AUTOMATIC
MODEL SELECTION

The linear regression model

Let

$$y=\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

be an n-dimensional random vector with mean vector $\mu=Ey$ and covariance matrix $\Sigma=\mathrm{var}(y)$.

For the standard linear model, we assume that

$$\mu=\begin{pmatrix} Ey_1 \\ \vdots \\ Ey_n \end{pmatrix}=\begin{pmatrix} \beta_1 x_{11}+...+\beta_k x_{1k} \\ \vdots \\ \beta_1 x_{n1}+...+\beta_k x_{nk} \end{pmatrix}=\underbrace{\begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix}}_{X}\underbrace{\begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}}_{\beta}=X\beta$$

and $$\Sigma=\begin{pmatrix} \mathrm{Var}(y_1) & \mathrm{Cov}(y_1,y_2) & \cdots & \mathrm{Cov}(y_1,y_n) \\ \mathrm{Cov}(y_2,y_1) & \mathrm{Var}(y_2) & \cdots & \mathrm{Cov}(y_2,y_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(y_n,y_1) & \mathrm{Cov}(y_n,y_2) & \cdots & \mathrm{Var}(y_n) \end{pmatrix}$$

$$=\begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix}=\sigma^2\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}=\sigma^2 I,$$

where the k columns of the matrix X (the k regressors) are linearly independent.

Likelihood function of the linear model

If $y=(y_1,\ldots,y_k)^T$ has a multivariate normal distribution with a diagonal covariance matrix, the multivariate normal density $f(y_1,\ldots,y_k)$ factors into n univariate normal densities:

$$f(y_1,\ldots,y_k)=\prod_{t=1}^{n} f_t\,(y_t)$$
$$=\prod_{t=1}^{n}\frac{1}{\sqrt{2\pi \mathrm{Var}(y_t)}}\exp\left(-\frac{(y_t-Ey_t)^2}{2\mathrm{Var}(y_t)}\right)$$

Under the assumptions $Ey=X\beta$ and $var(y)=\sigma^2 I$ of the linear model we have

$$f(y_1,\ldots,y_k)=\prod_{t=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y_t-Ey_t)^2}{2\sigma^2}\right)$$
$$=(2\pi\sigma^2)^{-\frac{n}{2}}\exp\left(-\tfrac{1}{2\sigma^2}\sum_{t=1}^{n}(y_t-\beta_1 x_{t1}-\ldots-\beta_k x_{tk})^2\right)$$

This density is a function of $y_1,\ldots,y_k$ with fixed model parameters $\beta_1,\ldots,\beta_k$, and σ^2. When we want to stress the dependence of the density on the model parameters, we write $f(y_1,\ldots,y_k;\beta_1,\ldots,\beta_k,\sigma^2)$ instead of $f(y_1,\ldots,y_k)$. Viewing $f(y_1,\ldots,y_k;\beta_1,\ldots,\beta_k,\sigma^2)$ as a function of $\beta_1,\ldots,\beta_k$, and σ^2 with $y_1,\ldots,y_k$ fixed, we obtain the likelihood function of the linear model:

$$L(\beta_1,\ldots,\beta_k,\sigma^2;y_1,\ldots,y_k)= f(y_1,\ldots,y_k;\beta_1,\ldots,\beta_k,\sigma^2)$$

ML estimators for the model parameters

The maximum likelihood (ML) estimators for the model parameters are obtained by maximizing the likelihood function or equivalently the log likelihood function

$$\log L(\beta_1,\ldots,\beta_k,\sigma^2;y_1,\ldots,y_k)$$

$$= -\frac{n}{2}\log(2\pi\sigma^2)-\frac{1}{2\sigma^2}\sum_{t=1}^{n}(y_t-\beta_1 x_{t1}-\ldots-\beta_k x_{tk})^2.$$

