Научный вестник НГТУ. – 2014. – № 1(54)

УДК 519.23

Планирование выборочных обследований в задаче идентификации полиномиальных структурных зависимостей*

А.Ю. ТИМОФЕЕВА

Новосибирск, Новосибирский государственный технический университет

В работе решается задача планирования выборочных обследований с целью идентификации структурных зависимостей путем построения D-оптимальных планов. Для оценивания полиномиальных моделей впервые получено выражение для информационной матрицы плана, учитывающее погрешности при регистрации значений входного фактора. Доказаны условия ее положительной полуопределенности в случае квадратичной зависимости. Во-первых, дисперсия объясняющей переменной должна быть больше дисперсии ошибки. Во-вторых, значение эксцесса распределения входного признака должно превышать его величину для нормального распределения. В противном случае для обеспечения неотрицательности определителя информационной матрицы требуется, чтобы отношение дисперсии ошибки к дисперсии входного фактора было меньше некоторой величины. Для решения задачи наиболее точного оценивания параметров полинома второй степени, на примере опорных точек плана, оптимального в условиях отсутствия погрешностей во входном факторе, (классический вариант) найдено аналитическое выражение для весов этих точек, обеспечивающих наибольшее значение определителя информационной матрицы при разной величине дисперсии ошибки объясняющей переменной. При значительном уровне погрешностей веса оптимального плана существенно отклоняются от классического варианта с равными весами, при этом для достижения максимальной информативности требуется увеличивать вес в точке 0. Для этого случая также получена граница положительной полуопределенности информационной матрицы оптимальных планов в виде ограничения на дисперсию ошибки входного фактора.

Ключевые слова: планирование эксперимента, выборочное обследование, структурная зависимость, информационная матрица, метод скорректированных наименьших квадратов, критерий оптимальности, положительная полуопределенность матрицы, нормальная ошибка.

ВВЕДЕНИЕ И ПОСТАНОВКА ЗАДАЧИ

Традиционно задачи экспериментирования и выборочного обследования разграничиваются с точки зрения целей их проведения. Обследования предпринимаются для описания существующей совокупности объектов посредством оценивания характеристик распределения их признаков. В ходе эксперимента исследуется некое соотношение, которое необязательно соответствует какой-либо совокупности, т. е. восстанавливается гипотетическая зависимость [1]. При этом естественно предполагается, что экспериментатор фиксирует требуемые значения входного признака без погрешности или с пренебрежимо малыми погрешностями. Такую схему будем называть классической.

Однако представляет интерес и задача восстановления зависимостей между признаками по данным выборочных обследований, для достижения наилучшей точности которого требуется комбинирование инструментария теории планирования эксперимента и методов формирования выборки. При этом существенной проблемой выступает случайный характер значений признаков, регистрируемых в ходе выборочных обследований, а также наличие некоторой доли погрешностей при измерении этих значений. Тем самым приходится иметь дело с так называемой структурной зависимостью [2]. Это сказывается как на результатах оценивания

* *Статья получена 26 декабря 2013 г.*
Работа выполнена при финансовой поддержке РФФИ в рамках научного проекта № 14-07-31171 мол_а

моделей, так и на информативности индивидуальных наблюдений. Следовательно, возникает потребность в адаптации теории планирования эксперимента для случая планирования выборочных обследований с целью наиболее точного оценивания структурных зависимостей.

Влияние наличия ошибок во входных факторах на построение оптимальных планов эксперимента рассмотрим на примере полиномиальной модели

$$Y = \theta_0 + \theta_1 X + \theta_2 X^2 + \ldots + \theta_m X^m, \tag{1}$$

где θ_i – некоторые постоянные величины, а обе переменные X и Y наблюдаются с некоторыми случайными погрешностями. Следовательно, вместо истинных значений Y_i и X_i фиксируются значения

$$y_i = Y_i + \varepsilon_i,\ x_i = X_i + \delta_i,\ i = \overline{1, N}, \tag{2}$$

