Statistics 5401
20. Population Principal Components

Gary W. Oehlert
School of Statistics
313B Ford Hall
612-625-1557
gary@stat.umn.edu

We have been talking about principal components for data. We can also do principal components for populations. Population principal components depend on Σ, the population variance matrix.
Let the random p-vector x have mean μ and covariance Σ. Note, x need not be normally distributed.
Let $v_1, v_2, \ldots, v_p$ be p-vectors of length 1.
The first principal component is the linear combination of x with maximal variance:

$$w_1 = v_1'x \text{ with maximal } v_1'\Sigma v_1$$

Subsequent principal components are maximal variance linear combinations uncorrelated with previous principal components:

$$w_j = v_j'x \text{ with maximal } v_j'\Sigma v_j$$

$$\text{subject to } v_j'\Sigma v_k = 0 \text{ for } k < j$$

The coefficients v_i are the eigenvectors of Σ, and the (maximal) variances are the eigenvalues.
If Σ has rank p (the usual case), then all eigenvalues are greater than 0.
Thus testing a null hypothesis that the number of positive eigenvalues is less than p is not usually helpful.
Nevertheless, having several nearly zero eigenvalues implies that we can capture most of the variation in the random variable with a lower rank model.
When

$$\frac{\sum_{j=1}^m \eta_j}{\sum_{j=1}^p \eta_j} \approx 1$$

or

$$\frac{\sum_{j=m+1}^p \eta_j}{\sum_{j=1}^p \eta_j} \approx 0$$

then $\Sigma \approx \sum_{j=1}^m \eta_j v_j v_j'$.
Note that capturing most of the variation from the p-dimensions in m linear combinations does not necessarily mean that the m principal components capture the aspects of the distribution that are important to us.
Overall

$$w = \mathbf{V}'(x - \mu)$$

or

$$x = \mathbf{V}w + \mu$$

The different components w_i are uncorrelated, but the principal components are correlated with the original variables.

$$Cov(w_j, x_i) = Cov(\sum_{k=1}^p v_{jk}x_k, x_i) = \sum_{k=1}^p v_{jk}\Sigma_{ki} = \eta_j v_{ji}$$

$$Cor(w_j, x_i) = \frac{Cov(w_j, x_i)}{\sqrt{Var(w_j)Var(x_i)}} = \frac{\eta_j v_{ji}}{\sqrt{\eta_j \Sigma_{ii}}} = \frac{v_{ji}\sqrt{\eta}}{\sqrt{\Sigma_{ii}}}$$

Some special cases.

• $\Sigma = \text{diag}(\sigma)$ where the variances σ_i are all different. Then the eigenvalues are the σ_i and the original components are the eigenvectors.

• $\Sigma = \mathbf{I}_p$ All of the eigenvalues equal 1, and any set of orthonormal vectors form the eigenvectors.

• $\Sigma_{ii} = \sigma^2; \Sigma_{ij} = \sigma^2\rho$ $(-1/(p-1) \leq \rho \leq 1)$. This is the "intraclass correlation model. One eigenvalue is $\sigma^2(1+(p-1)\rho)$ with eigenvector $\mathbf{1}$. The other eigenvalues are all $\sigma^2(1-\rho)$, and the remaining eigenvectors are any set of orthonormal contrasts among the p variables.

Inference. The only relatively simple inference for the eigenvalues η_i arises when x is multivariate normal and all the eigenvalues of Σ are different.

In that case, and for large n,

$$\sqrt{n}(\hat{\eta}_i - \eta_i) \approx N(0, 2\eta_i^2)$$

and the various $\hat{\eta}_i$s are asymptotically independent.

This is neat, but it's not really obvious what to do with this inference.

