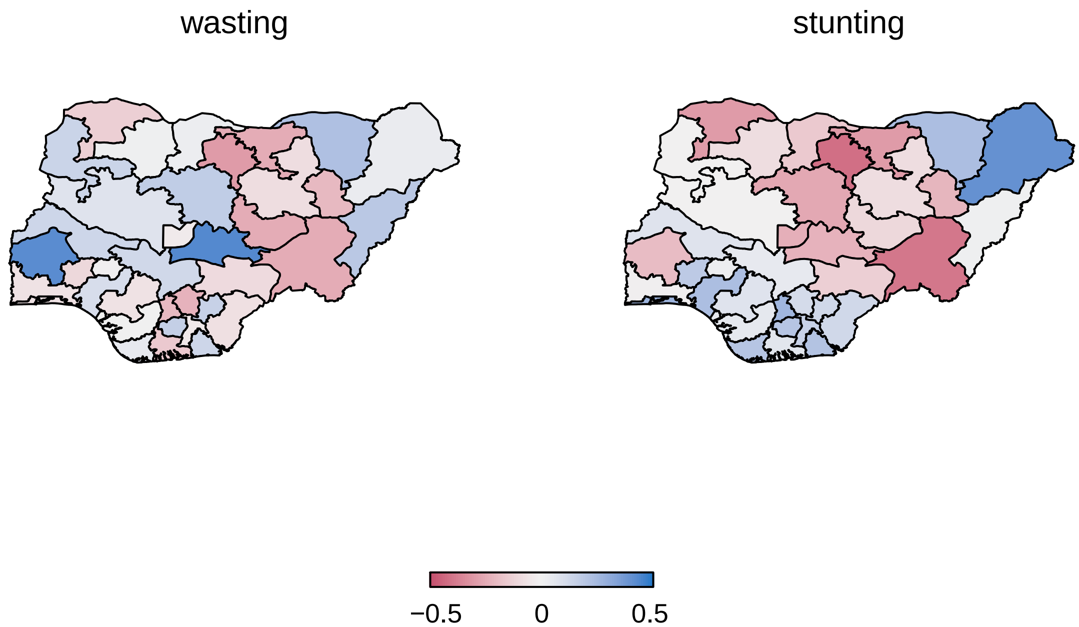

Figure 38: Estimated effects for the parameter α for the bivariate skew-t distribution in the application to childhood undernutrition.

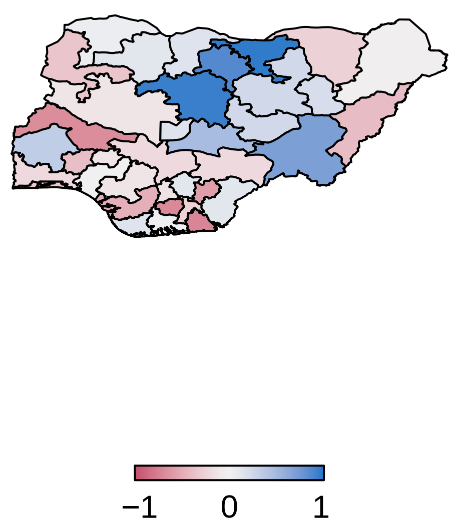

Figure 39: Estimated effects for the parameter ν for the bivariate skew-t distribution in the application to childhood undernutrition.

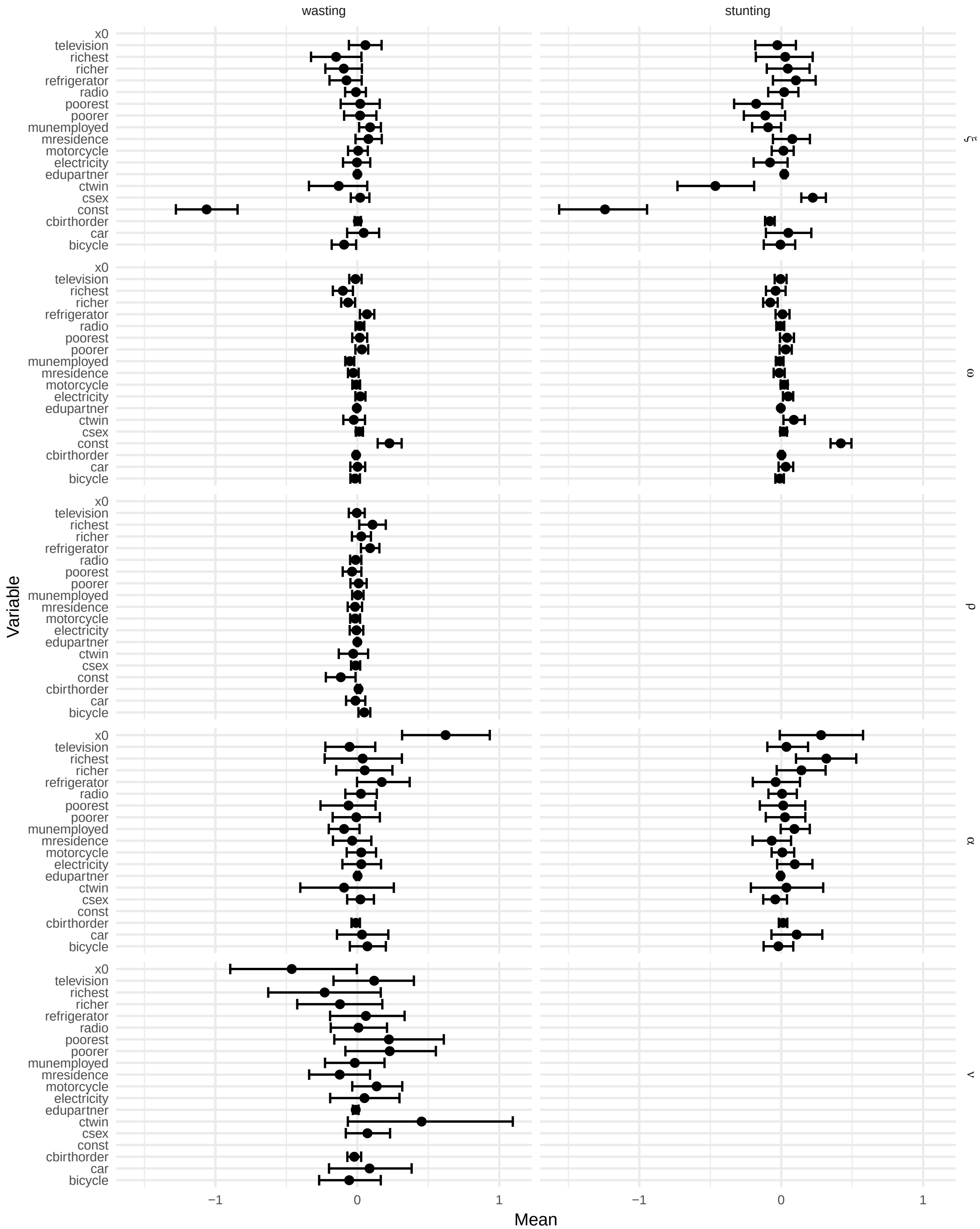

Figure 40: Estimated linear effects for all parameters in the application to childhood undernutrition.