Setting the partial derivatives of the log likelihood with respect to $\beta_1,\ldots,\beta_k$, and σ^2 to zero gives

$$-\frac{1}{2\sigma^2}\sum_{t=1}^{n}2(y_t-\beta_1 x_{t1}-\ldots-\beta_k x_{tk})(-x_{tj})=0,\ j=1,\ldots,k,$$

$$-\frac{n}{2}\frac{1}{\sigma^2}+\frac{1}{2\sigma^4}\sum_{t=1}^{n}(y_t-\beta_1 x_{t1}-\ldots-\beta_k x_{tk})^2=0,$$

which is equivalent to

$$-\frac{1}{\sigma^2}(y-X\beta)^T X=0,\ -\frac{n}{\sigma^2}+\frac{1}{\sigma^4}(y-X\beta)^T(y-X\beta)=0$$

and also to

$$X^T(y-X\beta)=0,\ -n\sigma^2+(y-X\beta)^T(y-X\beta)=0$$

and finally also to

$$X^T y=X^T X\beta,\ (y-X\beta)^T(y-X\beta)=n\sigma^2.$$

Thus

$$\hat\beta=(\hat\beta_1,\ldots,\hat\beta_k)^T=(X^T X)^{-1}X^T y,\ \hat\sigma^2=\frac{1}{n}(y-X\hat\beta)^T(y-X\hat\beta).$$

Geometrical interpretation

$X\hat{\beta}=X(X^TX)^{-1}X^Ty$ is the projection of y onto the subspace of $\mathbb{R}^n$ spanned by the columns $x_1,\ldots,x_k$ of X because

$$X\hat{\beta}=\begin{pmatrix}\hat{\beta}_1x_{11}+\ldots+\hat{\beta}_kx_{1k}\\ \vdots\\ \hat{\beta}_1x_{n1}+\ldots+\hat{\beta}_kx_{nk}\end{pmatrix}=\hat{\beta}_1\begin{pmatrix}x_{11}\\ \vdots\\ x_{n1}\end{pmatrix}+\ldots+\hat{\beta}_k\begin{pmatrix}x_{1k}\\ \vdots\\ x_{nk}\end{pmatrix}$$

is an element of span($x_1,\ldots,x_k$) and

$$\begin{pmatrix}x_1^T\\ \vdots\\ x_k^T\end{pmatrix}(y-X\hat{\beta})=X^T(y-X\hat{\beta})=X^Ty-X^TX(X^TX)^{-1}X^Ty=0,$$

which implies that $y-X\hat{\beta}$ is an element of the orthogonal complement of span($x_1,\ldots,x_k$).

Analogously, $y-X\hat{\beta}=(I-X(X^TX)^{-1}X^T)y$ is the projection of y onto the orthogonal complement of span($x_1,\ldots,x_k$) because

$$y-X\hat{\beta}\in(\text{span}(x_1,\ldots,x_k))^{\perp}$$

and

$$y-(y-X\hat{\beta})=X\hat{\beta}\in\text{span}(x_1,\ldots,x_k)=((\text{span}(x_1,\ldots,x_k))^{\perp})^{\perp}.$$

<u>Exercise:</u> Show that the matrices

$$P_X=X(X^TX)^{-1}X^T,\ P_{X^{\perp}}=I-X(X^TX)^{-1}X^T$$

are symmetric and idempotent.

Expected values of the ML estimators

$\hat\beta=(X^TX)^{-1}X^Ty$ is an unbiased estimator for β because

$$E\hat\beta=(X^TX)^{-1}X^TEy=(X^TX)^{-1}X^TX\beta=\beta.$$

Furthermore, using

$$E\,P_{X^\perp}\,y=E(y-X\hat\beta)=Ey-EX\hat\beta=X\beta-XE\hat\beta=X\beta-X\beta=0$$