True eigenvalues 5, 3, 1; sample sizes 30, 100, 300, and 1000; 10,000 random normal samples; average scaled eigenvalues

	30	100	300	1000
first	1.068	1.0181	1.006	1.0012
second	0.92009	0.97838	0.99289	0.99838
third	0.90394	0.97291	0.98981	0.99727

Average scaled variances:

	30	100	300	1000
first	0.92518	0.97063	0.97919	0.9972
second	0.78304	0.93409	0.99886	0.98772
third	0.88542	0.97593	0.98973	1.0166

p-values for testing normality of the distribution of the sample eigenvalues using rankit correlations

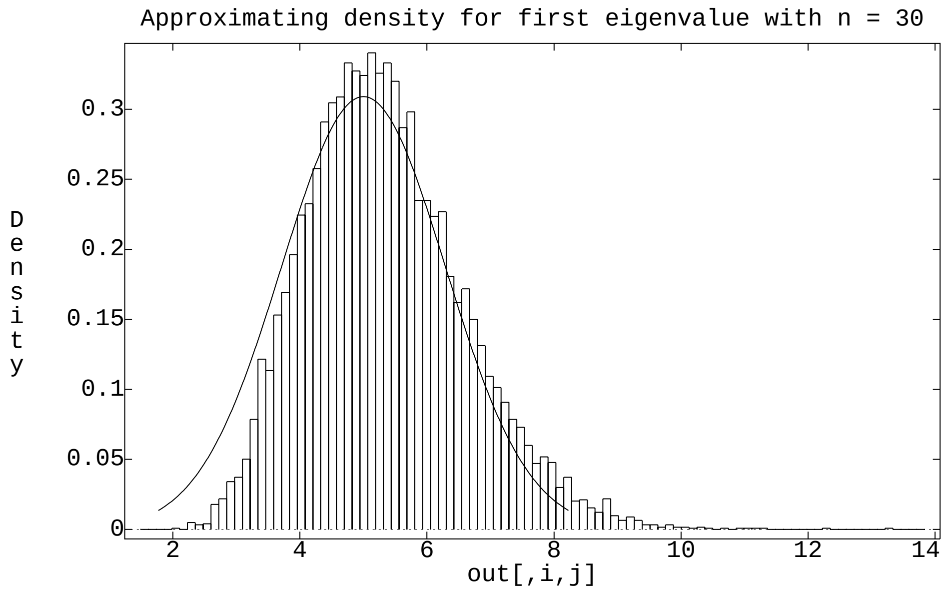
	30	100	300	1000
first	0	0	0	0.013
second	0	0	0	0.021
third	0	0	0	0.002

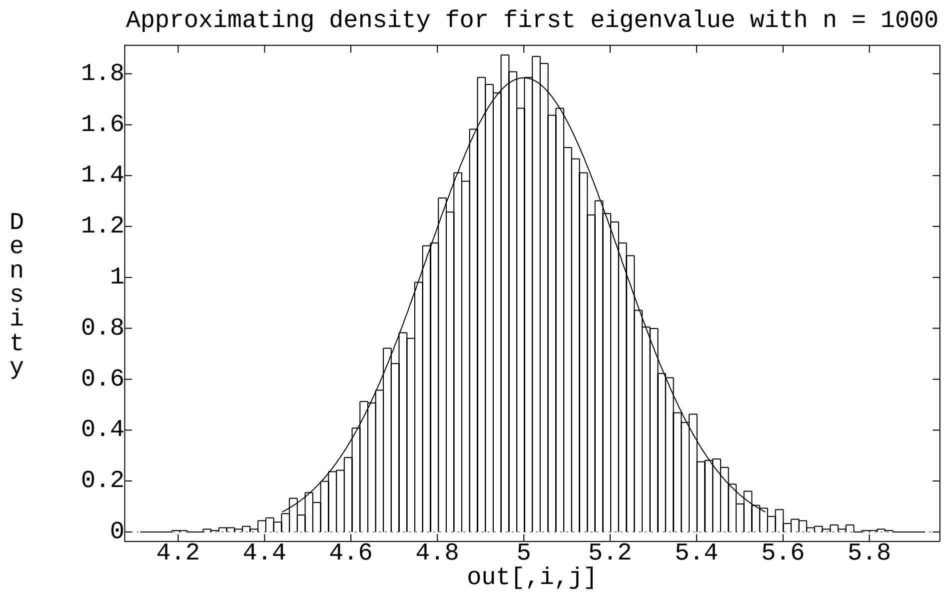
g1 skewness (mean zero, sd .024 under normal)