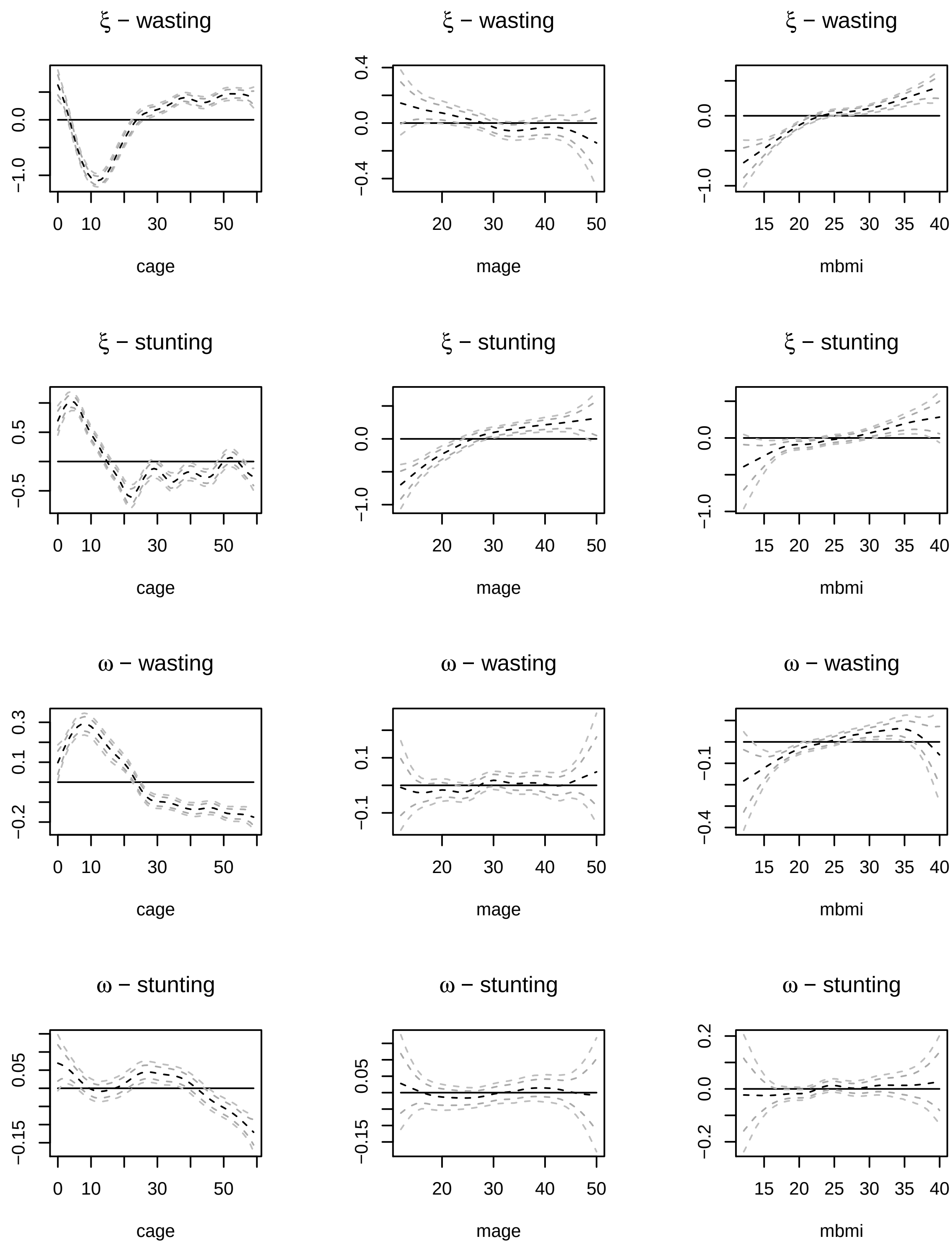

Figure 41: Estimated nonlinear effects with 80 % and 95 % credible intervals for all parameters in the application to childhood undernutrition (1).

