Multivariate Transformations

David Ruppert

October 27, 2000

Many of the classical techniques of multivariate analysis, e.g., principle components, factor analysis, canonical correlations, and linear discriminant analysis, are based upon the assumption that the data have been sampled from a multivariate normal distribution. These techniques can be quite nonrobust towards some types of deviations from normality, for example heavy-tails or skewness. When the data deviate from normality, multivariate transformation can re-express the data as a sample closer to being multivariate normal. In principle, the methodologies developed for transformation to multivariate normality could be used to transform to other multivariate distributions. However, the normal distribution plays such a central role in multivariate analysis that transformation to other multivariate targets has not been considered as important and will not be discussed here.

The multivariate normal distribution has a number of important properties. Let $\boldsymbol{X} = (X_1, \ldots, X_p)$ have a multivariate normal distribution. Then:

1. All marginal distributions, in particular, the univariate marginal distributions of X_1, $\ldots$, X_p, are normal.

2. The expectation of any subvector $\boldsymbol{X}_1$, given any other subvector $\boldsymbol{X}_2$, is linear. More precisely,

$$E(\boldsymbol{X}_1|\boldsymbol{X}_2) = E(\boldsymbol{X}_1) + \mathrm{Cov}(\boldsymbol{X}_1, \boldsymbol{X}_2)\mathrm{Var}(\boldsymbol{X}_2)^{-1}(\boldsymbol{X}_2 - E(\boldsymbol{X}_2)).$$

   where $\mathrm{Var}(\boldsymbol{X}_2)$ is the variance-covariance matrix of $\boldsymbol{X}_2$ and $\mathrm{Cov}(\boldsymbol{X}_1, \boldsymbol{X}_2)$ the matrix of covariances between $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$.

3. If $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are as above then the conditional variance matrix of $\boldsymbol{X}_1$ given $\boldsymbol{X}_2$ is a constant matrix, i.e., it does not depend on $\boldsymbol{X}_2$. This property is called conditional homoscedasticity.

Multivariate transformation aims to induce all three properties, but not all data sets can be transformed exactly to multivariate normality and some comprise between these properties may be necessary.

Multivariate transformation is accomplished by applying a possibily different univariate transformation to each of the components of the multivariate data. Therefore, it is best to start with univariate transformations. The most commonly used univariate transformation family is the Box-Cox (1964) power transformation family:

$$\begin{aligned} y^{(\lambda)} &= \frac{y^\lambda - 1}{\lambda}, \quad \lambda \neq 0; \\ &= \log(y), \quad \lambda = 0. \end{aligned}$$

This family is only appropriate for positive data. For data that are possibly negative, one can first transform them to positive values by adding a constant and then apply a Box-Cox transformation. Alternatively, Manly (1976) suggests applying a Box-Cox transformation to the exponentiated data. Thus, Manly's transformations is

$$\begin{aligned} h(y;\lambda) &= \frac{\exp(y\lambda) - 1}{\lambda}, \quad \lambda \neq 0; \\ &= y, \qquad\qquad\quad \lambda = 0. \end{aligned}$$

The parameters of a univariate transformation can be selected by trial and error to find parameter values giving, say, good normal probability plots. Alternatively, Box and Cox (1964) suggest that a power transformation can be selected by maximum likelihood or Bayesian estimation. Suppose we have data $y_1, \ldots, y_n$. Box and Cox assume that for some value of λ, $y_i^{(\lambda)}$ is normally distributed with mean μ and variance σ^2. All three parameters, λ, μ, and σ^2, can be estimated by maximum likelihood. Box and Cox show that the MLE of λ maximizes the profile likelihood

$$-\frac{\log\{\widehat{\sigma}^2(\lambda)\}}{n} + (\lambda - 1)\sum_{i=1}^{n} \log(y_i).$$

Here $\widehat{\sigma}^2(\lambda)$ and $\widehat{\mu}(\lambda)$ are the MLEs of σ^2 and μ when the value of λ is known. This estimation procedure can be applied to the exponentiated data to estimate the parameter in Manly's transformation family.

Now suppose that $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$ is a sample from a p-dimensional distribution. To transform the $\boldsymbol{Y}_i$ to multivariate normality we apply a possible different transforamtion to each coordinate. For example, to apply power transformations, we must select the appropriate power for each coordinate. If there is a transformation to multivariate normality, then this transformation also induces marginal normality. Therefore a transformation can be

selected by applying univariate maximum likelihood separately to each coordinate. However, using univariate maximum likelihood is somewhat inefficient. It only uses information about the marginal distributions, not, say, properties 2. and 3. of multivariate normal distributions. Andrews, Gnanadesikan, and Warner (1971) assume that there are power transformations of the coordinates of the data that transforms them to multivariate normality. Let $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)$ be the parameters of this multivariate transformation. Andrews et al. derive the likelihood for this model. They find that the MLE of $\boldsymbol{\lambda}$ maximizes the profile likelihood

$$-\frac{\log|\widehat{\boldsymbol{\Sigma}}(\boldsymbol{\lambda})|}{n} + \sum_{j=1}^{p}\left\{(\lambda_j - 1)\sum_{i=1}^{n}\log(Y_{ij})\right\}.$$

Here $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{ip})^\mathsf{T}$ and $\widehat{\boldsymbol{\Sigma}}(\boldsymbol{\lambda})$ is the MLE of the variance-covariance matrix whene $\boldsymbol{\lambda}$ is known.

