Getting Started with `DistFamilyFit`

Şule Şevval Karakaya Demir, Vilda Purutçuoğlu, Ahmet Bursalı

Version 0.1.1

1 Scope and input format

This package implements the methodology of [1], where the families below are compared on mobile application usage variables and Johnson transformations are used for dependence analysis.

The package fits continuous distribution families. Johnson [5] is the default; Pearson [7], Burr [2, 9] and the generalized lambda distribution (GLD) [8, 4] are selectable; background on these systems is given in [6]. Multivariate modeling currently uses Johnson marginals only. Independent fits to several columns must not be described as a joint distribution model.

A comma-separated values (CSV) or tab-separated values (TSV) input must be rectangular and contain a unique, nonempty header. Each row represents one observation and each selected numeric column represents one variable. Values in a multivariate row must refer to the same observation. ID, date and category fields may remain in the file but must be excluded through column selection. Set field separators and decimal marks explicitly. Do not put units, percentage symbols or thousands separators in numeric cells.

```
sample_id,variable_1,variable_2
s001,12.5,-1.7
s002,18.3,0.4
```

The tiny table illustrates format only. The default minimum sample size is 20; it is a software requirement, not a guarantee of reliable statistical inference. The bundled template contains 30 artificial observations:

```
> path <- system.file('extdata', 'input_template.csv', package = 'DistFamilyFit')
> prepared <- read_fit_data(path, columns = c('variable_1', 'variable_2'))
> head(prepared$data, 3)

  variable_1 variable_2
1   1.068939  0.3771947
2   1.142631  0.7183698
3   1.221403  0.9914710

> prepared$audit$n_retained

[1] 30
```

For a semicolon-delimited decimal-comma file, use `sep=';'` and `dec=','`. The reader never guesses these settings. Missing values are blank or `NA` by default. Omission is explicit and complete-case across selected columns; all removed original row numbers are retained in the audit. Zeros, negative values and outliers are not silently removed.

```
> raw <- data.frame(id = seq_len(60), value = sinh(rnorm(60)))
> clean <- prepare_fit_data(raw, columns = 'value')
> clean$audit$ignored_columns

[1] "id"
```

2 One variable and family selection

```
> x <- sinh(rnorm(120))
> u <- fit_univariate(x)
> u$ranking[, c('model', 'KS', 'status')]

      model         KS  status
1 JohnsonSU 0.03524710      OK
2 JohnsonSB 0.08192269 Warning
3 JohnsonSL 0.08192292      OK
```

The ranking uses the classical Kolmogorov–Smirnov (KS) distance. The plotting-position distance `D_CDF`, $\max_i |F(x_{(i)}) - (i - 0.5)/n|$, is reported in a separate column. MatchRate is 100 times one minus `D_CDF`; it is not classification accuracy, model probability or a goodness-of-fit p-value. All of these are descriptive in-sample measures.

```
> b <- fit_univariate(exp(x / 2), families = 'Burr')
> g <- fit_univariate(x, families = 'GLD')
> j <- fit_johnson_univariate(x, types = c('SU', 'SB'))
> comparison <- fit_univariate(x, families = c('Johnson', 'GLD'))
> if (requireNamespace('PearsonDS', quietly = TRUE)) {
+   p <- fit_univariate(x, families = 'Pearson')
+ }
```

Use `families='all'` to include every implemented family. The optional PearsonDS package supplies Pearson moment matching; the other families do not require it. Burr III/XII candidates require strictly positive data. A failed candidate is recorded rather than deleting observations for that candidate.

Johnson uses profiled likelihood and bounded numerical searches on standardized scales. No normal or lognormal fallback is performed. Reaching a numerical search boundary triggers a warning and suppresses AIC for that candidate. Johnson SL uses three identifiable parameters. GLD uses FKML quantile matching [3]; Pearson uses empirical moments with divisor n consistently. AIC is not reported for these non-likelihood estimators or for subsample estimation.