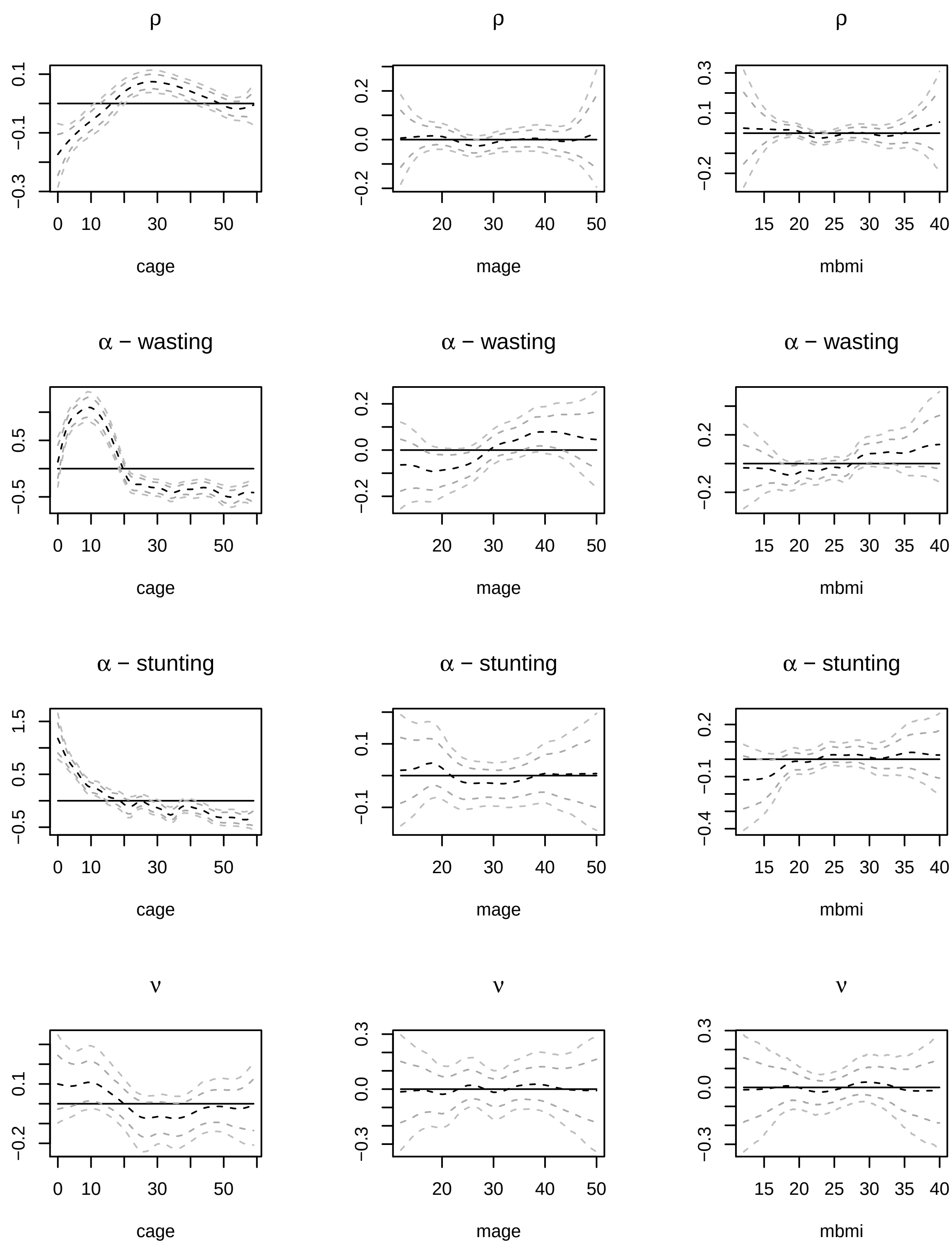

Figure 42: Estimated nonlinear effects with 80 % and 95 % credible intervals for all parameters in the application to childhood undernutrition (2).

3 Theory - Univariate Skew-Normal Distribution

In the following, the formulas used for the implementation will be given.

3.1 Direct Parametrization

The results given here can be found, for example, in Azzalini (1985).

The log-likelihood is given by

$$l(\xi, \omega, \alpha) = const. - \ln(\omega) - \frac{(x-\xi)^2}{2\omega^2} + \zeta_0\left(\alpha\frac{x-\xi}{\omega}\right)$$

where

$$\zeta_0(x) = \ln(2\Phi(x)).$$

Furthermore

$$\zeta_1(x) = \frac{\partial\zeta_0(x)}{\partial x} = \frac{\phi(x)}{\Phi(x)}$$

and

$$\zeta_2(x) = \frac{\partial^2\zeta_0(x)}{(\partial x)^2} = -\zeta_1(x)^2 - x\zeta_1(x).$$

The first partial derivatives of the log-likelihood are given by

$$\frac{\partial}{\partial\xi}l(\xi, \omega, \alpha) = \frac{z}{\omega} - \frac{\alpha}{\omega}\zeta_1(\alpha z)$$

$$\frac{\partial}{\partial\omega}l(\xi, \omega, \alpha) = -\frac{1}{\omega} + \frac{z^2}{\omega} - \frac{\alpha}{\omega}\zeta_1(\alpha z)z$$

$$\frac{\partial}{\partial\alpha}l(\xi, \omega, \alpha) = \zeta_1(\alpha z)z.$$

The second partial derivatives of the log-likelihood are given by

$$\frac{\partial^2}{(\partial\xi)^2}l(\xi, \omega, \alpha) = -\frac{1}{\omega^2} + \frac{\alpha^2}{\omega^2}\zeta_2(\alpha z)$$

$$\frac{\partial^2}{\partial\xi\partial\omega}l(\xi,\omega,\alpha) = -\frac{2z}{\omega^2} + \frac{\alpha}{\omega^2}\zeta_1(\alpha z) + \frac{\alpha^2}{\omega^2}\zeta_2(\alpha z)z$$

$$\frac{\partial^2}{\partial\xi\partial\alpha}l(\xi,\omega,\alpha) = -\frac{1}{\omega} - \frac{\alpha}{\omega}\zeta_2(\alpha z)z$$

$$\frac{\partial^2}{(\partial\omega)^2}l(\xi,\omega,\alpha) = \frac{1}{\omega^2} - \frac{3z^2}{\omega^2} + \frac{2\alpha}{\omega^2}\zeta_1(\alpha z)z + \frac{\alpha^2}{\omega^2}\zeta_2(\alpha z)z^2$$

$$\frac{\partial^2}{\partial\omega\partial\alpha}l(\xi,\omega,\alpha) = -\frac{1}{\omega}\zeta_1(\alpha z)z - \frac{\alpha}{\omega}\zeta_2(\alpha z)z^2$$

$$\frac{\partial^2}{(\partial\alpha)^2}l(\xi,\omega,\alpha) = \zeta_2(\alpha z)z^2.$$

The expected Fisher information is given by

$$E\left(-\frac{\partial^2}{(\partial\xi)^2}l(\xi,\omega,\alpha)\right) = \frac{1+\alpha^2 a_0(\alpha)}{\omega^2}$$

$$E\left(-\frac{\partial^2}{\partial\xi\partial\omega}l(\xi,\omega,\alpha)\right) = \frac{1}{\omega^2}\left(\frac{b\alpha(1+2\alpha^2)}{(1+\alpha^2)^{\frac{3}{2}}} + \alpha^2 a_1(\alpha)\right)$$

$$E\left(-\frac{\partial^2}{\partial\xi\partial\alpha}l(\xi,\omega,\alpha)\right) = \frac{1}{\omega}\left(\frac{b}{(1+\alpha^2)^{\frac{3}{2}}} - \alpha a_1(\alpha)\right)$$

$$E\left(-\frac{\partial^2}{(\partial\omega)^2}l(\xi,\omega,\alpha)\right) = \frac{2+\alpha^2 a_2(\alpha)}{\omega^2}$$

$$E\left(-\frac{\partial^2}{\partial\omega\partial\alpha}l(\xi,\omega,\alpha)\right) = -\frac{\alpha a_2(\alpha)}{\omega}$$

