An Introduction to Probabilistic Graphical Models

Michael I. Jordan
University of California, Berkeley

June 30, 2003

Chapter 14

Factor Analysis

In this chapter we present a latent variable model in which the latent variable is a continuous random vector. The problem that we address is one of density estimation, although the ideas that we describe can be exploited in regression and classification settings well.

In many density estimation problems, the measured data vector may be high-dimensional, but we may have reason to believe that the data lie near a lower-dimensional manifold. In such a setting it may be useful to model the data generation process as a two-stage process, in which (1) a point in the manifold is generated according to a (simple) probability density, and (2) the observed data are generated conditionally from another (simple) density that is centered on the point. The coordinates of this point form the components of a latent random vector. Assuming that we wish to parameterize a continuous manifold, the latent variable is a continuous-valued random vector.

When we assume that the manifold is a linear subspace, we obtain a model known as *factor analysis*. Figure 14.3 shows the geometry underlying the factor analysis model. The observed data are assumed to lie near a p-dimensional subspace $\mathcal{M}$ in $\Re^q$, where $p < q$. Given a set of basis vectors $\{\lambda_j\}$, a point in $\mathcal{M}$ can be represented as the product Λx, where x is a coordinate vector and Λ is the matrix whose columns are the basis vectors $\{\lambda_j\}$. We treat the coordinate vectors as values of a continuous random vector X, endowing X with a probability density. Specifically, in the case of factor analysis, we assume that X is a Gaussian random vector. Finally, given a point in $\mathcal{M}$, the observed data Y are assumed to be generated according to a Gaussian distribution centered around that point. We can view the resulting density as the convolution of the Gaussian density on the manifold with a Gaussian distribution extending into $\Re^q$, resulting in a "thick subspace" lying in $\Re^q$.

From the point of view of generating the "thick subspace," the basis vectors for $\mathcal{M}$ are not unique. Indeed, as we will see in Section 14.1, any orthogonal transformation of the basis vectors leaves the likelihood invariant. Historically, factor analysis arose in a setting in which the goal was often that of recovering unique basis vectors—these are the "factors" that give the model its name. The lack of identifiability of the basis vectors is a problem from this point of view, and factor analysis has often been seen as "controversial." From the point of view of density estimation, however, the interest is in the subspace and not any particular basis for describing the subspace. In this setting, factor analysis should not be viewed as any more "controversial" than any of the

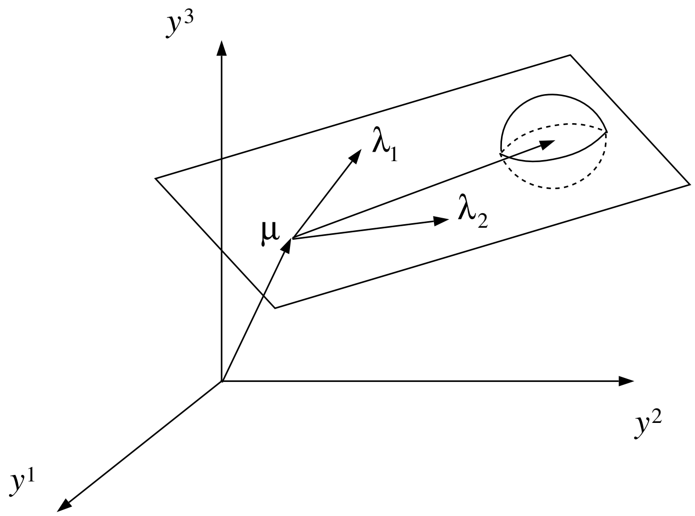

Figure 14.1: The geometry of the factor analysis model.

other models that we have described.

Factor analysis is closely related to *principal component analysis (PCA)*—another linear method for dimensionality reduction. Indeed, one can obtain PCA as a limiting case of factor analysis, much as one obtains the K-means algorithm as a limiting case of the Gaussian mixture model. We discuss PCA and some of the links between PCA and factor analysis in Section 14.4.

14.1 The model

The factor analysis model is shown as a graphical model in Figure 14.2. The model is comprised of a latent Gaussian variable X and an observable variable Y, where X is a p-dimensional random vector and Y a q-dimensional random vector, and where we assume $p < q$.