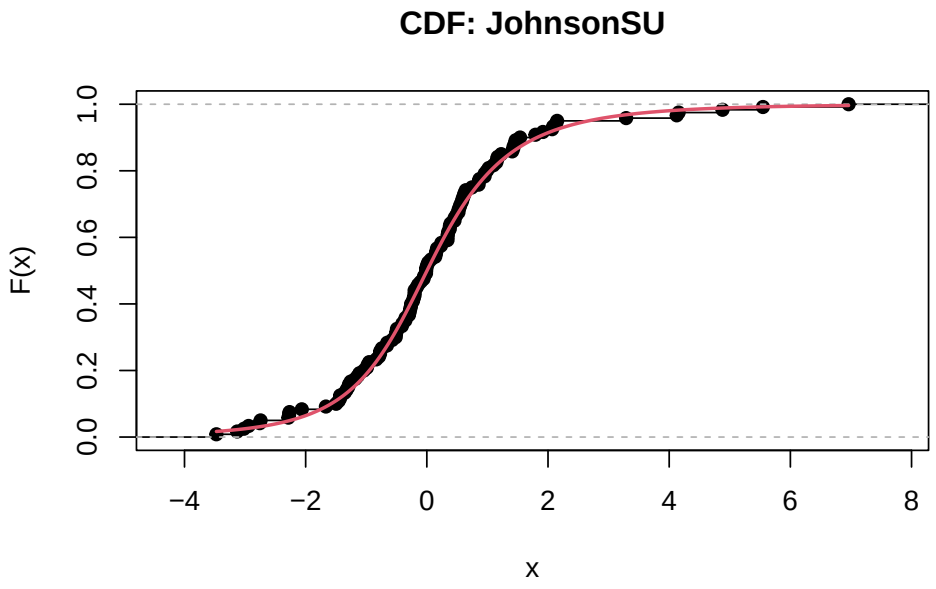
3 Distribution evaluation and separate plots

```
> cdf_fit(u, c(-1, 0, 1))

[1] 0.1891232 0.4989535 0.7906095

> quantile_fit(u, c(.1, .5, .9))

[1] -1.580260314  0.002855057  1.817877067
```

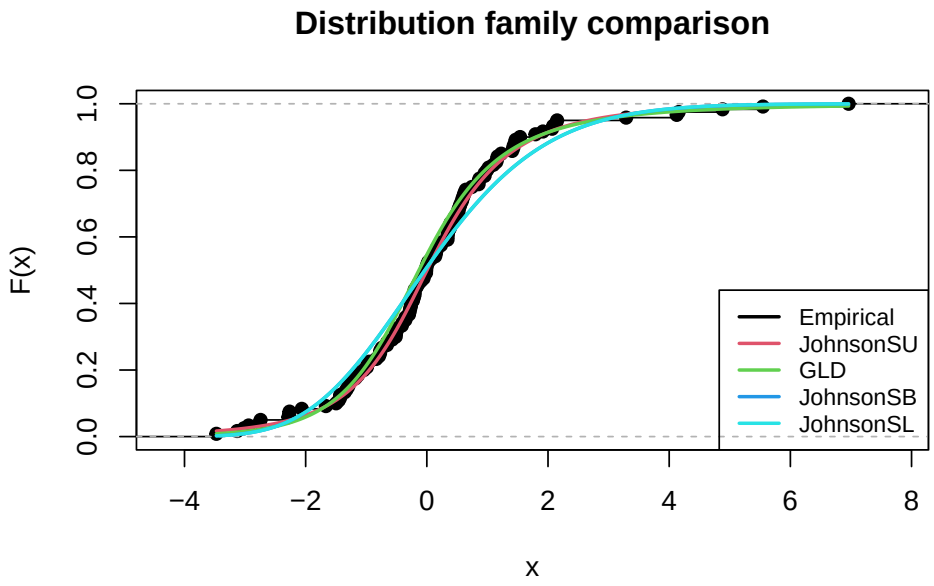
```
> density_fit(u, c(-1, 0, 1))

[1] 0.2086361 0.3666016 0.1946876

> transform_johnson(u, c(-1, 0, 1))

[1] -0.881132000 -0.002623218  0.808537759
```

Parameters are internally standardized, but evaluation functions use the original units. Fitting functions do not create plots. Call a plotting function explicitly:

```
> (function() {
+   oldpar <- par(no.readonly = TRUE)
+   on.exit({
+     par(oldpar)
+     par(mfg = oldpar$mfg, new = oldpar$new)
+   }, add = TRUE)
+ plot_fit(u, type = 'cdf')
+ })()
>
```




```
> (function() {
+   oldpar <- par(no.readonly = TRUE)
+   on.exit({
+     par(oldpar)
+     par(mfg = oldpar$mfg, new = oldpar$new)
+   }, add = TRUE)
+ plot_comparison(comparison)
+ })()
>
```



Quantile–quantile (Q–Q) and density plots are available through `plot_fit` with the corresponding type argument.