$$E\left(-\frac{\partial^2}{(\partial\alpha)^2}l(\xi,\omega,\alpha)\right) = a_2(\alpha)$$

where

$$a_k(\alpha) = E\left(Z^k\zeta_1(\alpha Z)^2\right). \tag{1}$$

As it is takes a lot of computation time to evaluate the expectations $a_k(\alpha)$, modified versions of the approximations given by Bayes and Branco in Bayes and Branco (2007). In Bayes and Branco (2007) the approximations are given by

$$\tilde{a}_k = \frac{k!}{2^{\frac{k}{2}-1}\pi(\frac{k}{2})!}\left(1+\frac{2\alpha^2}{\frac{\pi^2}{4}}\right)^{-\frac{k+1}{2}}$$

if k is even and

$$\begin{aligned}\tilde{a}_k = & -\frac{2^{\frac{k}{2}+1}}{\pi^{\frac{3}{2}}}\left(\frac{k-1}{2}\right)!\alpha\left(1+\frac{2\alpha^2}{\frac{\pi^2}{4}}\right)^{-\frac{k}{2}-1}\left(1+\alpha^2\left(1+\frac{2\alpha^2}{\frac{\pi^2}{4}}\right)^{-1}\right)^{-\frac{1}{3}} \\ & \cdot\left(\sum_{j=0}^{\frac{k-1}{2}} c_j\alpha^j\left(1+\frac{2\alpha^2}{\frac{\pi^2}{4}}\right)^{-\frac{j}{2}}\right)\end{aligned}$$

if k is odd, while

$$c_j = \frac{2j-1}{2j}c_{j-1},\ c_0 = 1$$

are used as a replacement. The modified approximations for $k = 0, 1, 2$ are given by

$$\tilde{a}_0 = \frac{2}{\pi}\left(1+\frac{8(0.96\alpha)^2}{\pi^2}\right)^{-\frac{1}{2}},$$

$$\tilde{a}_1 = -0.975\left(\frac{2}{\pi}\right)^{\frac{3}{2}}\alpha\left(1+\frac{8\alpha^2}{\pi^2}\right)^{-\frac{3}{2}}\left(1+\alpha^2\left(1+\frac{8\alpha^2}{\pi^2}\right)^{-1}\right)^{-\frac{1}{3}}$$

and

$$\tilde{a}_2 = \frac{2}{\pi}\left(1+\frac{8(0.96\alpha)^2}{\pi^2}\right)^{-\frac{3}{2}}.$$

The modifications are necessary to ensure positive values for the expected Fisher information for all parameters in both parameterizations.

In the implementation the following link functions are used

$$
\begin{aligned}
\xi &: \quad \text{identity} \\
\omega &: \quad \log \\
\alpha &: \quad \text{identity}
\end{aligned}
$$

and therefore the first partial derivatives of the log-likelihood with respect to the predictors for the parameters are given by

$$
\begin{aligned}
v_\xi &= \frac{z}{\omega} - \frac{\alpha}{\omega}\zeta_1(\alpha z) \\
v_\omega &= -\frac{1}{\omega^2} + \frac{z^2}{\omega^2} - \frac{\alpha}{\omega^2}\zeta_1(\alpha z)z \\
v_\alpha &= \zeta_1(\alpha z)z
\end{aligned}
$$

and diagonal elements of the expected Fisher information are

$$
\begin{aligned}
w_\xi &= \frac{1 + \alpha^2 a_0(\alpha)}{\omega^2} \\
w_\omega &= 2 + \alpha^2 a_2(\alpha) \\
w_\alpha &= a_2(\alpha)
\end{aligned}
$$

which are all positive as each variable in the formulas is positive.

3.2 Centered Parametrization

The direct parameters can be obtained using:

$$
\begin{aligned}
\xi &= \mu - b\omega\delta \\
\omega &= \frac{\sigma}{\sqrt{1 - b^2\delta^2}} \\
\alpha &= \frac{R}{\sqrt{b^2 - (1 - b^2)R^2}} \\
R &= \left(\frac{2\gamma}{4 - \pi}\right)^{\frac{1}{3}} \\
\delta &= \frac{\alpha}{\sqrt{1 + \alpha^2}}
\end{aligned}
$$

The derivatives of the parameter transfomations are given by (only non-zero derivatives are given):

$$
\begin{aligned}
\frac{\partial \xi}{\partial \mu} =& 1 \\
\frac{\partial \xi}{\partial \sigma} =& -\frac{b\delta}{\sqrt{1-b^2\delta^2}} \\
\frac{\partial \xi}{\partial \gamma} =& \frac{\partial \xi}{\partial \delta}\frac{\partial \delta}{\partial \alpha}\frac{\partial \alpha}{\partial \gamma} \\
=& \left(-\frac{b\sigma}{\sqrt{1-b^2\delta^2}} - \frac{b^3\sigma\delta^2}{(1-b^2\delta^2)^{\frac{3}{2}}}\right)\left(\frac{1}{(1+\alpha^2)^{\frac{3}{2}}}\right) \\
&\cdot \left(\frac{1}{\sqrt{b^2-(1-b^2)R^2}} + \frac{(1-b^2)R^2}{(b^2-(1-b^2)R^2)^{\frac{3}{2}}}\right)\left(\left(\frac{2}{(4-\pi)\gamma^2}\right)^{\frac{1}{3}}\frac{1}{3}\right)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \omega}{\partial \sigma} =& \frac{1}{\sqrt{1-b^2\delta^2}} \\
\frac{\partial \omega}{\partial \gamma} =& \frac{\partial \omega}{\partial \delta}\frac{\partial \delta}{\partial \alpha}\frac{\partial \alpha}{\partial \gamma} \\
=& \left(\frac{b^2\sigma\delta}{(1-b^2\delta^2)^{\frac{3}{2}}}\right)\left(\frac{1}{(1+\alpha^2)^{\frac{3}{2}}}\right) \\
&\cdot \left(\frac{1}{\sqrt{b^2-(1-b^2)R^2}} + \frac{(1-b^2)R^2}{(b^2-(1-b^2)R^2)^{\frac{3}{2}}}\right)\left(\left(\frac{2}{(4-\pi)\gamma^2}\right)^{\frac{1}{3}}\frac{1}{3}\right)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \alpha}{\partial \gamma} =& \frac{\partial \alpha}{\partial R}\frac{\partial R}{\partial \gamma} \\
=& \left(\frac{1}{\sqrt{b^2-(1-b^2)R^2}} + \frac{(1-b^2)R^2}{(b^2-(1-b^2)R^2)^{\frac{3}{2}}}\right)\left(\left(\frac{2}{(4-\pi)\gamma^2}\right)^{\frac{1}{3}}\frac{1}{3}\right)
\end{aligned}
$$