The model is parameterized as follows. Let X have a marginal Gaussian distribution:

$$X \sim \mathcal{N}(0, I), \tag{14.1}$$

with zero mean and an identity covariance matrix. Let the conditional distribution of Y be Gaussian, with mean $\mu + \Lambda x$:

$$Y \sim \mathcal{N}(\mu + \Lambda x, \Psi), \tag{14.2}$$

where Ψ is a diagonal covariance matrix.

The product of Gaussian distributions is Gaussian, and thus the joint distribution of X and Y is Gaussian, as are all marginals and conditionals computed under this joint. In particular, let us calculate the marginal of Y and the conditional distribution of X given Y.

We first calculate the marginal probability of Y, doing the calculation in two ways. The first approach is based on expressing Y as a sum:

$$Y = \mu + \Lambda x + W, \tag{14.3}$$

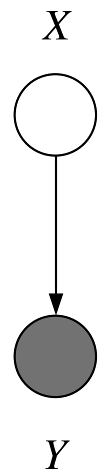

Figure 14.2: The factor analysis model as a graphical model.

where W is distributed as $\mathcal{N}(0, \Psi)$, and is independent of X. It is easy to verify that this representation yields the conditional distribution for Y in Eq. (14.2). We now make use of this representation to calculate the unconditional mean of Y:

$$\begin{aligned} E(Y) &= E(\mu + \Lambda X + W) && (14.4) \\ &= \mu + \Lambda EX + EW && (14.5) \\ &= \mu, && (14.6) \end{aligned}$$

and the unconditional covariance matrix of Y:

$$\begin{aligned} \mathrm{Var}(Y) &= E[(\mu + \Lambda X + W - \mu)(\mu + \Lambda X + W - \mu)^T] && (14.7) \\ &= E[(\Lambda X + W)(\Lambda X + W)^T] && (14.8) \\ &= \Lambda E(XX^T)\Lambda^T + E(WW^T) && (14.9) \\ &= \Lambda\Lambda^T + \Psi. && (14.10) \end{aligned}$$

Given that Y is Gaussian these two results determine the marginal distribution of Y.

There is another way to obtain these results that is based on the relationship between the conditional mean and covariance of a random vector Y and the unconditional mean and covariance. This approach will be particularly useful in Chapter 15. Recall that the conditional expectation, $E(Y \mid X)$, is a random variable, and thus it is meaningful to compute means and variances of $E(Y \mid X)$. In particular, in Appendix XXX we derive the following relationship:

$$E(Y) = E(E(Y \mid X)), \tag{14.11}$$

a result known as the *iterated expectation theorem*, and an analogous result for the variance:

$$\mathrm{Var}(Y) = \mathrm{Var}(E(Y \mid X)) + E(\mathrm{Var}(Y \mid X)). \tag{14.12}$$

Let us now make use of these relationships to derive the unconditional distribution of Y for factor analysis. Using Eq. (14.11), we have:

$$E(Y) = E(\mu + \Lambda X) = \mu. \tag{14.13}$$

and from Eq. (14.12), we obtain:

$$\begin{aligned} \mathrm{Var}(Y) &= \mathrm{Var}(\mu + \Lambda X) + E\Psi & (14.14)\\ &= E[(\Lambda X)(\Lambda X)^T] + \Psi & (14.15)\\ &= \Lambda\Lambda^T + \Psi; & (14.16) \end{aligned}$$

the same results as before.

We also need to calculate the covariance of X and Y, which we can compute using Eq. (14.3):

$$\begin{aligned} \mathrm{Cov}(X, Y) &= E[X(\mu + \Lambda X + W - \mu)^T] & (14.17)\\ &= E[X(\Lambda X + W)^T] & (14.18)\\ &= \Lambda^T. & (14.19) \end{aligned}$$

We can also obtain this expression using the following result from Appendix XXX:

$$\mathrm{Cov}(X, Y) = \mathrm{Cov}(X, E(Y \mid X)), \tag{14.20}$$

which yields:

$$\mathrm{Cov}(X, Y) = \mathrm{Cov}(X, \mu + \Lambda X) = \Lambda^T \tag{14.21}$$

as before.