где N – объемы выборки, ε_i, δ_i – случайные ошибки с распределением, не зависящем от номера наблюдения i, относительно которых предполагается:

$$\left.\begin{array}{c} E(\varepsilon_i) = E(\delta_i) = 0,\ D(\varepsilon_i) = \sigma_\varepsilon^2,\ D(\delta_i) = \sigma_\delta^2,\ \forall i \\ \operatorname{cov}(\varepsilon_i, \varepsilon_j) = \operatorname{cov}(\delta_i, \delta_j) = 0,\ \forall i \neq j,\ \operatorname{cov}(\varepsilon_i, \delta_j) = 0,\ \forall i, j. \end{array}\right\} \tag{3}$$

В моделях с погрешностями во входных факторах различают два случая: с детерминированными (функциональный случай) и стохастическими (структурный случай) X_i [2]. Далее будем рассматривать X_i как одинаково распределенные случайные величины с r конечными начальными моментами $V_1, \ldots, V_r$. Дополнительно требуется ввести предположения о некоррелированности ошибок с истинными переменными:

$$\operatorname{cov}(X_i, \delta_j) = \operatorname{cov}(X_i, \varepsilon_j) = \operatorname{cov}(Y_i, \delta_j) = \operatorname{cov}(Y_i, \varepsilon_j) = 0,\ \forall i, j. \tag{4}$$

Пусть наблюдаемые значения входного фактора фиксируются в соответствии с некоторым заданным дискретным распределением с r конечными начальными моментами $\nu_1, \ldots, \nu_r$ и с коэффициентом эксцесса γ_2. Иными словами, обследование проводится в соответствии с нормированным планом:

$$\xi = \begin{Bmatrix} x_1 & \ldots & x_n \\ p_1 & \ldots & p_n \end{Bmatrix}, \tag{5}$$

где $x_1, \ldots, x_n$ – опорные точки (спектр) плана, приведенные к промежутку [–1, 1], $p_i = \dfrac{n_i}{N}$ – вес точки x_i, определяемый как доля повторных наблюдений n_i в i-й точке к общему числу наблюдений N.

Задача состоит в построении оптимального плана (5), обеспечивающего наилучшую точность оценивания параметров модели (1) по критерию D-оптимальности [3]:

$$\xi^* = Arg \max_{\xi} \left| M(\xi) \right|,$$

где M – информационная матрица плана, которая будет определена далее.

1. ВЫЧИСЛЕНИЕ ИНФОРМАЦИОННОЙ МАТРИЦЫ

При выполнении условий регулярности [4] информационная матрица плана выражается следующим образом:

$$M = -E\left[\frac{\partial^2 L}{\partial \theta^2}\right], \tag{6}$$

где L – логарифмическая функция правдоподобия, которая в предположении совместного нормального распределения ошибок модели (1, 2) имеет вид [2]

$$L = const - N\ln\sigma_\varepsilon - N\ln\sigma_\delta - \frac{1}{2\sigma_\delta^2}\sum_{i=1}^{N}\left(x_i - X_i\right)^2 -$$
$$-\frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{N}\left(y_i - \theta_0 - \theta_1 X_i - \theta_2 X_i^2 - \ldots - \theta_m X_i^m\right)^2.$$

Тогда (k, j)-й элемент информационной матрицы (6) определяется как

$$M_{kj} = -E\left[\frac{\partial^2 L}{\partial\theta_k\partial\theta_j}\right] = -\frac{1}{\sigma_\varepsilon^2}E\left[\frac{\partial}{\partial\theta_j}\sum_{i=1}^{N}X_i^k\left(y_i - \sum_{l=0}^{m}\theta_l X_i^l\right)\right] =$$
$$= \frac{1}{\sigma_\varepsilon^2}E\left[\sum_{i=1}^{N}X_i^{k+j}\right] = \frac{N}{\sigma_\varepsilon^2}V_{k+j},\ k, j = \overline{0, m}. \tag{7}$$

Для вычисления элементов информационной матрицы необходимо определить V_r, $r = \overline{0, 2m}$. Для этого предлагается воспользоваться идеей метода скорректированных наименьших квадратов (Adjusted least squares), предложенного в [5]. При фиксированной σ_δ^2 математическое ожидание наблюдаемой случайной величины x^r можно представить в виде

$$E\left[x^r\right] = \nu_r = E\left[\left(X + \delta\right)^r\right].$$

Воспользовавшись формулой бинома, в предположениях (3, 4) получаем

$$\nu_r = \sum_{j=0}^{r} C_r^j V_j E\left[\delta^{r-j}\right].$$

Далее выражая начальные моменты ненаблюдаемой случайной величины, приходим к рекуррентному соотношению

$$V_r = \nu_r - \sum_{j=0}^{r-1} C_r^j V_j E\left[\delta^{r-j}\right],\ r = \overline{0, 2m}, \tag{8}$$

где $V_0 = 1$. При нормальном распределении ошибки δ его нечетные моменты равны нулю, а четные можно определить как $E\left[\delta^{2k}\right] = \sigma_\delta^{2k}\left(2k-1\right)!!$.