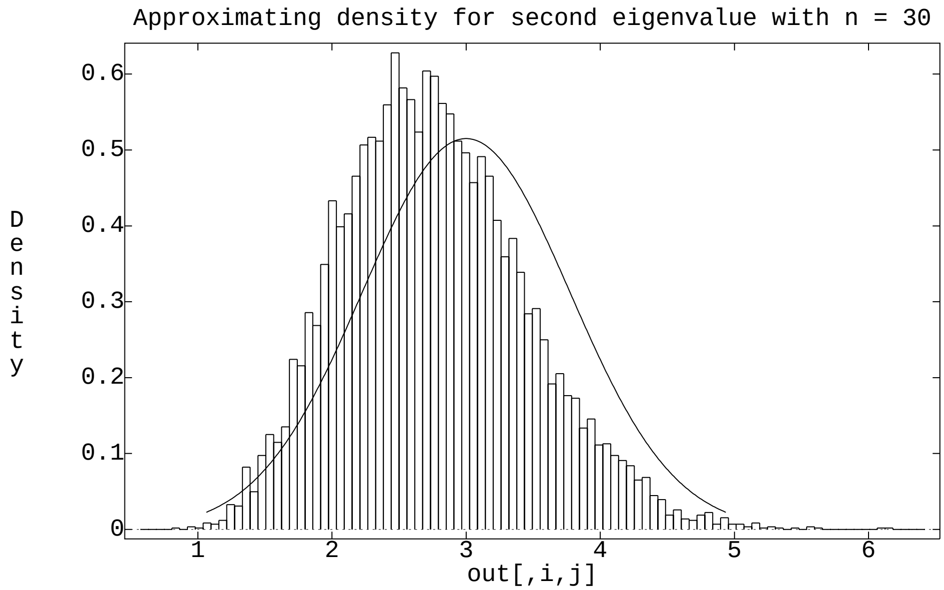
	30	100	300	1000
first	0.56793	0.29325	0.20918	0.078688
second	0.41542	0.28895	0.16387	0.066393
third	0.52642	0.32936	0.16598	0.087569

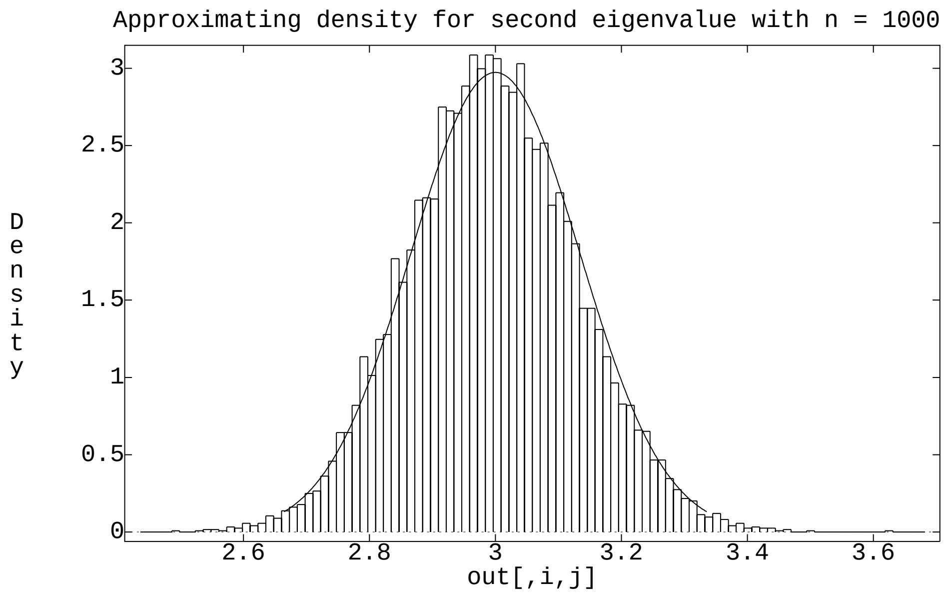
g2 kurtosis (mean zero, sd .049 under normal)