Collecting together the results thus far, we have shown that the joint distribution of X and Y is a Gaussian with mean vector $(0, \mu^T)^T$ and covariance matrix:

$$\Sigma = \begin{bmatrix} I & \Lambda^T \\ \Lambda & \Lambda\Lambda^T + \Psi \end{bmatrix}. \tag{14.22}$$

We can now calculate the conditional distribution of X given Y, using the results from Chapter 13. We begin by calculating the conditional mean of X:

$$E(X \mid y) = \Lambda^T(\Lambda\Lambda^T + \Psi)^{-1}(y - \mu). \tag{14.23}$$

Note that the matrix that must be inverted in this calculation is a $(q \times q)$-dimensional matrix. It is also possible to exploit Eq. (**??**) and invert a $(p \times p)$-dimensional matrix instead:

$$E(X \mid y) = (I + \Lambda^T\Psi^{-1}\Lambda)^{-1}\Lambda^T\Psi^{-1}(y - \mu). \tag{14.24}$$

In the context of factor analysis, in which $p < q$, Eq. (14.24) is the preferred way to compute the conditional expectation.

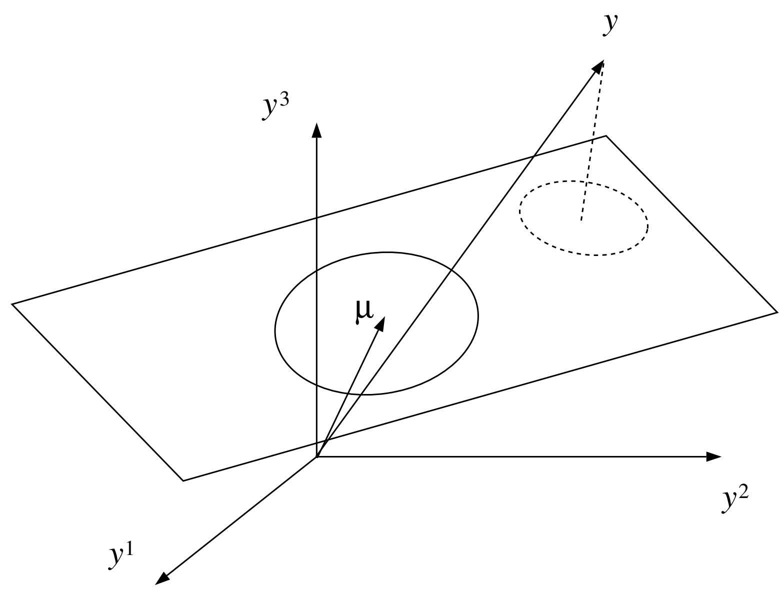

Figure 14.3: The solid ellipse corresponds to the Gaussian distribution of the latent variable X prior to the observation of Y. After $Y = y$ is observed, the distribution of X is depicted as a dotted ellipse. The mean of the updated distribution given by Eq. (14.24) and the covariance is given by Eq. (14.26).

We also compute the conditional variance of X:

$$\begin{aligned} \mathrm{Var}(X \mid y) &= I - \Lambda^T(\Lambda\Lambda^T + \Psi)^{-1}\Lambda && (14.25) \\ &= (I + \Lambda^T\Psi^{-1}\Lambda)^{-1}, && (14.26) \end{aligned}$$

where we have used Eq. (**??**) in the second step. Again, for dimensionality reasons, we usually prefer to implement Eq. (14.26).

Figure **??** summarizes our results from a geometric point of view. Before observing Y, the distribution of X is a Gaussian centered around the origin of the latent variable subspace. After an observation $Y = y$, we obtain an updated distribution for X, where Eq. (14.24) determines the mean and Eq. (14.26) determines the updated covariance matrix of the updated distribution. We in essence "project" y onto the latent subspace, obtaining not only a point projection, but an estimate of uncertainty as well.

In summary, we have obtained the probability distribution of the latent variable given the observed variable—the analog of the calculation of the posterior probability τ in Chapter **??**.

14.2 Maximum likelihood estimation

We now turn to the problem of finding maximum likelihood estimates of the parameters of the factor analysis model.