Величина ν_r определяется исходя из нормированного плана как

$$\nu_r = \sum_{i=1}^{n} p_i x_i^r\ . \tag{9}$$

Таким образом, подставив (8, 9) в (6), выразим информационную матрицу через опорные точки плана и их веса:

$$M = \frac{N}{\sigma_\varepsilon^2}\sum_{i=1}^{n} p_i \mu(i),$$

где $\mu(i)$ – вклад в информационную матрицу i-й точки плана, (k,l) -й элемент этой матрицы определяется как

$$\mu_{kl}(i) = x_i^{k+l} - \sum_{j=0}^{k+l-1} C_r^j \mu_{j0}(i) E\left[\delta^{k+l-j}\right],\ \mu_{00}(i) = 1.$$

Согласно введенным предположениям (3), дисперсия ошибки отклика постоянна в каждой точке плана. Без потери общности положим $\sigma_\varepsilon^2 = 1$. Общее число экспериментов также будем считать постоянным. Далее в ходе анализа рассматривается информационная матрица на одно наблюдение.

2. СВОЙСТВО ПОЛОЖИТЕЛЬНОЙ ПОЛУОПРЕДЕЛЕННОСТИ ИНФОРМАЦИОННОЙ МАТРИЦЫ

В классической схеме информационная матрица всегда обладает свойством положительной полуопределенности. Однако при наличии ошибок при регистрации входных факторов это свойство может нарушаться. Исследуем эту проблему на примере полинома второй степени.

В этом случае согласно соотношениям (7, 8) информационная матрица на единицу наблюдений рассчитывается следующим образом:

$$M = \begin{pmatrix} 1 & \nu_1 & \nu_2 - \sigma_\delta^2 \\ \nu_1 & \nu_2 - \sigma_\delta^2 & \nu_3 - 3\nu_1\sigma_\delta^2 \\ \nu_2 - \sigma_\delta^2 & \nu_3 - 3\nu_1\sigma_\delta^2 & \nu_4 - 6\nu_2\sigma_\delta^2 + 3\sigma_\delta^4 \end{pmatrix}.$$

Без потери общности, предположим, что значения x центрированы, т.е. $\nu_1 = 0$. Тогда определитель информационной матрицы имеет вид

$$|M| = \left(\nu_2 - \sigma_\delta^2\right)\left(\nu_4 - 6\nu_2\sigma_\delta^2 + 3\sigma_\delta^4 - \left(\nu_2 - \sigma_\delta^2\right)^2\right) - \nu_3^2. \qquad (10)$$

Очевидно, что увеличение абсолютного значения момента третьего порядка распределения входного признака приводит к уменьшению значений определителя (10) и снижению информативности наблюдений в соответствии с критерием D-оптимальности. Следовательно, имеет смысл ограничиться рассмотрением только симметричных распределений x.

Существенным ограничением корректности задачи планирования с ошибкой во входном факторе выступает отношение

$$\gamma_x = \frac{\sigma_\delta^2}{\nu_2}, \qquad (11)$$

которое обозначим как уровень шума плана. Далее докажем условия положительной полуопределенности информационной матрицы.

Утверждение. Для того чтобы информационная матрица для полиномиальной модели (1, 2) при $m = 2$ и предположениях относительно ошибок (3, 4) с симметричным распределением x

была положительно полуопределенной, необходимо и достаточно, чтобы одновременно выполнялись следующие условия:

$$1.\ \gamma_x \le 1, \tag{12}$$

$$2.\ \gamma_2 \ge 3 \text{ или } \begin{cases} \gamma_2 < 3, \\ \gamma_x \le 1 - \sqrt{\dfrac{3-\gamma_2}{2}}. \end{cases} \tag{13}$$

Доказательство. Согласно теореме из [6] для того, чтобы матрица M была положительно полуопределенной, необходимо и достаточно, чтобы все ее главные миноры были неотрицательны.