4 Johnson multivariate model

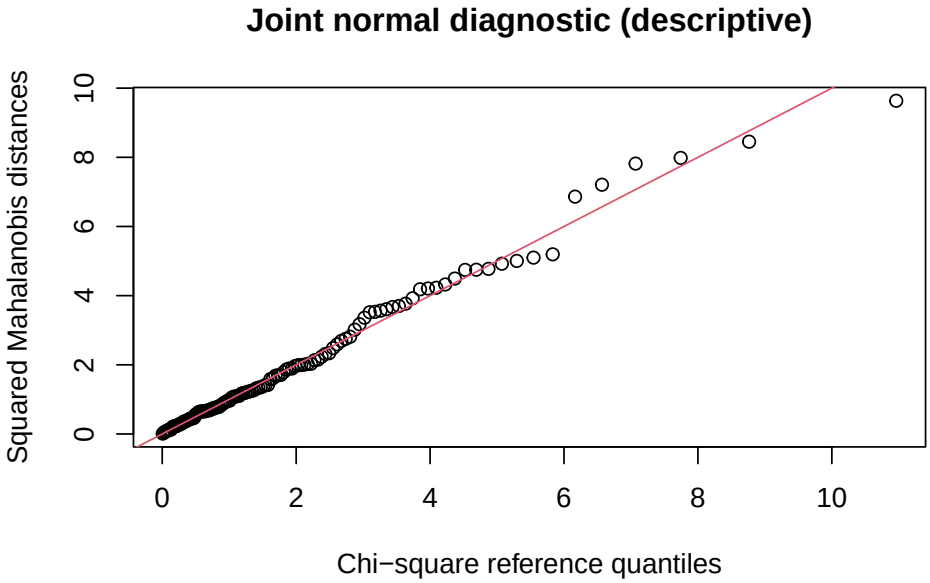
For joint fitting, rows are paired observations. Complete-case processing takes place before marginal estimation. This differs from estimating each marginal from a different subset and only later taking complete transformed rows.

```
> z <- rnorm(120)
> w <- .6 * z + sqrt(1 - .6^2) * rnorm(120)
> paired <- data.frame(first = sinh(z), second = sinh(w))
> m <- fit_johnson_multivariate(paired, types = 'SU')
> m$marginal_diagnostics[, c('model', 'KS', 'status')]

           model         KS status
first  JohnsonSU 0.04383982     OK
second JohnsonSU 0.04186191     OK

> round(m$correlation_Z, 3)

       first second
first   1.00   0.56
second  0.56   1.00
```

The covariance and correlation describe transformed Z variables, not the original measurements. Marginal normality does not imply joint normality. The sample covariance uses divisor n-1; the normal maximum likelihood estimation (MLE) covariance uses n. Singular covariance is reported and regularization must be explicitly requested through ridge. Information criteria in criteria_Z describe only the conditional Z-normal model, not the original-scale model including

estimated marginal parameters. They must not be used to compare models with different fitted transformations.

```
> (function() {
+   oldpar <- par(no.readonly = TRUE)
+   on.exit({
+     par(oldpar)
+     par(mfg = oldpar$mfg, new = oldpar$new)
+   }, add = TRUE)
+ plot_transformed(m, type = 'mahalanobis')
+ })()
>
```



The chi-square reference in this plot is descriptive because transformations and parameters were estimated. It is not a calibrated hypothesis test.

5 Diagnostics and limitations

OK means that the implemented computation checks passed; it does not establish good statistical fit. Warning candidates may be selected, so their notes must be reviewed. Unstable candidates fail numerical consistency checks and are excluded from automatic selection. Failed candidates retain their failure reason. The smallest KS among usable candidates does not prove model acceptance.

Count data and repeated point masses are only continuous approximations in this version. Explicit positive-component fitting describes a conditional distribution, not a zero-inflated model. Natural support restrictions such as [0,1] are not imposed on every family automatically. Time series and repeated measurements need additional consideration of dependence between rows. Bootstrap-calibrated acceptance tests and general multivariate Pearson, Burr or GLD models are not implemented.

6 Estimation and extreme tails

Pearson uses moment matching, while FKML GLD uses quantile matching and adaptive bisection to invert its quantile function for CDF evaluation. Johnson and Burr use numerical maximum likelihood. Log-space Burr calculations stabilize the same model. Finite precision limits extreme-tail evaluation; routine round-trip diagnostics principally cover the 1%–99% quantile range and do not certify more extreme tails. A search-boundary warning does not indicate substitution by another distribution.