14.2.1 The log likelihood

As in the case of the mixture model, the likelihood for the factor analysis model is a marginal probability. We have a seeming advantage in the case of the factor analysis, however, because the marginal probability can be calculated analytically. Indeed, as we have seen, the marginal probability of Y is a Gaussian with mean μ and covariance matrix $\Lambda\Lambda^T + \Psi$. The log likelihood is therefore a Gaussian log likelihood:

$$l(\theta \mid D) = -\frac{N}{2}\log|\Lambda\Lambda^T + \Psi| - \frac{1}{2}\left\{\sum_n (y_n - \mu)^T(\Lambda\Lambda^T + \Psi)^{-1}(y_n - \mu)\right\}, \qquad (14.27)$$

where as usual we assume an IID data set $\mathcal{D} = \{y_n : n = 1, \ldots, N\}$.

We noted earlier that the likelihood remains invariant to orthogonal transformations of Λ. To show this, let R be an orthogonal matrix, and let $\tilde{\Lambda} = \Lambda R$. We have:

$$\tilde{\Lambda}\tilde{\Lambda}^T = \Lambda R(\Lambda R)^T = \Lambda R R^T \Lambda^T = \Lambda\Lambda^T. \qquad (14.28)$$

Thus the likelihood, which depends on Λ only through the product $\Lambda\Lambda^T$ is not changed if Λ is postmultiplied by R.

One useful consequence of Eq. (14.27) is an analytical formula for estimating the mean μ. Indeed, from the point of view of estimating μ, Eq. (14.27) is simply a Gaussian log likelihood, and we obtain the usual maximum likelihood estimate:

$$\hat{\mu}_{ML} = \frac{1}{N}\sum_n y_n, \qquad (14.29)$$

by differentiating with respect to μ and setting to zero.

In the remainder of this section, we will omit reference to the mean μ in order to simplify our notation. In practice we estimate μ according to Eq. (14.29) and then subtract the estimate from the data vectors. The resulting centered variables play the role of the data vectors y_n in the remainder of our discussion.

Unfortunately further progress in estimating the parameters is stymied—the parameters Λ and Ψ are coupled in Eq. (14.27), both by the determinant and the matrix inverse. There are no closed form expressions for the maxima of the log likelihood with respect to these parameters.

To decouple the parameters and obtain a simple algorithm for maximum likelihood estimation in factor analysis, we again make use of the EM algorithm.[1]

14.2.2 An EM algorithm

To derive an EM algorithm, we follow the recipe discussed in Chapter 11. In particular, we write down the complete log likelihood, take the expectation, and maximize the resulting expected complete log likelihood with respect to the parameters.

Before immersing ourselves in the algebra, however, let us step back and consider the results that we expect to obtain. Suppose in particular that we have "complete data"—pairs of observations of X and Y. Clearly the estimation of the distribution of X would reduce to a Gaussian density estimation problem in this case, although given our assumption that X has zero mean and identity covariance matrix, we have no parameters to estimate for the distribution of X. From Eq. (14.3) we see that Y is a linear function of x, with additive white Gaussian noise W. Thus, if both X and Y were observed, we would have a linear regression problem.

This argument suggests that if we can "fill in" X in the E step, we should find that the M step reduces to estimating Λ and Ψ using linear regression. This is in fact correct, but we need to take care in defining what we mean by "fill in." In particular, it is not correct to simply replace X with its conditional expectation in the linear regression formulas. To obtain the correct result, we need to calculate the expected complete log likelihood and identify the expected sufficient statistics for our problem.

14.2.3 The E step

Given complete data, $\mathcal{D}_c = \{(x_n, y_n) : n = 1, \ldots, N\}$, the complete likelihood is simply a product of Gaussian distributions. Taking the logarithm, we obtain:

$$l_c(\theta \,|\, \mathcal{D}_c) = -\frac{N}{2} \log |\Psi| - \frac{1}{2} \sum_n x_n^T x_n - \frac{1}{2} \sum_n (y_n - \Lambda x_n)^T \Psi^{-1} (y_n - \Lambda x_n). \tag{14.30}$$

Although this expression may appear daunting, it has the important property that Λ and Ψ are decoupled, although this fact may not yet be clear. It is in fact a much simpler expression to maximize than the log likelihood.

[1] An alternative approach is to use a nonlinear optimization algorithm such as conjugate gradients. See [?] for a presentation of this approach.