we obtain

$$\begin{aligned}E(y-X\hat\beta)^T(y-X\hat\beta)&=E(P_{X^\perp}y)^T(P_{X^\perp}y)=E\,\mathrm{tr}\,(P_{X^\perp}y)^TP_{X^\perp}y\\&=E\,\mathrm{tr}\,P_{X^\perp}y(P_{X^\perp}y)^T=\mathrm{tr}\,E\,P_{X^\perp}y(P_{X^\perp}y)^T\\&=\mathrm{tr}\,\mathrm{var}(P_{X^\perp}y)=\mathrm{tr}\,P_{X^\perp}\mathrm{var}(y)P_{X^\perp}^T\\&=\mathrm{tr}\,P_{X^\perp}\sigma^2I\,P_{X^\perp}=\sigma^2\,\mathrm{tr}\,P_{X^\perp}P_{X^\perp}\\&=\sigma^2\,\mathrm{tr}\,P_{X^\perp}=\sigma^2\,\mathrm{tr}(I-X(X^TX)^{-1}X^T)\\&=\sigma^2(\mathrm{tr}\,I-\mathrm{tr}\,X(X^TX)^{-1}X^T)\\&=\sigma^2(\mathrm{tr}\underbrace{I}_{n\times n}-\mathrm{tr}\underbrace{X^TX(X^TX)^{-1}}_{k\times k})=\sigma^2(n-k).\end{aligned}$$

Thus

$$E\hat\sigma^2=E\tfrac{1}{n}(y-X\hat\beta)^T(y-X\hat\beta)=\tfrac{n-k}{n}\sigma^2.$$

<u>Exercise:</u> Show that

$$\mathrm{Cov}(X\hat\beta,y-X\hat\beta)=0.$$

The final prediction error criterion

Let y and z be independent and identically distributed (i.i.d.) normal random vectors with mean vector $X\beta$ and covariance matrix $\sigma^2 I$.

Using the ML estimate $\hat{\beta}=(X^TX)^{-1}X^Ty$ obtained from y we may predict z by $X\hat{\beta}$. It follows from

$$
\begin{aligned}
2\sigma^2 I&=\mathrm{Var}(z)+\mathrm{Var}(y)=\mathrm{Var}(z-y)=\mathrm{Var}((z-X\hat{\beta})-(y-X\hat{\beta}))\\
&=\mathrm{Var}(z-X\hat{\beta})-2\mathrm{Cov}(z-X\hat{\beta},y-X\hat{\beta})+\mathrm{Var}(y-X\hat{\beta})\\
&=\mathrm{Var}(z-X\hat{\beta})-2\mathrm{Cov}(z,y-X\hat{\beta})+2\mathrm{Cov}(X\hat{\beta},y-X\hat{\beta})+\mathrm{Var}(y-X\hat{\beta})\\
&=\mathrm{Var}(z-X\hat{\beta})+\mathrm{Var}(y-X\hat{\beta})\\
&=E(z-X\hat{\beta})(z-X\hat{\beta})^T+E(y-X\hat{\beta})(y-X\hat{\beta})^T
\end{aligned}
$$

that

$$
\begin{aligned}
2n\sigma^2=\mathrm{tr}(2\sigma^2 I)&=E\,\mathrm{tr}(z-X\hat{\beta})(z-X\hat{\beta})^T+E\,\mathrm{tr}(y-X\hat{\beta})(y-X\hat{\beta})^T\\
&=E\,\mathrm{tr}(z-X\hat{\beta})^T(z-X\hat{\beta})+E\,\mathrm{tr}(y-X\hat{\beta})^T(y-X\hat{\beta})\\
&=E(z-X\hat{\beta})^T(z-X\hat{\beta})+E(y-X\hat{\beta})^T(y-X\hat{\beta})\\
&=E(z-X\hat{\beta})^T(z-X\hat{\beta})+(n-k)\sigma^2.
\end{aligned}
$$

Thus, the mean squared prediction error is given by

$$\frac{1}{n}E(z-X\hat{\beta})^T(z-X\hat{\beta})=\frac{n+k}{n}\sigma^2$$

and an unbiased estimator for it is

$$\mathrm{FPE}(k)=\frac{n+k}{n}\frac{n}{n-k}\hat{\sigma}^2=\frac{n+k}{n-k}\hat{\sigma}^2=\left(1+\frac{2k}{n-k}\right)\hat{\sigma}^2.$$