To demonstrate the greater efficiency of multivariate maximum likelihood relative to marginal maximum likelihood applied separately to the components of $\boldsymbol{Y}$, a small simulation was performed. Sample of 50 six-dimensional vectors was generated so that the logarithms of these vectors were normal with a common mean of 0 and variance of 1 and with all pairwise correlations equal to 0.8. Thus, $\lambda_j = 0$ for $j = 1, \ldots, 6$. The $\widehat{\lambda}_j$ are identically distributed though correlated. Using 500 simulated samples, it was found that the mean squared errors of the marginal likelihood estimators of the λ_j was 0.0156 while the mean squared errors of the multivariate likelihood estimators was only 0.0092; therefore the relative efficiency of multivariate to marginal likelihood was 170%.

Despite the higher efficiency of multivariate maximum likelihood, in practice multivariate and marginal MLEs are usually quite similar; this is seen in the example below. Therefore, univariate maximum likelihood is often used for computational expediency. It is rather easy to maximize the univariate likelihood by grid-search or bisection, but to maximize the multivariate likelihood efficiently a function maximizing routine is needed—the univariate MLEs can be used as starting values for this routine.

As an example, we use data on `wages` (wages in dollars per hours) and `exper` (years of work experience). There are 534 observations taken from the Current Population Survey of 1985. The data are available from Statlib (http://lib.stat.cmu.edu/datasets/). The original source is Berndt (1991).

Figure 1 illustrates the untransformed data. Panels (a) and (b) are boxplots of `wages` and `exper`. It is evident that both variates are right-skewed with `wages` the more skewed

of the two. Panel (c) is a plot of `wages` versus `exper` with a scatterplot smooth added. Clearly, the conditional expectation of `wages` given `exper` is not linear nor even monotonic. Panel (d) shows the absolute residuals from the fit in panel (c). This plot is used to diagnose heteroscedasticity; see Carroll and Ruppert (1988). The absolute residuals tend to be larger in the middle of the range of `exper` indicating that the conditional variance of `wages` given `exper` is greatest in this region. However, this heteroscedasticity is rather mild. Panels (e) and (f) are the same as (c) and (d) except that `exper` and `wages` have been interchanged. Nonlinearity of the conditional expectation and mild heteroscedasticity are again evident.

The multivariate power transformation was estimated by maximum likelihood and the MLEs were $\widehat{\lambda}_1 = 0.40$ and $\widehat{\lambda}_2 = -0.07$. The univariate MLE of λ_2 was identical to the multivariate MLE while the univariate MLE of λ_1 was 0.43, quite similar to the multivariate MLE. Figure 2 shows the same plots as Figure 1 except now for the transformed data. One can see from the boxplots that the data are now much less skewed than before transformation. Also, there is little evidence of heteroscedasticity in the plots of absolute residuals in panels (d) and (f). However, the conditional expectations that are estimated by the fits in panels (c) and (d) still appear somewhat nonlinear.

As discussed in Section 4.2 of Carroll and Ruppert (1988), monotonic transformations can be very effective for removing skewness. They also can remove heteroscedasticity of the type where the conditional variance is a monotonic function of the conditonal mean. Finally, transformation can linearize a monotonic conditional expectation. However, transformation cannot in general convert a non-monotonic conditional expectation into a monotonic, e.g., linear, one. Therefore, it is not surprising that in this example, transformation removes skewness but cannot linearize the non-monotonic conditional expectations. Moreover, switching to another transformation family would not solve this problems. There is a limit to what transformation can achieve. However, the plots, Figures 1(c) and 2(c), of `exper` versus `wages` show that the transformed data are much closer to being elliptically contoured, as required by multivariate normality, than the original data.

The estimation of transformations using data containing possible outliers requires some care. In some situations, outlying data are "accommodated' by the transformation, that is, they are not outlying on the transformed scale. For example, lognormal data often have extreme outliers in the right tail but these data points are usually not outlying after a log transformation. On the other hand, data points that are not outlying before a transforma-

tion, may be outlying after transformation. For example, observations very close to zero may be highly negative after a log transformation. Atkinson (1995) is an excellent reference on diagnostics for univariate transformations, and Velilla (1995) discusses diagnostics and robust estimators for multivariate transformations. Atkinson and Riani (1997) propose a highly robust method for detecting influential observations when estimating multivariate transformations and illustrate their method with an interesting example.