To make further progress we use the "trace trick." Rewriting the quadratic forms, we have:

$$\begin{aligned}
l_c(\theta \,|\, \mathcal{D}_c) &= -\frac{N}{2}\log|\Psi| - \frac{1}{2}\sum_n \mathrm{tr}(x_n^T x_n) - \frac{1}{2}\sum_n \mathrm{tr}[(y_n - \Lambda x_n)^T \Psi^{-1}(y_n - \Lambda x_n)] \\
&= -\frac{N}{2}\log|\Psi| - \frac{1}{2}\sum_n \mathrm{tr}(x_n x_n^T) - \frac{1}{2}\sum_n \mathrm{tr}[(y_n - \Lambda x_n)(y_n - \Lambda x_n)^T \Psi^{-1}] \\
&= -\frac{N}{2}\log|\Psi| - \frac{N}{2}\mathrm{tr}(S\Psi^{-1}), \qquad (14.31)
\end{aligned}$$

where we have defined:

$$S \triangleq \frac{1}{N}\sum_n (y_n - \Lambda x_n)(y_n - \Lambda x_n)^T. \qquad (14.32)$$

Note that this matrix has the form of a sample covariance matrix, although it depends on the unknown parameter Λ.

We now take the conditional expectation of the complete log likelihood, conditioning on the observed data y and the current parameter vector $\theta^{(t)}$. Using the operator notation $\langle\cdot\rangle$ to denote this conditional expectation, we obtain:

$$Q(\theta \,|\, \theta^{(t)}) = -\frac{N}{2}\log|\Psi| - \frac{N}{2}\mathrm{tr}(\langle S\rangle \Psi^{-1}), \qquad (14.33)$$

We now calculate the conditional expectation $\langle S\rangle$—where we substitute the random variable X_n for x_n in the definition of S and treat S as a random quantity. We have:

$$\begin{aligned}
\langle S\rangle &= \frac{1}{N}\sum_n \langle y_n y_n^T - y_n X_n^T \Lambda^T - \Lambda X_n y_n^T + \Lambda X_n X_n^T \Lambda^T\rangle \qquad (14.34) \\
&= \frac{1}{N}\sum_n \left(y_n y_n^T - y_n \langle X_n^T\rangle \Lambda^T - \Lambda\langle X_n\rangle y_n^T + \Lambda\langle X_n X_n^T\rangle \Lambda^T \right), \qquad (14.35)
\end{aligned}$$

where we see that we require the conditional expectations $\langle X_n\rangle$ and $\langle X_n X_n^T\rangle$.

In summary, the expected sufficient statistics that we require for the E step are the conditional expectations $\langle X_n\rangle$ and $\langle X_n X_n^T\rangle$. We have already obtained these expectations in Section 14.1. Thus:

$$\begin{aligned}
\langle X_n\rangle &= E(X_n \,|\, Y_n) \qquad (14.36) \\
\langle X_n X_n^T\rangle &= \mathrm{Var}(X_n \,|\, y_n) + E(X_n \,|\, y_n)E(X_n \,|\, y_n)^T. \qquad (14.37)
\end{aligned}$$

With these equations we "fill in" the conditional distribution of the latent variable X_n.

14.2.4 The M step

We now turn to the M step. To calculate the necessary derivatives, let us recall from Section 13.5.2 how to take derivatives of log determinants:

$$\frac{\partial}{\partial A}\log|A| = A^{-T}, \qquad (14.38)$$

and derivatives of traces:

$$\frac{\partial}{\partial A}\mathrm{tr}[BA] = B^T. \tag{14.39}$$

In Appendix A we also derive a slight extension of the latter result:

$$\frac{\partial}{\partial A}\mathrm{tr}[BA^TCA] = 2CAB, \tag{14.40}$$

when B and C are symmetric.

We compute the derivative of Q with respect to Λ. The relevant terms are:

$$Q(\Lambda \,|\, \theta^{(t)}) = -\frac{1}{2}\sum_n \mathrm{tr}\left\{(y_n y_n^T - y_n\langle X_n^T\rangle\Lambda^T - \Lambda\langle X_n\rangle y_n^T + \Lambda\langle X_n X_n^T\rangle\Lambda^T)\Psi^{-1}]\right\}. \tag{14.41}$$

Taking the derivative, we obtain:

$$\frac{\partial Q}{\partial \Lambda} = \sum_n \Psi^{-1} y_n\langle X_n^T\rangle - \sum_n \Psi^{-1}\Lambda\langle X_n X_n^T\rangle, \tag{14.42}$$

where we have used the fact that $\mathrm{tr}[A] = \mathrm{tr}[A^T]$ and the circulation property of the trace. Setting to zero, we obtain:

$$\Lambda^{(t+1)} = \left(\sum_n y_n\langle X_n^T\rangle\right)\left(\sum_n \langle X_n X_n^T\rangle\right)^{-1}. \tag{14.43}$$