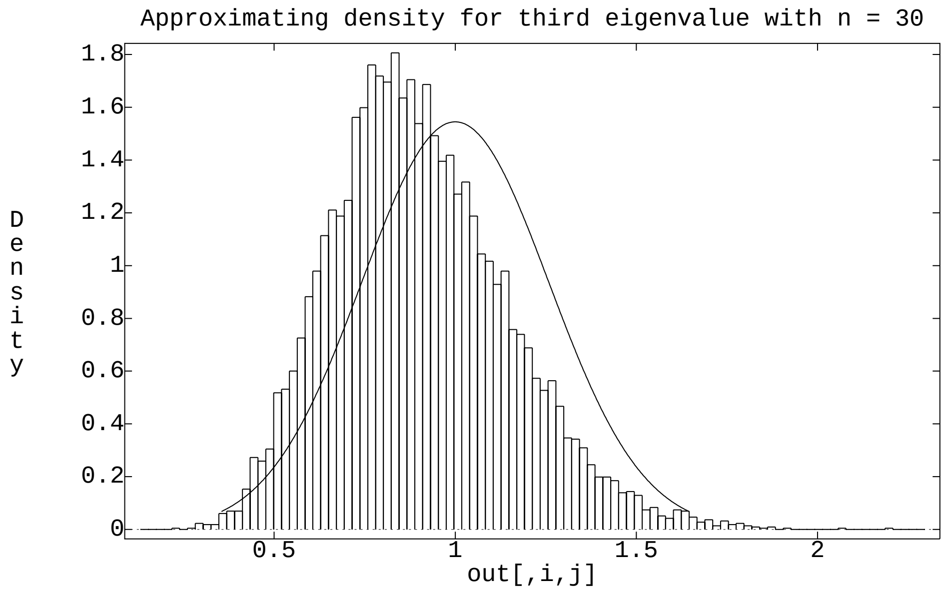
	30	100	300	1000
first	0.58399	0.11373	0.18838	0.030802
second	0.16045	0.10817	0.056552	0.061182
third	0.38673	0.14261	-0.06519	0.04203