The corrected AIC

Let y and z be i.i.d. $N(X\beta,\sigma^2 I)$ and

$$\hat\beta=(X^TX)^{-1}X^Ty,\ \hat\sigma^2=\tfrac{1}{n}(y-X\hat\beta)^T(y-X\hat\beta).$$

A measure that is somehow related to the mean squared prediction error is

$$E(-2\log f(z;\hat\beta,\hat\sigma^2))=E(n\log(2\pi)+n\log\hat\sigma^2+\frac{(z-X\hat\beta)^T(z-X\hat\beta)}{\hat\sigma^2}).$$

Here $f(z;\hat\beta,\hat\sigma^2)$ is viewed as a function of z, $\hat\beta$, and $\hat\sigma^2$. Clearly, the naïve estimator

$$-2\log f(y;\hat\beta,\hat\sigma^2)$$

underestimates $E(-2\log f(z;\hat\beta,\hat\sigma^2))$. It follows from

$$E[-2\log f(z;\hat\beta,\hat\sigma^2)] - E[-2\log f(y;\hat\beta,\hat\sigma^2)]$$
$$=E\frac{(z-X\hat\beta)^T(z-X\hat\beta)}{\hat\sigma^2} - E\frac{(y-X\hat\beta)^T(y-X\hat\beta)}{\hat\sigma^2}$$
$$=E(z-X\hat\beta)^T(z-X\hat\beta)\frac{n}{\sigma^2}E\Big(\frac{1}{\frac{n\hat\sigma^2}{\sigma^2}}\Big) - n$$
$$=\mathrm{tr}(\mathrm{var}(z)+\mathrm{var}(X\hat\beta))\frac{n}{\sigma^2}\frac{1}{n-k-2} - n$$
$$=(n+k)\sigma^2\frac{n}{\sigma^2}\frac{1}{n-k-2} - n$$
$$=2(k+1)+\frac{2k^2+6k+4}{n-k-2}$$

that

$$AIC_C(k)=-2\log f(y;\hat\beta,\hat\sigma^2)+2(k+1)+\frac{2k^2+6k+4}{n-k-2}$$

is an unbiased estimator for $E[-2\log f(z;\hat\beta,\hat\sigma^2)]$.

Exercise: Check each step in the derivation of AIC_C. You may use the following facts:

(i) The statistics $\hat{\beta}$ and $\hat{\sigma}^2$ are independent.

(ii) $\hat{\beta} \sim N(\beta, \sigma^2 (X^T X)^{-1})$, $n\hat{\sigma}^2/\sigma^2 \sim \chi^2(n-k)$

(iii) $X \sim \chi^2(j) \Rightarrow E\frac{1}{X} = \frac{1}{j-2}$

Up to now we have never questioned the assumption that the design matrix X containing the regressors is given. In practice, we rarely know a priori which regressors should be included in a regression model and must therefore select the design matrix from a set of candidate matrices. A possible strategy is to select that $n \times k$ matrix which minimizes FPE(k) or $AIC_C(k)$.

While $\hat{\sigma}^2$ can only decrease if additional variables are included, the terms

$$1+\frac{2k}{n-k}$$

and

$$2(k+1)+\frac{2k^2+6k+4}{n-k-2}$$

occurring in FPE(k) and $AIC_C(k)$, respectively, increase as the number of regressors k increases and therefore serve as penalty terms to prevent overparametrization.

An apparent flaw of this model selection approach is that FPE(k) and $AIC_C(k)$ have been derived under the assumption that the mean of y can be written as a linear combination of the columns of X. Why should all candidate matrices satisfy this assumption?

At second glance, model selection with FPE(k) or $AIC_C(k)$ is not so absurd after all, because the chances of selecting a too small (misspecified) model disappear as n increases. So the real challenge is to avoid choosing a too large model. But FPE(k) and $AIC_C(k)$ are particularly suitable for comparing the correct model with larger models, because all of these models are correctly specified.