This is the expected result—we have obtained the normal equations from linear regression.[2]

Finally we compute the derivative of Q with respect to Ψ. The terms in the expected complete log likelihood that depend on Ψ are:

$$Q(\Psi \,|\, \theta^{(t)}) = -\frac{N}{2}\log|\Psi| - \frac{N}{2}\mathrm{tr}(\langle S\rangle\Psi^{-1}). \tag{14.44}$$

Recall that Ψ is a diagonal matrix. To calculate the derivative of Q with respect to Ψ, we take the usual matrix derivative but retain only the diagonal terms. Taking the derivative with respect to Ψ^{-1}, setting to zero and retaining only the diagonal terms yields:

$$\Psi^{(t+1)} = \mathrm{diag}\left(\langle S\rangle\right). \tag{14.45}$$

Recall that S depends on the parameter Λ. Thus we must substitute $\Lambda^{(t+1)}$ from Eq. (14.43) into the expression for $\langle S\rangle$. Carrying out this substitution, we find that we in fact obtain a simplification:

$$\begin{aligned}\langle S\rangle &= \frac{1}{N}\sum_n\left(y_n y_n^T - y_n\langle X_n^T\rangle\Lambda^{(t+1)T} - \Lambda^{(t+1)}\langle X_n\rangle y_n^T + \Lambda^{(t+1)}\langle X_n X_n^T\rangle\Lambda^{(t+1)T}\right) && (14.46)\\ &= \frac{1}{N}\left(\sum_n y_n y_n^T - \Lambda^{(t+1)}\langle X_n\rangle y_n^T\right), && (14.47)\end{aligned}$$

[2]Note that in univariate regression the parameter vector enters into the regression model as a *row* vector. Writing the regression model using the notation of this chapter, we have: $Y = \mu + \theta^T x + W$, where Y is a scalar. In factor analysis each row of the matrix Λ corresponds to a parameter vector in univariate regression: $Y = \mu + \Lambda x + W$, where Y is a vector. Thus we should expect to obtain the tranpose of the usual normal equations, and this is indeed what we obtain in Eq. (14.43).

and thus we have:

$$\Psi^{(t+1)} = \frac{1}{N}\mathrm{diag}\left\{\sum_n y_n y_n^T - \Lambda^{(t+1)} \sum_n \langle X_n \rangle y_n^T\right\}, \tag{14.48}$$

as the M step for Ψ.

14.3 Mixtures of factor analyzers

[Section not yet written].

14.4 Principal components analysis and factor analysis

[Section not yet written].

14.5 Appendix A

In this section we show how to calculate the derivative of the expression $\mathrm{tr}[BA^TCA]$ with respect to A. The calculation is based on the equation:

$$\frac{\partial}{\partial A}\mathrm{tr}[BA] = B^T \tag{14.49}$$

that we established in Section 13.5.1.

We use the product rule, first holding A^T constant and then holding A constant. Placing a bar over a matrix to indicate that it is being treated as a constant, we have:

$$\begin{aligned}
\frac{\partial}{\partial A}\mathrm{tr}[BA^TCA] &= \frac{\partial}{\partial A}\mathrm{tr}[B\bar{A}^TCA] + \frac{\partial}{\partial A}\mathrm{tr}[BA^TC\bar{A}] && (14.50)\\
&= (B\bar{A}^TC)^T + \frac{\partial}{\partial A}\mathrm{tr}[\bar{A}^TC^TAB^T] && (14.51)\\
&= C^TAB^T + \frac{\partial}{\partial A}\mathrm{tr}[B^T\bar{A}^TC^TA] && (14.52)\\
&= C^TAB^T + (B^TA^TC^T)^T && (14.53)\\
&= C^TAB^T + CAB. && (14.54)
\end{aligned}$$

A special case of this result is obtained when B and C are symmetric:

$$\frac{\partial}{\partial A}\mathrm{tr}[BA^TCA] = 2CAB, \tag{14.55}$$

which is the result that we require in Section 14.2.4.