References

Andrews, D. F., Gnanadesikan, R., and Warner, J. L. (1971). Transformations of multivariate data, *Biometrics*, 27, 825–840.

Atkinson, A. C. (1985). *Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis.* Oxford University Press, New York.

Atkinson, A. C., and Riani, M. (1977). Bivariate boxplots, multiple outliers, multivariate transformations and discriminant analysis: the 1997 Hunter lecture. *Environmetrics,* 8, 583–602.

Berndt, E. R. (1991). *The Practice of Econometrics.* Addison-Wesley, NY.

Box, G. E. P., and Cox, D. R. (1964). An analysis of transformations (with discussion). *J. Roy. Statist. Soc., Ser. B.* 2, 211–46.

Carroll, R. J., and Ruppert, D. (1988). *Transformation and Weighting in Regression*, Chapman and Hall, New York and London.

Manly, B. F. J. (1976). Exponential data transformations. *Statistician.* 25, 37–42.

Velilla, S. (1995). Diagnostics and robust estimation in multivariate data transformations. *J. of the Amer. Statist. Assoc.*, 90, 945–951.

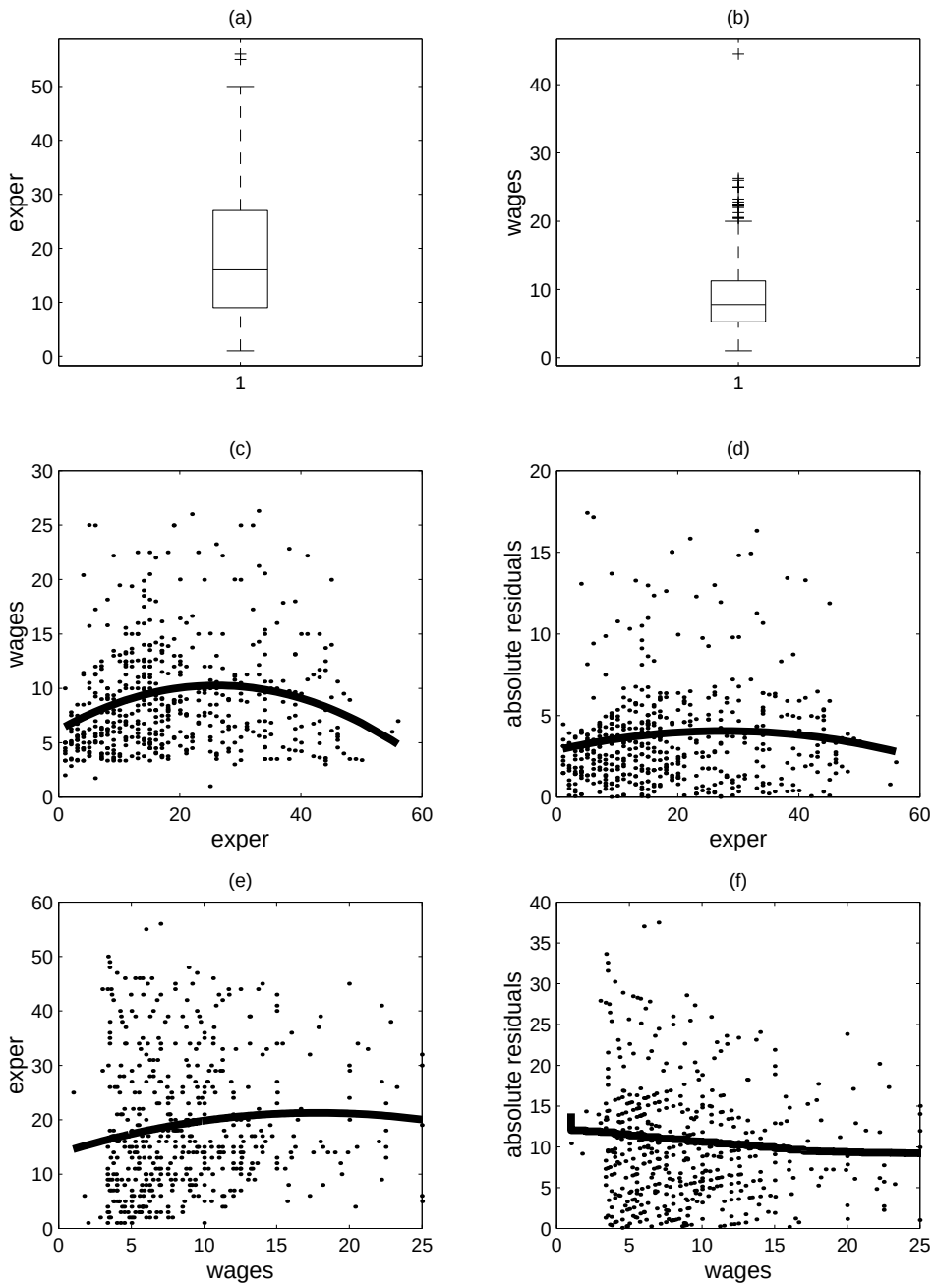

Figure 1: *Current Population Survey untransformed data. (a) Boxplot of* `exper`*. (b) Boxplot of* `wages`*. (c) Scatterplot of* `wages` *and* `exper` *with a spline smooth. (d) Scatterplot of absolute residuals from the fit in (c) versus* `exper` *with a spline smooth. (e) and (f) are the same as (c) and (d) but with variates interchanges.*

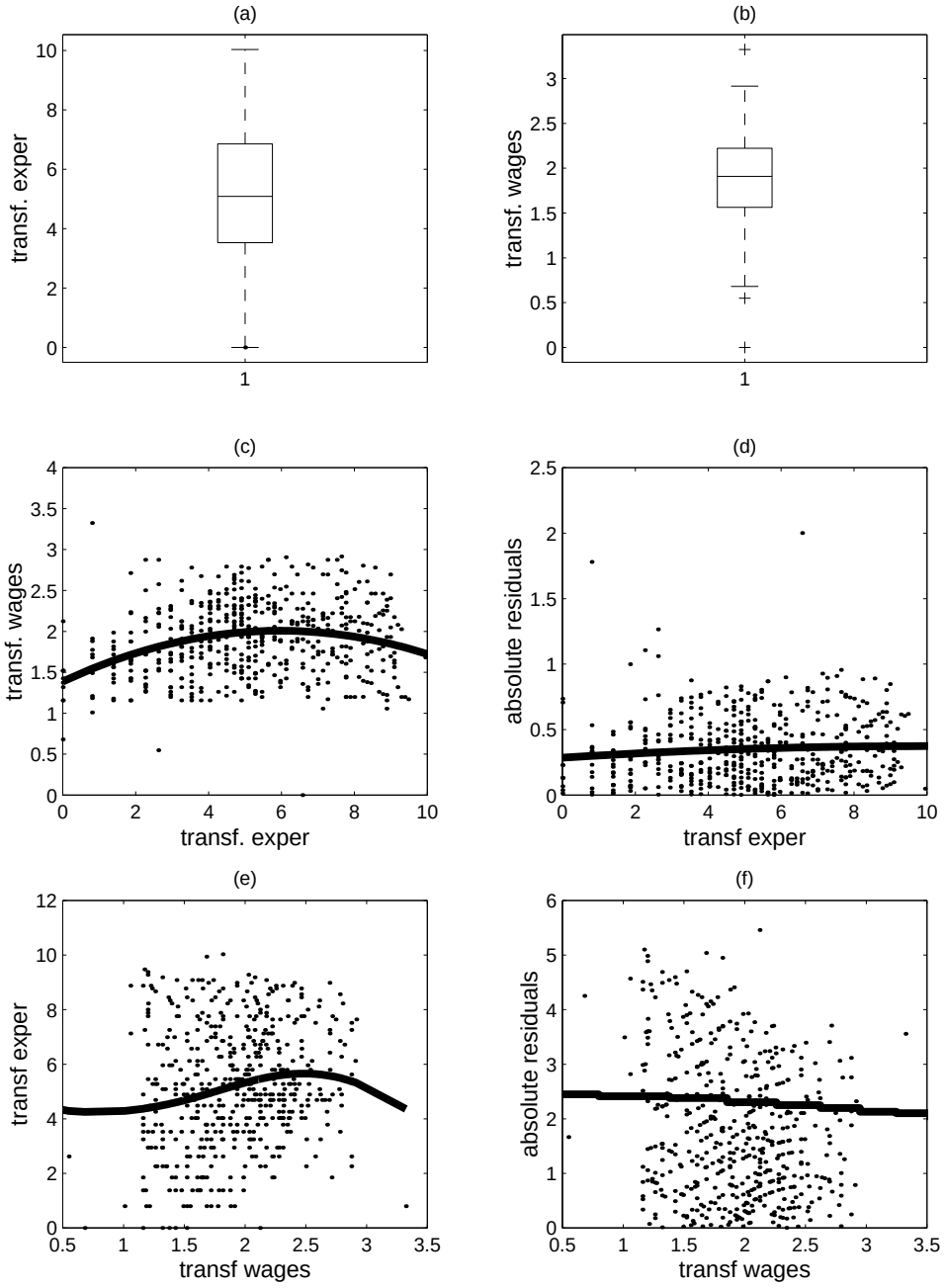

Figure 2: *Current Population Survey transformed data. (a)–(e) are as in Figure 1.*