In the implementation the following link functions are used

$$
\begin{aligned}
\mu: &\quad \text{identity} \\
\sigma &\quad \log \\
\gamma: &\quad \gamma_{MAX} \cdot \frac{.}{\sqrt{1+.^2}}
\end{aligned}
$$

where $\gamma_{MAX} \approx 0.9952716$ is the maximum of the skewness. Using the derivatives given above and results for the direct parametrization, the score function and expected Fisher information for the predictors are given by

$$v_\mu = v_\xi \frac{\partial \xi}{\partial \mu} \frac{\partial \mu}{\partial \eta_\mu} = v_\xi = \frac{z}{\omega} - \frac{\alpha}{\omega} \zeta_1(\alpha z)$$

$$v_\sigma = \left(v_\xi \frac{\partial \xi}{\partial \sigma} + v_\omega \frac{\partial \omega}{\partial \sigma} \right) \frac{\partial \sigma}{\partial \eta_\sigma} = \frac{z^2 - b\delta z - 1 - \alpha \zeta_1(\alpha z)(z - b\delta)}{\sqrt{1 - b^2\delta^2}\omega} \exp(\eta_\sigma)$$

$$\begin{aligned} v_\gamma &= \left(v_\xi \frac{\partial \xi}{\partial \gamma} + v_\omega \frac{\partial \omega}{\partial \gamma} + v_\alpha \frac{\partial \alpha}{\partial \gamma} \right) \frac{\partial \gamma}{\partial \eta_\gamma} \\ &= \Bigg\{ \Bigg[\alpha b \zeta_1(\alpha z) - zb + \frac{b^2\delta}{1 - b^2\delta^2} \Bigg(- zb\delta + \alpha(b\delta - z)\zeta_1(\alpha z) - 1 + z^2 \Bigg) \Bigg] \\ &\quad \cdot \frac{1}{(1+\alpha)^{\frac{3}{2}}} + z\zeta_1(\alpha z) \Bigg\} \frac{\partial \alpha}{\partial \gamma} \frac{\partial \gamma}{\partial \eta_\gamma} \end{aligned}$$

$$w_{\mu,\mu} = w_{\xi,\xi} \left(\frac{\partial \xi}{\partial \mu} \right)^2 = w_{\xi,\xi} = \frac{1 + \alpha^2 a_0(\alpha)}{\omega^2}$$

$$\begin{aligned} w_{\sigma,\sigma} &= \left(w_{\xi,\xi} \left(\frac{\partial \xi}{\partial \sigma} \right)^2 + 2 w_{\xi,\omega} \frac{\partial \xi}{\partial \sigma} \frac{\partial \omega}{\partial \sigma} + w_{\omega,\omega} \left(\frac{\partial \omega}{\partial \sigma} \right)^2 \right) \left(\frac{\partial \sigma}{\partial \eta_\sigma} \right)^2 \\ &= \left\{ b^2\delta \left[\delta \left(1 + \alpha^2 a_0(\alpha) - \frac{2(1-\alpha^2)}{1+\alpha^2} \right) - 2\alpha^2 a_1(\alpha) \right] + 2 + \alpha^2 a_2(\alpha) \right\} \\ &\quad \cdot \frac{1}{\omega^2(1 - b^2\delta^2)} \left(\frac{\partial \sigma}{\partial \eta_\sigma} \right)^2 \end{aligned}$$

$$
\begin{aligned}
w_{\gamma,\gamma} = &\left(w_{\xi,\xi} \left(\frac{\partial \xi}{\partial \gamma} \right)^2 + 2 w_{\xi,\omega} \frac{\partial \xi}{\partial \gamma} \frac{\partial \omega}{\partial \gamma} + 2 w_{\xi,\alpha} \frac{\partial \xi}{\partial \gamma} \frac{\partial \alpha}{\partial \gamma} \right. \\
&\quad + w_{\omega,\omega} \left(\frac{\partial \omega}{\partial \gamma} \right)^2 + 2 w_{\omega,\alpha} \frac{\partial \omega}{\partial \gamma} \frac{\partial \alpha}{\partial \gamma} \\
&\quad \left. + w_{\alpha,\alpha} \left(\frac{\partial \alpha}{\partial \gamma} \right)^2 \right) \left(\frac{\partial \gamma}{\partial \eta_\gamma} \right)^2 \\
= &\Bigg[a_2(\alpha) + \frac{2}{(1+\alpha^2)^{\frac{3}{2}}} \left\{ b\alpha a_1(\alpha) + (b\delta\alpha a_1(\alpha) - \alpha a_2(\alpha)) \frac{b^2\delta}{1-b^2\delta^2} \right\} \\
&\quad + \frac{1}{(1+\alpha^2)^3} \Bigg\{ - b^2(1 - \alpha^2 a_0(\alpha)) \\
&\quad - \frac{2b^3\delta}{1-b^2\delta^2} \left(b\delta^3 + \alpha^2 a_1(\alpha) + b\delta - b\delta\alpha^2 a_0(\alpha) \right) \\
&\quad + \frac{b^4\delta^2}{(1-b^2\delta^2)^2} \Big(2 - b^2\delta^2(1+2\delta^2) \\
&\quad + \alpha^2 (a_2(\alpha) - 2b\delta a_1(\alpha) + b^2\delta^2 a_0(\alpha) \Big) \Bigg\} \Bigg] \\
&\quad \left(\frac{1}{\sqrt{b^2 - (1-b^2)R^2}} + \frac{(1-b^2)R^2}{(b^2-(1-b^2)R^2)^{\frac{3}{2}}} \right)^2 \\
&\quad \left(\left(\frac{2}{(4-\pi)\gamma^2} \right)^{\frac{1}{3}} \frac{1}{3} \right)^2 \left(\frac{\partial \gamma}{\partial \eta_\gamma} \right)^2
\end{aligned}
$$