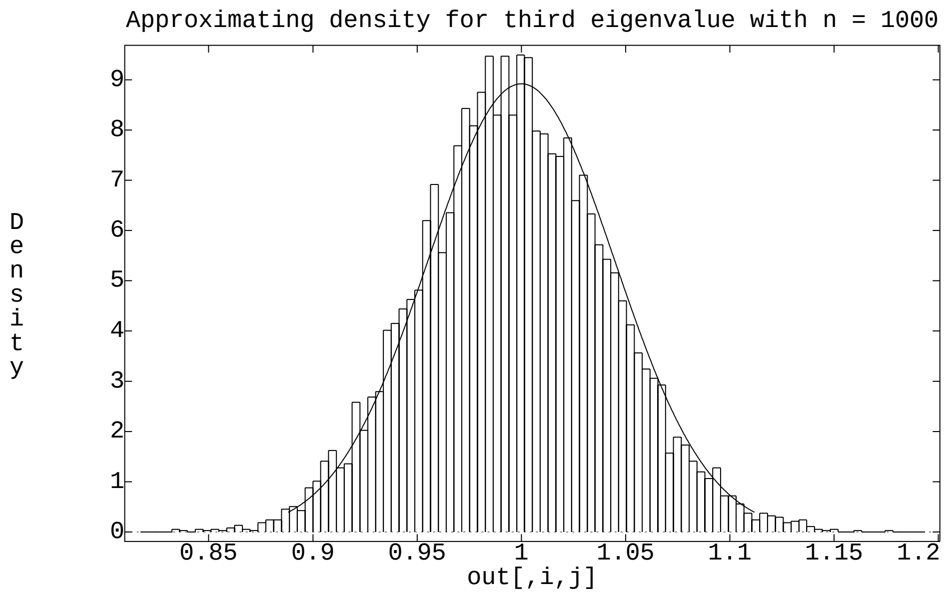

Approximating density for first eigenvalue with n = 30
Density
0.3
0.25
0.2
0.15
0.1
0.05
0
2
4
6
8
10
12
14
out[,i,j]


Approximating density for first eigenvalue with n = 1000
Density
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
4.2
4.4
4.6
4.8
5
5.2
5.4
5.6
5.8
out[,i,j]


Approximating density for second eigenvalue with n = 30
Density
0.6
0.5
0.4
0.3
0.2
0.1
0
1
2
3
4
5
6
out[,i,j]


Approximating density for second eigenvalue with n = 1000
Density
3
2.5
2
1.5
1
0.5
0
2.6
2.8
3
3.2
3.4
3.6
out[,i,j]





Just a little nonnormality in the original data messes up all these eigenvalue results rather dramatically. Here we give 5% of the data a varianace 9 times as large. Repeat the above analysis.

Average scaled means (nonnormal data)

	30	100	300	1000
first	1.0989	1.0314	1.0058	1.0027
second	0.86781	0.96156	0.99225	0.99747
third	0.87804	0.9704	0.99203	0.99802

Not too bad.

Average scaled variances:

	30	100	300	1000
first	3.1406	3.1532	3.2472	3.3557
second	1.4049	2.4456	3.1024	3.3273
third	1.986	3.0418	3.3247	3.3089

WAY off.

p-values for testing normality of the distribution of the sample eigenvalues using rankit correlations

	30	100	300	1000
first	0	0	0	0
second	0	0	0	0
third	0	0	0	0

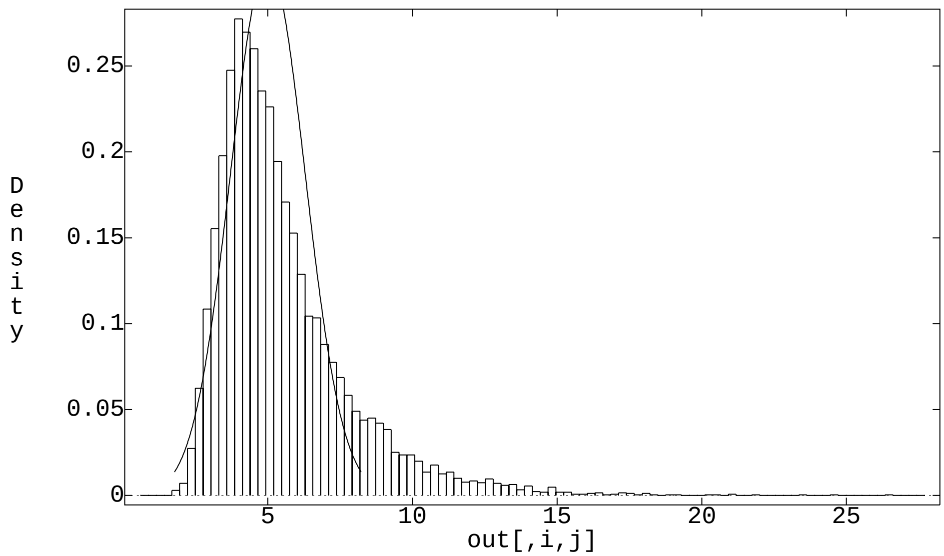
g1 skewness (mean zero, sd .024 under normal)