7 Large samples and random numbers

The argument `max_fit_n` estimates parameters on a random subsample when the data are larger than this size. By default (`seed = NULL`) the subsample uses the current random number stream, so `set.seed()` called beforehand makes it reproducible. Supplying `seed` applies a seed locally and restores the previous random number state afterwards.

```
> big <- sinh(rnorm(2000))
> s1 <- fit_univariate(big, johnson_types = 'SU', max_fit_n = 300, seed = 1)
> s1$audit$n_estimation
```

```
[1] 300
```

References

[1] Karakaya Demir, Ş. Ş., Purutçuoğlu, V. and Bursalı, A. (2026). Modeling Mobile Application Usage with Distribution Families. Accepted at the 3rd International Conference on Advanced Technology in Electronic and Electrical Engineering (ICATEEE 2026), Hammamet, Tunisia.

[2] Burr, I. W. (1942). Cumulative frequency functions. *The Annals of Mathematical Statistics*, 13(2), 215–232.

[3] Chalabi, Y., Scott, D. J. and Wuertz, D. (2012). Flexible distribution modeling with the generalized lambda distribution. MPRA Paper No. 43333.

[4] Freimer, M., Kollia, G., Mudholkar, G. S. and Lin, C. T. (1988). A study of the generalized Tukey lambda family. *Communications in Statistics – Theory and Methods*, 17(10), 3547–3567.

[5] Johnson, N. L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika*, 36(1–2), 149–176.

[6] Johnson, N. L., Kotz, S. and Balakrishnan, N. (1994). *Continuous Univariate Distributions*, Vol. 1, 2nd ed. Wiley.

[7] Pearson, K. (1895). Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London. A*, 186, 343–414.

[8] Ramberg, J. S. and Schmeiser, B. W. (1974). An approximate method for generating asymmetric random variables. *Communications of the ACM*, 17(2), 78–82.

[9] Tadikamalla, P. R. (1980). A look at the Burr and related distributions. *International Statistical Review*, 48(3), 337–344.

8 Detailed terminology and interpretation

CDF and ECDF. A cumulative distribution function (CDF) is the model probability $P(X \leq x)$. An empirical cumulative distribution function (ECDF) is the fraction of observations at or below x. A density is different from an exact-value probability for continuous data.

KS. The Kolmogorov–Smirnov (KS) distance is the largest fitted-versus- empirical cumulative probability difference, including both sides of sample steps. It is a descriptive distance here; the package does not supply a standard fitted-parameter significance-test p-value.

MLE and NLL. Maximum likelihood estimation (MLE) maximizes model likelihood; negative log-likelihood (NLL) is minus the sum of log densities. The Johnson and Burr candidates use numerical likelihood estimation. Numerical convergence does not prove a global optimum.

AIC and BIC. The Akaike information criterion (AIC) is $2k - 2\ell$; the Bayesian information criterion (BIC) is $k \log(n) - 2\ell$, where k is the compatible model's parameter count, n is the number of observations and ℓ is its log-likelihood. Boundary, subsampling, non-likelihood and regularization restrictions are documented in the function help. Joint criteria describe only a normal model on a fixed transformed Z data set; they exclude marginal estimation and cannot compare different fitted transformations.

RMSE. Root mean square error (RMSE) is the square root of the average squared discrepancy. The `CDF_quantile_RMSE` field compares fitted cumulative probabilities at empirical quantiles with the probabilities 0.1, 0.25, 0.5, 0.75, 0.9, 0.95 and 0.99. It is on the probability scale, not in original measurement units.

SU, SL and SB. These Johnson labels mean unbounded, lognormal and bounded forms. The unbounded form extends over the real line, the lognormal form has a finite lower endpoint, and the bounded form has two finite endpoints. They describe support and transformation families, not levels of fit quality.

GLD and FKML. The generalized lambda distribution (GLD) uses the Freimer–Kollia–Mudholkar–Lin (FKML) parameterization. It is fitted by quantile matching, not by maximum likelihood. The cumulative distribution function is evaluated by numerical inversion of its quantile function.

L-BFGS-B. This is the limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm with box constraints. It numerically optimizes profiled Johnson parameters within finite search guards; those guards are not true support endpoints.

Q–Q. A quantile–quantile (Q–Q) plot compares model and observed quantiles. Extreme-point deviations can remain when central agreement is good. The squared Mahalanobis reference plot is also descriptive.

RNG. A random number generator (RNG) produces pseudorandom draws. With the default `seed = NULL`, subsampling uses and advances the current random number generator state. An explicit seed is applied locally and the previous state is restored. Reproducibility does not eliminate sampling uncertainty.

The installed glossary, available as `help("DistFamilyFit_glossary")`, also explains file formats, encoding, R-specific identifiers and remaining output labels.