4 Theory - Bivariate Skew-Normal Distribution

The log-likelihood is given by

$$
\begin{aligned}
l(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha}) = &const. - \frac{1}{2} \log \left(\omega_1^2 \omega_2^2 (1-\rho^2) \right) \\
&- \frac{1}{2} \left\{ \frac{\omega_2^2 (y_1 - \xi_1)^2 - 2\omega_1\omega_2\rho(y_1-\xi_1)(y_2-\xi_2) + \omega_1^2 (y_2-\xi_2)^2}{\omega_1^2\omega_2^2(1-\rho^2)} \right\} \\
&+ \zeta_0 \left(\frac{\alpha_1}{\omega_1}(y_1 - \xi_1) + \frac{\alpha_2}{\omega_2}(y_2 - \xi_2) \right).
\end{aligned}
$$

The first partial derivatives are given by

$$\begin{aligned}\frac{\partial}{\partial \xi_1} l(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha}) =& \frac{y_1 - \xi_1}{\omega_1^2(1-\rho^2)} - \frac{\rho(y_2 - \xi_2)}{\omega_1\omega_2(1-\rho^2)} \\ & - \frac{\alpha_1}{\omega_1}\zeta_1\left(\frac{\alpha_1}{\omega_1}(y_1 - \xi_1) + \frac{\alpha_2}{\omega_2}(y_2 - \xi_2)\right)\end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial \omega_1} l(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha}) =& -\frac{1}{\omega_1} + \frac{(y_1 - \xi_1)^2}{\omega_1^3(1-\rho^2)} - \frac{\rho(y_1 - \xi_1)(y_2 - \xi_2)}{\omega_1^2\omega_2(1-\rho^2)} \\ & - \frac{\alpha_1(y_1 - \xi_1)}{\omega_1^2}\zeta_1\left(\frac{\alpha_1}{\omega_1}(y_1 - \xi_1) + \frac{\alpha_2}{\omega_2}(y_2 - \xi_2)\right)\end{aligned}$$

$$\begin{aligned}\frac{\partial}{\partial \rho} l(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha}) =& \frac{(y_1 - \xi_1)(y_2 - \xi_2)}{\omega_1\omega_2(1-\rho^2)} + \cdot \frac{\rho}{1-\rho^2}\Bigg\{1 - \\ & \frac{\omega_2^2(y_1 - \xi_1)^2 - 2\omega_1\omega_2\rho(y_1 - \xi_1)(y_2 - \xi_2) + \omega_1^2(y_2 - \xi_2)^2}{\omega_1^2\omega_2^2(1-\rho^2)}\Bigg\}\end{aligned}$$

$$\frac{\partial}{\partial \alpha_1} l(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha}) = \frac{y_1 - \xi_1}{\omega_1}\zeta_1\left(\frac{\alpha_1}{\omega_1}(y_1 - \xi_1) + \frac{\alpha_2}{\omega_2}(y_2 - \xi_2)\right)$$

where the derivatives with respect to ξ_2, ω_2 and α_2 can be obtained by exchanging the ones and twos. Derivatives for vector valued parameters can be found, for example, in Azzalini (2013) or Arellano-Valle and Azzalini (2008).

The necessary second partial derivatives are given by

$$\frac{\partial^2}{(\partial \xi_1)^2} l(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha}) = -\frac{1}{\omega_1^2(1-\rho^2)} + \frac{\alpha_1^2}{\omega_1^2}\zeta_2\left(\frac{\alpha_1}{\omega_1}(y_1 - \xi_1) + \frac{\alpha_2}{\omega_2}(y_2 - \xi_2)\right)$$

$$\begin{aligned}\frac{\partial^2}{(\partial \omega_1)^2} l(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha}) =& \frac{1}{\omega_1^2} - 3\frac{(y_1 - \xi_1)^2}{\omega_1^4(1-\rho^2)} + 2\frac{\rho(y_1 - \xi_1)(y_2 - \xi_2)}{\omega_1^3\omega_2(1-\rho^2)} \\ & + 2\alpha_1\frac{y_1 - \xi_1}{\omega_1^3}\zeta_1\left(\frac{\alpha_1}{\omega_1}(y_1 - \xi_1) + \frac{\alpha_2}{\omega_2}(y_2 - \xi_2)\right) \\ & + \alpha_1^2\frac{(y_1 - \xi_1)^2}{\omega_1^4}\zeta_2\left(\frac{\alpha_1}{\omega_1}(y_1 - \xi_1) + \frac{\alpha_2}{\omega_2}(y_2 - \xi_2)\right)\end{aligned}$$

$$\begin{aligned}\frac{\partial^2}{(\partial\rho)^2} l(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha}) = &\left\{\frac{1}{1-\rho^2} + \frac{4\rho^2}{(1-\rho^2)^2}\right\} \\ &\cdot \left\{1 - \frac{(y_1-\xi_1)^2}{\omega_1^2(1-\rho^2)} + 2\frac{\rho(y_1-\xi_1)(y_2-\xi_2)}{\omega_1\omega_2(1-\rho^2)} - \frac{(y_2-\xi_2)^2}{\omega_2^2(1-\rho^2)}\right\} \\ &- \frac{2\rho^2}{(1-\rho^2)^2} + \frac{4\rho(y_1-\xi_1)(y_2-\xi_2)}{\omega_1\omega_2(1-\rho^2)^2}\end{aligned}$$

$$\frac{\partial^2}{(\partial\alpha_1)^2} l(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha}) = \frac{(y_1-\xi_1)^2}{\omega_1^2}\zeta_2\left(\frac{\alpha_1}{\omega_1}(y_1-\xi_1) + \frac{\alpha_2}{\omega_2}(y_2-\xi_2)\right)$$

Using

$$c_k = \frac{2}{(2\pi)^{\frac{k}{2}}\sqrt{1+k\eta'\boldsymbol{\Omega}\eta}}$$

$$A_0(\rho, \boldsymbol{\alpha}) = E\left(\frac{1}{\Phi(U)}\right)$$

$$A_1(\boldsymbol{\Omega}, \boldsymbol{\alpha}) = E\left(\frac{U}{\Phi(U)}\mu_c\right)$$

$$A_2(\boldsymbol{\Omega}, \boldsymbol{\alpha}) = E\left(\frac{1}{\Phi(U)}(U^2\mu_c\mu_c' + \Omega_c)\right)$$

where $U \sim N(0, \bar{\alpha}^2)$,

$$\bar{\alpha}^2 = \frac{\tilde{\eta}^2}{1+2\tilde{\eta}^2},$$

$$\tilde{\eta}^2 = \eta'\boldsymbol{\Omega}\eta = \alpha_1^2 + 2\rho\alpha_1\alpha_2 + \alpha_2^2,$$