	30	100	300	1000
first	1.8504	1.0724	0.66204	0.3557
second	1.4151	0.81224	0.55184	0.37944
third	1.3593	1.0372	0.60801	0.35338

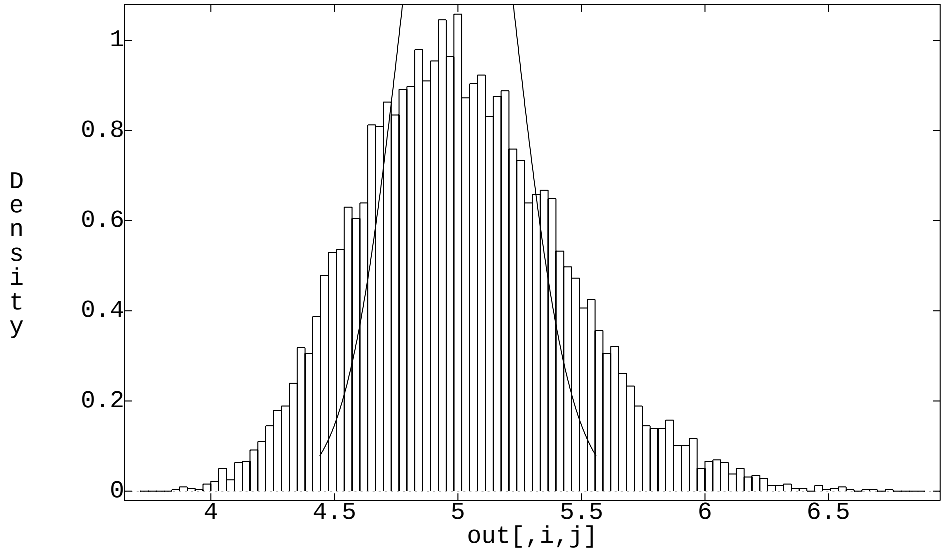
g2 kurtosis (mean zero, sd .049 under normal)

	30	100	300	1000
first	5.7039	1.6449	0.73783	0.09948
second	3.688	0.93771	0.31616	0.34205
third	2.5028	1.5674	0.50106	0.27868

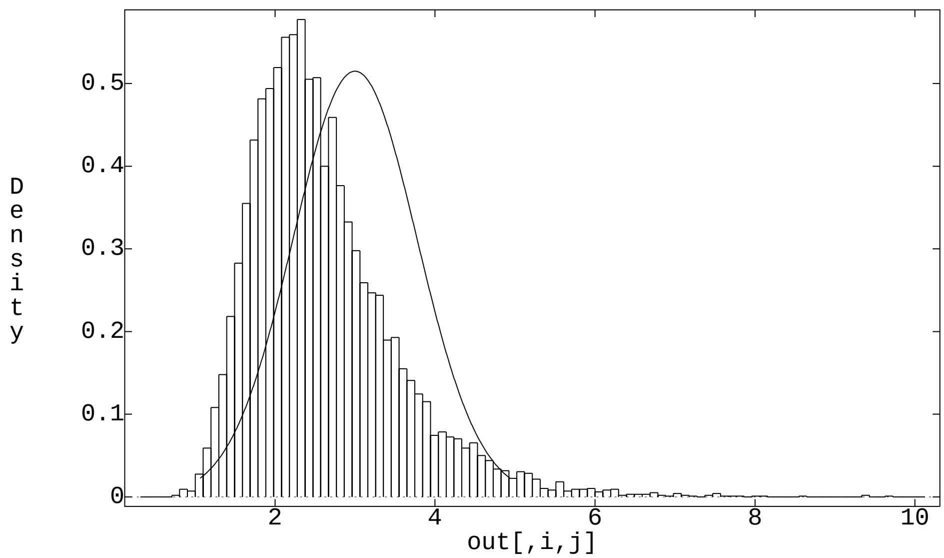
Approximating density for first eigenvalue with n = 30 (nonnormal data)