Exercise: Show that the minimization of

$$AIC_C(k) = -2\log f(y;\hat{\beta},\hat{\sigma}^2) + 2(k+1) + \frac{2k^2+6k+4}{n-k-2}$$

is equivalent to the minimization of

$$n\log\hat{\sigma}^2 + 2(k+1) + \frac{2k^2+6k+4}{n-k-2}.$$

If we ignore the last term occurring in $AIC_C(k)$, which vanishes as n increases, we obtain

$$AIC(k) = -2\log f(y;\hat{\beta},\hat{\sigma}^2) + 2(k+1).$$

Here the penalty term is just two times the number of model parameters. (The parameters in the linear regression model are $\beta_1,\ldots,\beta_k$, and σ^2.)

Exercise: Show that the minimization of

$$FPE(k) = (1+\frac{2k}{n-k})\hat{\sigma}^2$$

is roughly equivalent to the minimization of AIC(k).

Hint: $\log(1+\varepsilon)\approx\varepsilon$

We might expect that

$$\mathrm{AIC}(k)=-2\log f(y;\hat{\beta},\hat{\sigma}^2)+2(k+1)$$

which has been derived as an asymptotically unbiased estimator for

$$E[-2\log f(z;\hat{\beta},\hat{\sigma}^2)]$$

in the framework of the linear regression model

$$y_t=\beta_1 x_{t1}+\ldots+\beta_k x_{tk}+u_t,$$

can also be used when $y=(y_1,\ldots,y_n)^T$ comes from a Gaussian AR(p) model

$$y_t=\phi_1 y_{t-1}+\ldots+\phi_p y_{t-p}+u_t$$

with parameters $\phi_1,\ldots,\phi_p$, and $\sigma^2=\mathrm{var}(u_t)$.

Indeed, if $\hat{\phi}=(\hat{\phi}_1,\ldots,\hat{\phi}_p)^T$ and $\hat{\sigma}^2$ are the ML estimators for the model parameters and $z=(z_1,\ldots,z_n)^T$ is an independent series from the same AR(p) model, then

$$\mathrm{AIC}(p)=-2\log f(y;\hat{\phi},\hat{\sigma}^2)+2(p+1)$$

is an approximately unbiased estimator for

$$E[-2\log f(z;\hat{\phi},\hat{\sigma}^2)].$$

Analogously, in the case of an ARMA(p,q) model

$$y_t=\phi_1 y_{t-1}+\ldots+\phi_p y_{t-p}+u_t+\theta_1 u_{t-1}+\ldots+\theta_q u_{t-q}$$

we may use

$$\mathrm{AIC}(p,q)=-2\log f(y;\hat{\phi},\hat{\theta},\hat{\sigma}^2)+2(p+q+1).$$

The likelihood function for an AR(1) model

Suppose that $y=(y_1,\dots,y_n)^T$ comes from a Gaussian AR(1) model represented by

$$y_t=\phi y_{t-1}+u_t$$

or, equivalently, by

$$y_t=\phi(\phi y_{t-2}+u_{t-1})+u_t=\phi(\phi(\phi y_{t-3}+u_{t-2})+u_{t-1})+u_t=\dots=\sum_{j=0}^{\infty}\phi^j u_{t-j}$$

where $|\phi|<1$ and the errors u_t are i.i.d. $N(0,\sigma^2)$. Then

$$Ey_t=\sum_{j=0}^{\infty}\phi^j Eu_{t-j}=0$$

and for $h\geq 0$

$$\begin{aligned}\gamma(h)&=cov(y_t,y_{t-h})=Ey_ty_{t-h}\\&=E(u_t+\phi u_{t-1}+\phi^2u_{t-2}+\dots)(u_{t-h}+\phi u_{t-h-1}+\phi^2u_{t-h-2}+\dots)\\&=\sigma^2(\phi^h\phi^0+\phi^{h+1}\phi^1+\phi^{h+2}\phi^2+\dots)=\sigma^2\phi^h\sum_{j=0}^{\infty}(\phi^2)^j=\frac{\sigma^2}{1-\phi^2}\phi^h.\end{aligned}$$

The ML estimates are obtained by finding the values of ϕ and σ^2 which maximize

$$f(y_1,\dots,y_n;\phi,\sigma^2)=(2\pi)^{-\frac{n}{2}}(\det\Gamma)^{-\frac{1}{2}}\exp(-\tfrac{1}{2}y^T\Gamma^{-1}y),$$

where $\Gamma=Eyy^T$ depends on ϕ and σ^2.