$$\mu_c = \tilde{\eta}^{-2}\Omega\eta$$

and

$$\Omega_c = \Omega - \tilde{\eta}^{-2}\Omega\eta\eta'\Omega$$

the diagonal elements of the expected Fisher information matrix are given by

$$E\left(-\frac{\partial^2}{(\partial\xi_1)^2} l(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha})\right) = \frac{1}{\omega_1^2(1-\rho^2)} + \frac{\alpha_1^2}{\omega_1^2}c_2 A_0(\rho, \boldsymbol{\alpha})$$

$$
\begin{aligned}
E\left(-\frac{\partial^2}{(\partial\omega_1)^2} l(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha})\right) =& -\frac{1}{\omega_1^2} + 3\frac{1}{\omega_1^2(1-\rho^2)} - 2\frac{\rho^2}{\omega_1^2(1-\rho^2)} \\
&+ \frac{\alpha_1^2}{\omega_1^4} A_2(\boldsymbol{\Omega}, \boldsymbol{\alpha})_{(1,1)}
\end{aligned}
$$

$$
E\left(-\frac{\partial^2}{(\partial\rho)^2} l(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha})\right) = \frac{1}{1-\rho^2} + \frac{2\rho^2}{(1-\rho^2)^2}
$$

$$
E\left(-\frac{\partial^2}{(\partial\alpha_1)^2} l(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha})\right) = \frac{c_2}{\omega_1^2} A_2(\boldsymbol{\Omega}, \boldsymbol{\alpha})_{(1,1)}
$$

The expectations A_0 and A_2 are replaced by approximations in the implementation to improve the speed of the algorithm.

5 Theory - Univariate Skew-t Distribution

In Azzalini (2013) the basis essentials of the following calculations can be found.

The log-likelihood is given by

$$
\begin{aligned}
l(x; \xi, \omega, \alpha, \nu) =& const. - \ln(\omega) - \frac{1}{2}\ln(\nu) + \ln\left(\Gamma\left(\frac{\nu+1}{2}\right)\right) - \ln\left(\Gamma\left(\frac{\nu}{2}\right)\right) \\
&- \frac{\nu+1}{2}\ln\left(1 + \frac{z^2}{\nu}\right) + \ln(T(\alpha z\tau; \nu+1)).
\end{aligned}
$$

The first derivatives are given by

$$
\frac{\partial}{\partial\xi} l(\xi, \omega, \alpha, \nu) = \frac{z\tau^2}{\omega} - \frac{\alpha\tau\nu}{\omega(z^2+\nu)}\lambda
$$

$$
\frac{\partial}{\partial\omega} l(\xi, \omega, \alpha, \nu) = -\frac{1}{\omega} + \frac{z^2\tau^2}{\omega} - \frac{\alpha z\tau\nu}{\omega(z^2+\nu)}\lambda
$$

$$
\frac{\partial}{\partial\alpha} l(\xi, \omega, \alpha, \nu) = z\tau\lambda
$$

$$
\begin{aligned}
\frac{\partial}{\partial \nu} l(\xi, \omega, \alpha, \nu) =& \frac{1}{2}\Bigg\{ \Psi\left(\frac{\nu}{2}+1\right) - \Psi\left(\frac{\nu}{2}\right) - \frac{2\nu+1}{\nu(\nu+1)} - \ln\left(1+\frac{z^2}{\nu}\right) \\
&+ \frac{z^2\tau^2}{\nu} + \frac{\alpha z(z^2-1)}{(\nu+z^2)^2\tau}\lambda + \frac{\iota}{T(\alpha z\tau|\nu+1)} \Bigg\}
\end{aligned}
$$

where

$$
\lambda = \frac{t(\alpha z\tau|\nu+1)}{T(\alpha z\tau|\nu+1)}
$$

and

$$
\iota = \int_{-\infty}^{\alpha z\tau} \left\{ \frac{(\nu+2)u^2}{(\nu+1)(\nu+1+u^2)} - \ln\left(1+\frac{u^2}{\nu+1}\right) \right\} t(u|\nu+1)du.
$$

6 Theory - Bivariate Skew-t Distribution

The log-likelihood is given by

$$
\begin{aligned}
l(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha}, \nu) =& \log(2) - \log(\nu) - \log(\pi) - \log(\omega_1) - \log(\omega_2) - \frac{1}{2}\log(1-\rho^2) \\
&+ \log\left(\Gamma\left(\frac{\nu+2}{2}\right)\right) - \log\left(\Gamma\left(\frac{\nu}{2}\right)\right) - \frac{\nu+2}{2}\log\left(1+\frac{Q}{\nu}\right) \\
&+ \log\left(T\left(\boldsymbol{\alpha}'\mathbf{z}\sqrt{\frac{\nu+2}{\nu+Q}}; \nu+2\right)\right)
\end{aligned}
$$

The first partial derivatives are given by

$$
\begin{aligned}
\frac{\partial}{\partial \xi_1} l(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha}, \nu) =& \frac{-1}{\omega_1}\Bigg\{ -\frac{\nu+2}{2}\frac{2z_1 - 2\rho z_2}{\left(1+\frac{Q}{\nu}\right)\nu\left(1-\rho^2\right)} \\
&+ \frac{t\left(\boldsymbol{\alpha}'\mathbf{z}\sqrt{\frac{\nu+2}{\nu+Q}}; \nu+2\right)}{T\left(\boldsymbol{\alpha}'\mathbf{z}\sqrt{\frac{\nu+2}{\nu+Q}}; \nu+2\right)} \\
&\cdot \left(\alpha_1\sqrt{\frac{\nu+2}{\nu+Q}} - \boldsymbol{\alpha}'\mathbf{z}\frac{\sqrt{\nu+2}\left(z_1-\rho z_2\right)}{(\nu+Q)^{1.5}}\right) \Bigg\}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial}{\partial \omega_1} l(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha}, \nu) =& -\frac{1}{\omega_1} - \frac{z_1}{\omega_1}\Bigg\{ -\frac{\nu+2}{2}\frac{2z_1 - 2\rho z_2}{\left(1+\frac{Q}{\nu}\right)\nu\left(1-\rho^2\right)} \\
&+ \frac{t\left(\boldsymbol{\alpha}'\mathbf{z}\sqrt{\frac{\nu+2}{\nu+Q}}; \nu+2\right)}{T\left(\boldsymbol{\alpha}'\mathbf{z}\sqrt{\frac{\nu+2}{\nu+Q}}; \nu+2\right)} \\
&\cdot \left(\alpha_1\sqrt{\frac{\nu+2}{\nu+Q}} - \boldsymbol{\alpha}'\mathbf{z}\frac{\sqrt{\nu+2}\,(z_1 - \rho z_2)}{(\nu+Q)^{1.5}}\right)\Bigg\}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial}{\partial \rho} l(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha}, \nu) =& -\left\{(\nu+2)\frac{1}{\nu+Q} + \frac{t\left(\boldsymbol{\alpha}'\mathbf{z}\sqrt{\frac{\nu+2}{\nu+Q}}; \nu+2\right)}{T\left(\boldsymbol{\alpha}'\mathbf{z}\sqrt{\frac{\nu+2}{\nu+Q}}; \nu+2\right)}\boldsymbol{\alpha}'\mathbf{z}\frac{\sqrt{\nu+2}}{(\nu+Q)^{\frac{3}{2}}}\right\} \\
&\cdot \left\{\frac{\rho}{(1-\rho^2)^2}\left(z_1^2 - 2\rho z_1 z_2 + z_2^2\right) - \frac{1}{1-\rho^2}z_1 z_2\right\} + \frac{\rho}{1-\rho^2}
\end{aligned}
$$