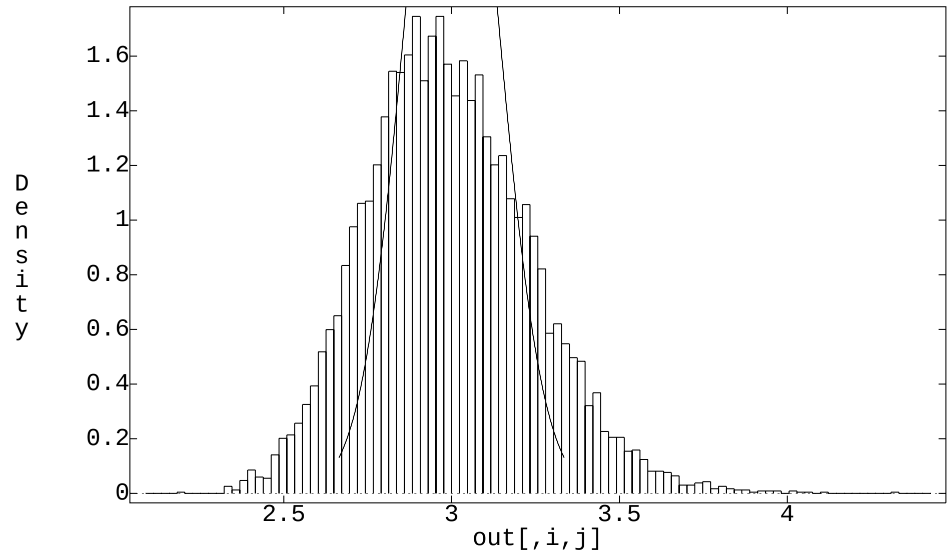
Approximating density for first eigenvalue with n = 1000 (nonnormal data)



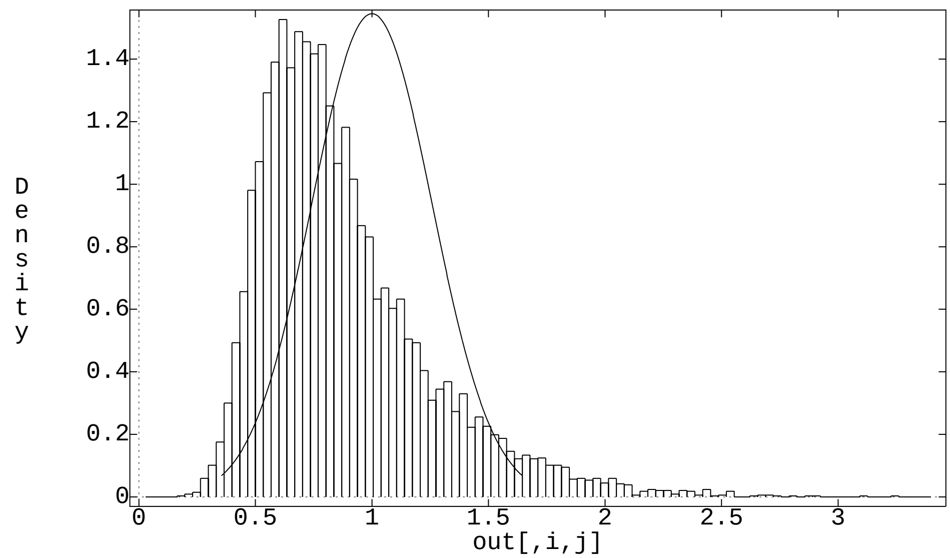
Approximating density for second eigenvalue with n = 30 (nonnormal data)



Approximating density for second eigenvalue with n = 1000 (nonnormal data)



Approximating density for third eigenvalue with n = 30 (nonnormal data)



Approximating density for third eigenvalue with n = 1000 (nonnormal data)