```
                              Xn,
                              Y,
                              cv=kfold,
                              scoring='neg_mean_squared_error')
-cv_null['test_score'].mean()
```

Out[53]: 204139.31

The explained_variance_ratio_ attribute of our PCA object provides the *percentage of variance explained* in the predictors and in the response using different numbers of components. This concept is discussed in greater detail in Section 12.2.

In [54]:
```
pipe.named_steps['pca'].explained_variance_ratio_
```

Out[54]: array([0.3831424 , 0.21841076])

Briefly, we can think of this as the amount of information about the predictors that is captured using M principal components. For example, setting $M = 1$ only captures 38.31% of the variance, while $M = 2$ captures an additional 21.84%, for a total of 60.15% of the variance. By $M = 6$ it increases to 88.63%. Beyond this the increments continue to diminish, until we use all $M = p = 19$ components, which captures all 100% of the variance.

Partial Least Squares

Partial least squares (PLS) is implemented in the PLSRegression() function.

PLS Regression()

In [55]:
```
pls = PLSRegression(n_components=2,
                    scale=True)
pls.fit(X, Y)
```

As was the case in PCR, we will want to use CV to choose the number of components.

In [56]:
```
param_grid = {'n_components':range(1, 20)}
grid = skm.GridSearchCV(pls,
                        param_grid,
                        cv=kfold,
                        scoring='neg_mean_squared_error')
grid.fit(X, Y)
```

As for our other methods, we plot the MSE.

In [57]:
```
pls_fig, ax = subplots(figsize=(8,8))
n_comp = param_grid['n_components']
ax.errorbar(n_comp,
            -grid.cv_results_['mean_test_score'],
            grid.cv_results_['std_test_score'] / np.sqrt(K))
ax.set_ylabel('Cross-validated MSE', fontsize=20)
ax.set_xlabel('# principal components', fontsize=20)
ax.set_xticks(n_comp[::2])
ax.set_ylim([50000,250000]);
```

CV error is minimized at 12, though there is little noticable difference between this point and a much lower number like 2 or 3 components.

6.6 Exercises

Conceptual

1. We perform best subset, forward stepwise, and backward stepwise selection on a single data set. For each approach, we obtain $p + 1$ models, containing $0, 1, 2, \ldots, p$ predictors. Explain your answers:

   (a) Which of the three models with k predictors has the smallest *training* RSS?

   (b) Which of the three models with k predictors has the smallest *test* RSS?

   (c) True or False:

   i. The predictors in the k-variable model identified by forward stepwise are a subset of the predictors in the $(k+1)$-variable model identified by forward stepwise selection.

   ii. The predictors in the k-variable model identified by backward stepwise are a subset of the predictors in the $(k + 1)$-variable model identified by backward stepwise selection.

   iii. The predictors in the k-variable model identified by backward stepwise are a subset of the predictors in the $(k + 1)$-variable model identified by forward stepwise selection.

   iv. The predictors in the k-variable model identified by forward stepwise are a subset of the predictors in the $(k+1)$-variable model identified by backward stepwise selection.

   v. The predictors in the k-variable model identified by best subset are a subset of the predictors in the $(k + 1)$-variable model identified by best subset selection.

2. For parts (a) through (c), indicate which of i. through iv. is correct. Justify your answer.

   (a) The lasso, relative to least squares, is:

   i. More flexible and hence will give improved prediction accuracy when its increase in bias is less than its decrease in variance.

   ii. More flexible and hence will give improved prediction accuracy when its increase in variance is less than its decrease in bias.

   iii. Less flexible and hence will give improved prediction accuracy when its increase in bias is less than its decrease in variance.

   iv. Less flexible and hence will give improved prediction accuracy when its increase in variance is less than its decrease in bias.

   (b) Repeat (a) for ridge regression relative to least squares.

   (c) Repeat (a) for non-linear methods relative to least squares.

3. Suppose we estimate the regression coefficients in a linear regression model by minimizing

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p}|\beta_j|\le s$$

for a particular value of s. For parts (a) through (e), indicate which of i. through v. is correct. Justify your answer.

   (a) As we increase s from 0, the training RSS will:
      i. Increase initially, and then eventually start decreasing in an inverted U shape.
      ii. Decrease initially, and then eventually start increasing in a U shape.
      iii. Steadily increase.
      iv. Steadily decrease.
      v. Remain constant.
   (b) Repeat (a) for test RSS.
   (c) Repeat (a) for variance.
   (d) Repeat (a) for (squared) bias.
   (e) Repeat (a) for the irreducible error.

4. Suppose we estimate the regression coefficients in a linear regression model by minimizing

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2$$

for a particular value of λ. For parts (a) through (e), indicate which of i. through v. is correct. Justify your answer.

   (a) As we increase λ from 0, the training RSS will:
      i. Increase initially, and then eventually start decreasing in an inverted U shape.
      ii. Decrease initially, and then eventually start increasing in a U shape.
      iii. Steadily increase.
      iv. Steadily decrease.
      v. Remain constant.
   (b) Repeat (a) for test RSS.
   (c) Repeat (a) for variance.
   (d) Repeat (a) for (squared) bias.
   (e) Repeat (a) for the irreducible error.