$$
\frac{\partial}{\partial \alpha_1} l(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha}, \nu) = \frac{t\left(\boldsymbol{\alpha}'\mathbf{z}\sqrt{\frac{\nu+2}{\nu+Q}}; \nu+2\right)}{T\left(\boldsymbol{\alpha}'\mathbf{z}\sqrt{\frac{\nu+2}{\nu+Q}}; \nu+2\right)} z_1\sqrt{\frac{\nu+2}{\nu+Q}}
$$

$$
\begin{aligned}
\frac{\partial}{\partial \nu} l(\boldsymbol{\xi}, \boldsymbol{\Omega}, \boldsymbol{\alpha}, \nu) =& -\frac{1}{\nu} + \frac{1}{2}\Bigg\{\psi\left(\frac{\nu+2}{2}\right) - \psi\left(\frac{\nu}{2}\right) - \log\left(1+\frac{Q}{\nu}\right) \\
&- (\nu+2)\frac{Q}{\left(1+\frac{Q}{\nu}\right)\nu^2} \\
&+ \frac{t\left(\boldsymbol{\alpha}'\mathbf{z}\sqrt{\frac{\nu+2}{\nu+Q}}; \nu+2\right)}{T\left(\boldsymbol{\alpha}'\mathbf{z}\sqrt{\frac{\nu+2}{\nu+Q}}; \nu+2\right)} \\
&\cdot \left(\frac{\boldsymbol{\alpha}'\mathbf{z}}{\sqrt{(\nu+2)(\nu+Q)}} - \frac{\boldsymbol{\alpha}'\mathbf{z}\sqrt{\nu+2}}{(\nu+Q)^{\frac{3}{2}}}\right) \\
&+ \frac{\iota}{T\left(\boldsymbol{\alpha}'\mathbf{z}\sqrt{\frac{\nu+2}{\nu+Q}}; \nu+2\right)}\Bigg\}
\end{aligned}
$$

where

$$\iota = \int_{-\infty}^{\boldsymbol{\alpha}'\mathbf{z}\sqrt{\frac{\nu+2}{\nu+Q}}} \Bigg\{ \psi\left(\frac{\nu+3}{2}\right) - \psi\left(\frac{\nu+2}{2}\right) \frac{1}{\nu+2} \frac{u^2(\nu+3)}{(\nu+2)(\nu+2+u^2)} - \log\left(1 + \frac{u^2}{\nu+2}\right) \Bigg\} t(u;\nu+2)\, du$$

The derivatives for ξ_2, ω_2 and α_2 are the same as for ξ_1, ω_1 and α_1 when 1 and 2 change places. In Azzalini (2013) the derivatives with respect to vector based parameters can be found.

The diagonal elements of the expected Fisher information matrix are evaluated numerically for a multitude of parameters combinations and stored to accelerate the algorithm.

7 MCMC Algorithm

The general structure of the algorithm is given as follows:

Algorithm

1. Find initial values for each parameter using a backfitting algorithm with all hyperparameters fixed at appropriate starting values.

2. For each MCMC iteration $t = 1, \dots, T$, complete the following steps:

   (a) For every vector of coefficients $\boldsymbol{\beta}_j^{\vartheta_k}$, $j = 1, \dots, J_k$, of every parameter ϑ_k, $k = 1, \dots, K$, of the response distribution

   - draw a new parameter from the proposal distribution q

   $$\boldsymbol{\beta}_j^{\vartheta_k *} \sim N\left(\boldsymbol{\mu}_j^{\vartheta_k}, \left(\boldsymbol{P}_j^{\vartheta_k}\right)^{-1}\right),$$

   where $\boldsymbol{\mu}_j^{\vartheta_k}$ and $\boldsymbol{P}_j^{\vartheta_k}$ are given in the main part. To ensure identifiability, the parameters are constrained to fulfill $\mathbf{A}\boldsymbol{\beta}_j^{\vartheta_k *} = \mathbf{0}$ using a suitable matrix $\mathbf{A}$ as given in Rue and Held (2005).

- Accept the new parameter with probability

$$\pi = \min\left\{1, \frac{p\left(\boldsymbol{\beta}_j^{\vartheta_k *}|.\right) q\left(\boldsymbol{\beta}_j^{\vartheta_k}, \boldsymbol{\beta}_j^{\vartheta_k *}\right)}{p\left(\boldsymbol{\beta}_j^{\vartheta_k}|.\right) q\left(\boldsymbol{\beta}_j^{\vartheta_k *}, \boldsymbol{\beta}_j^{\vartheta_k}\right)}\right\}.$$

- update the hyperparameters.

References

Arellano-Valle, R. B. and A. Azzalini (2008). The centred parametrization for the multivariate skew-normal distribution. *Journal of Multivariate Analysis 99*(7), 1362–1382.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian journal of statistics*, 171–178.

Azzalini, A. (2013). *The skew-normal and related families*, Volume 3. Cambridge University Press.

Bayes, C. L. and M. Branco (2007). Bayesian inference for the skewness parameter of the scalar skew-normal distribution. *Brazilian Journal of Probability and Statistics 21*(2), 141–163.

Rue, H. and L. Held (2005). *Gaussian Markov random fields: theory and applications*. New York/Boca Raton: Chapman & Hall/CRC.