Условие неотрицательности второго главного минора $\nu_2 - \sigma_\delta^2 \ge 0$ с учетом (11) совпадает с (12). Такое ограничение естественно, поскольку дисперсия наблюдаемого признака должна быть больше дисперсии ошибки наблюдения.

Из условия симметричности распределения случайной величины x имеем $\nu_3 = 0$. Определим условия его неотрицательности определителя (10). С учетом неотрицательности второго главного минора информационной матрицы первый сомножитель из (10) опустим. После некоторых упрощений получим

$$\sigma_\delta^4 - 2\nu_2\sigma_\delta^2 + \frac{1}{2}\left(\nu_4 - \nu_2^2\right) \ge 0 .$$

Полученное неравенство выполняется для любых σ_δ^2 в случае, если дискриминант квадратного уравнения относительно σ_δ^2 неположительный:

$$3\nu_2^2 - \nu_4 \le 0 .$$

Если это условие нарушается, то можно определить границы интервала изменения дисперсии ошибки объясняющей переменной, при котором определитель информационной матрицы становится отрицательным:

$$\sigma_{\delta_{1,2}}^2 = \nu_2 \pm \sqrt{\frac{3\nu_2^2 - \nu_4}{2}} .$$

Поскольку при превышении дисперсии ошибки нижней границы $\sigma_{\delta_2}^2$ определитель информационной матрицы становится отрицательным, то при условии $3\nu_2^2 - \nu_4 > 0$ получаем ограничение на $\sigma_\delta^2 \le \sigma_{\delta_2}^2$. С учетом того, что коэффициент эксцесса при введенных предположениях определяется как

$$\gamma_2 = \frac{\nu_4}{\nu_2^2} ,$$

полученное условие совпадает с (13). Утверждение полностью доказано.

Из доказанного утверждения следует, что в условиях, когда коэффициент эксцесса распределения x превышает 3, определитель информационной матрицы будет положительным при любом значении дисперсии ошибки, не превышающем дисперсию входного фактора, т. е. уровне шума плана γ_x, не превосходящем 1.

Если коэффициент эксцесса распределения объясняющей переменной меньше 3 (в частности, при равномерном распределении), то для обеспечения положительности определителя

информационной матрицы должно быть выполнено ограничение сверху на уровень шума по x. Так при равномерном распределении ($\gamma_2 = 1.8$) получаем, что $\gamma_x \le 1 - \sqrt{0.6} \approx 0.2254$, т. е. дисперсия входного фактора должна более чем в четыре раза превышать дисперсию ошибки. Далее перейдем к построению оптимальных планов.

3. ПОСТРОЕНИЕ D-ОПТИМАЛЬНЫХ ПЛАНОВ

Полученное выражение (10) указывает на то, что при фиксированной дисперсии ошибки наблюдения определитель информационной матрицы зависит от второго и четвертого моментов распределения входного фактора. Очевидно, что между моментами различного порядка существует взаимосвязь, определяемая видом выбранного распределения. Следовательно, задача построения оптимального плана может быть решена для выбранного класса распределений, описывающих план выборки.

Предположим, что выборка формируется в соответствии с планом:

$$\xi = \begin{Bmatrix} -1 & 0 & 1 \\ p_1 & p_2 & p_3 \end{Bmatrix},$$

который в классической схеме является D-оптимальным при $p_1 = p_2 = p_3 = 1/3$.

Для того чтобы обеспечить симметричность распределения и нулевое математическое ожидание входного фактора, необходимо выполнение условия $p_1 = p_3$. Тогда с учетом единичной суммы весов имеем $p_2 = 1 - 2p_1$.

Взаимосвязь между моментами входного признака выражается как

$$\nu_2 = \nu_4 = 2p_1 = 1 - p_2 . \tag{14}$$

С учетом этих равенств после некоторых упрощений выражение для определителя информационной матрицы (10) примет вид

$$|M| = p_2^3 - p_2^2\left(2 + 3\sigma_\delta^2\right) + p_2\left(1 + 7\sigma_\delta^2 - 6\sigma_\delta^4\right) - 4\sigma_\delta^2 + 6\sigma_\delta^4 - 2\sigma_\delta^6 ,$$

что представляет собой кубическую параболу по весу, точка локального минимума которой соответствует значению

$$p_2^* = \frac{2}{3} + \sigma_\delta^2 - \frac{1}{3}\sqrt{1 - 9\sigma_\delta^2 + 27\sigma_\delta^4} . \tag{15}$$

Очевидно, при отсутствии ошибки во входном факторе ($\sigma_\delta^2 = 0$) оптимальным будет вес 1/3, что согласуется с классическим D-оптимальным планом. На рис. 1 изображена полученная зависимость. Для удобства интерпретации по оси абсцисс отложены значения стандартного отклонения ошибки.

Проверим выполнение условий положительной определенности информационной матрицы. Условие (12) с учетом (14) может быть выражено как

$$p_2 < 1 - \sigma_\delta^2 .$$

На рис. 1 область, в которой это условия нарушается, заштрихована. Точка пересечения границы этого условия с (15) соответствует дисперсии ошибки входного фактора, равной 1/3.

Для проверки выполнения условий (13) они также выражены через дисперсию ошибки входного фактора и вес оптимального плана в точке 0 и отображены на рис. 1. Штрихпунктирная линия соотвествует эксцессу, равному 3. Пунктирной линией представлена граница ограничения на уровень шума плана из (13). Очевидно, что при $\sigma_\delta^2 \le 1/3$ неравенства (13) выполняются.

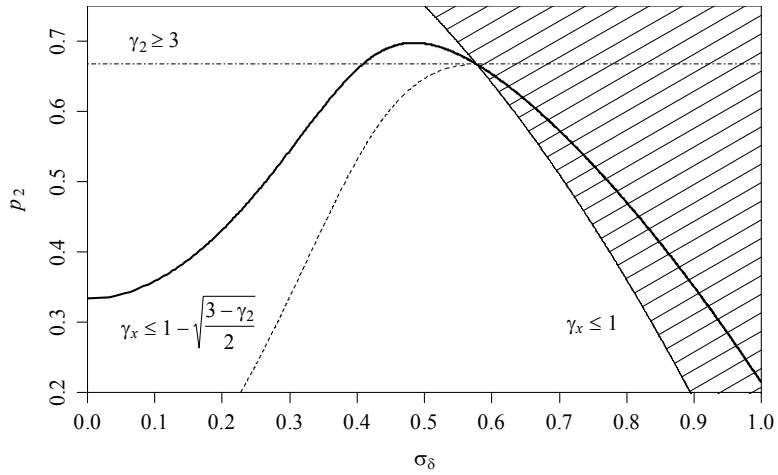

Рис. 1. Зависимость веса в точке 0 D-оптимального плана от стандартного отклонения входного признака

Таким образом, в условиях, когда дисперсия ошибки превышает 1/3, в данном классе распределений входного признака нельзя подобрать план, обеспечивающий положительное приращение информации.

ЗАКЛЮЧЕНИЕ

В рамках постановки задачи построения оптимальных планов выборки для наиболее точного оценивания параметров полиномиальных структурных моделей впервые получено выражение для информационной матрицы, учитывающее наличие случайной ошибки во входном факторе. Для полинома второй степени доказаны условия положительной полуопределенности информационной матрицы, которые выражаются в некоторых ограничениях на величину эксцесса распределения входного признака и уровень шума плана. Для сравнения с классическим случаем определены веса точек плана {–1,0,1} при различном уровне дисперсии ошибки входной переменной. Установлено, что при ее росте для достижения оптимальности плана требуется увеличение веса в точке 0. При этом показано, что при уровне дисперсии ошибки объясняющей переменной, превышающем 1/3, нельзя подобрать оптимальный план эксперимента, поскольку информационная матрица теряет свойство положительной полуопределенности. Полученные результаты могут быть применены при решении различных практических задач, например, представленных в [7].

СПИСОК ЛИТЕРАТУРЫ

[1] **Кендалл М.** Многомерный статистический анализ и временные ряды / М. Кендалл, А. Стьюарт. – М.: Наука, 1976. – 736 с.

[2] **Кендалл М.** Статистические выводы и связи / М. Кендалл, А. Стьюарт. – М.: Наука, 1973. – 899 с.

[3] **Федоров В.В.** Теория оптимального планирования эксперимента. / В.В. Федоров. – М. Наука, 1971. – 312 с.

[4] **Закс Ш.** Теория статистических выводов / Ш. Закс. – М.: Мир, 1975. – 776 с.

[5] **Cheng C.-L.** Polynomial regression with errors in the variables / C.-L. Cheng, H. Schneeweiss // Journal of the Royal Statistical Society: Series B. – 1998. – Vol. 60. – P. 189–199.

[6] **Гантмахер Ф.Р.** Теория матриц / Ф.Р. Гантмахер. – М.: Наука, 1966. – 576 с.

[7] **Денисов В.И.** Устойчивое оценивание нелинейных структурных зависимостей / В.И. Денисов, А.Ю. Тимофеева, Е.А. Хайленко, О.И. Бузмакова // Сибирский журнал индустриальной математики. – 2013. – № 4. – С. 47–60.

REFERENCES

[1] **Kendall M., St'iuart A.** The Advanced Theory of Statistics: Design and analysis, and time-series. Charles Griffin and Co., Ltd., London, 1961. (Russ. ed.: Kendall M., St'iuart A. Mnogomernyi statisticheskii analiz i vremennye riady. Moscow, Nauka Publ., 1976, 736 p.).

[2] **Kendall M., St'iuart A.** The Advanced Theory of Statistics: Inference and relationship. Charles Griffin and Co., Ltd., London, 1961. (Russ. ed.: Kendall M., St'iuart A. Statisticheskie vyvody i sviazi. Moscow, Nauka Publ., 1973, 899 p.).

[3] **Fedorov V.V.** Teoriya optimalnogo planirovaniya eksperimenta [Theory of optimal design of experiments]. Moscow, Nauka Publ., 1971, 312 p.

[4] **Zacks S.** The Theory of Statistical Inference. John Wiley & Sons Inc, 1971, 626 p. (Russ. ed.: Zaks Sh. Teoriia statisticheskikh vyvodov. Moscow, Mir Publ., 1975, 776 p.).

[5] **Cheng C.-L.** Polynomial regression with errors in the variables / C.-L. Cheng, H. Schneeweiss // *Journal of the Royal Statistical Society: Series B*, 1998, Vol. 60, pp. 189–199. doi: 10.1111/1467-9868.00118

[6] **Gantmakher F.R.** Teoriia matrits [Matrix theory]. Moscow, Nauka Publ., 1966, 576 p.

[7] **Denisov V.I., Timofeeva A.Iu., Khailenko, Buzmakova O.I.** Ustoichivoe otsenivanie nelineinykh strukturnykh zavisimostei [Robust estimation of nonlinear structural models]. *Sibirskii zhurnal industrial'noi matematiki – Journal of Applied and Industrial Mathematics*, 2013, no. 4, pp. 47–60.

Тимофеева Анастасия Юрьевна, кандидат экономических наук, старший преподаватель кафедры экономической информатики Новосибирского государственного технического университета. Основное направление научных исследований – развитие методов статистического анализа объектов стохастической природы, в том числе социально-экономических явлений. Имеет 25 публикаций. E-mail: supernasty@mail.ru.

A.Yu. Timofeeva

Sample survey design for identification of polynomial structural relationship

In this paper the problem of sample survey design for identification of structural relation on basis of the D-optimal approach is solved. For the estimation of polynomial models the expression for information matrix is extracted. For the first time it takes into consideration experimental error of input factor. The conditions of positive semidefiniteness of information matrix for quadratic dependence are proved. It is required first the input factor variance to be greater than the error variance. Secondly for nonnegative determinant of information matrix it is required value of kurtosis of the input factor distribution to rank over value of normal kurtosis, otherwise the ratio of the error variance to the input factor variance to be less than certain value. For solving the problem of the most accurate parameter estimation of second-degree polynomial by example of no random disturbance optimal control points (classic version) the analytical expression for spectrum weights maximizing determinant of information matrix by various input factor error variance is discovered. If level of error is considerable the weights of optimal design essentially diverge from the classical version with equal weights. For descriptiveness increasing weight of point 0 must be larger. In addition for this case the boundary of positive semidefiniteness of information matrix is derived in the form of restriction on the error variance of the input factor.

Key words: design of experiment, sample survey, structural relation, information matrix, adjusted least squares, optimality criterion, positive semidefiniteness of matrix, normal error.