This maximization problem can only be solved numerically but not analytically.

<u>Exercise:</u> Show that

$$\Gamma=\frac{\sigma^2}{1-\phi^2}\begin{pmatrix} 1 & \phi & \cdots & \phi^{n-1} \\ \phi & 1 & \cdots & \phi^{n-2} \\ \vdots & \vdots & \ddots & \vdots \\ \phi^{n-1} & \phi^{n-2} & \cdots & 1 \end{pmatrix}$$

and

$$\Gamma^{-1}=\frac{1}{\sigma^2}\begin{pmatrix} 1 & -\phi & 0 & \cdots & 0 \\ -\phi & 1+\phi^2 & -\phi & \cdots & 0 \\ 0 & -\phi & 1+\phi^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}.$$

<u>Exercise:</u> Show that $\Gamma^{-1}=L^TL$, where

$$L=\frac{1}{\sigma}\begin{pmatrix} \sqrt{1-\phi^2} & 0 & 0 & \cdots & 0 \\ -\phi & 1 & 0 & \cdots & 0 \\ 0 & -\phi & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}.$$

<u>Exercise:</u> Show that $\det\Gamma=\frac{\sigma^{2n}}{1-\phi^2}$.

The conditional likelihood function

The joint density of a sample $y=(y_1,\ldots,y_n)^T$ from a Gaussian AR(1) model represented by

$$y_t=\phi y_{t-1}+u_t$$

can be written as

$$\begin{aligned} f(y_1,\ldots,y_n) &= f(y_n|y_1,\ldots,y_{n-1})f(y_1,\ldots,y_{n-1}) \\ &= f(y_n|y_1,\ldots,y_{n-1})f(y_{n-1}|y_1,\ldots,y_{n-2})f(y_1,\ldots,y_{n-2}) \\ &\vdots \\ &= f(y_n|y_1,\ldots,y_{n-1})\ldots f(y_2|y_1)f(y_1). \end{aligned}$$

If $u_t\sim N(0,\sigma^2)$, y_{t-1} is fixed, and $y_t=\phi y_{t-1}+u_t$, then

$$y_t\sim N(\phi y_{t-1},\sigma^2).$$

Thus,

$$\begin{aligned} f(y_t|y_1,\ldots,y_{t-1}) &= f(y_t|y_{t-1}) \\ &= (2\pi\sigma^2)^{-\frac{1}{2}}\exp\left(-\frac{1}{2\sigma^2}(y_t-\phi y_{t-1})^2\right) \end{aligned}$$

and

$$\begin{aligned} f(y_n,\ldots,y_2|y_1) &= f(y_n|y_1,\ldots,y_{n-1})\ldots f(y_2|y_1) \\ &= f(y_n|y_{n-1})\ldots f(y_2|y_1) \\ &= (2\pi\sigma^2)^{-\frac{n-1}{2}}\exp\left(-\frac{1}{2\sigma^2}\sum_{t=2}^{n}(y_t-\phi y_{t-1})^2\right). \end{aligned}$$

Exercise: Show that maximizing

$$\log f(y_n,\ldots,y_2|y_1)=-\frac{n-1}{2}\log(2\pi\sigma^2)-\frac{1}{2\sigma^2}\sum_{t=2}^{n}(y_t-\phi y_{t-1})^2$$

gives the ordinary least squares (OLS) estimate

$$\hat{\phi}=\frac{\sum_{t=2}^{n} y_t y_{t-1}}{\sum_{t=2}^{n} y_{t-1}^2}.$$

Multiplying the conditional likelihood function by $f(y_1)$ we obtain the full likelihood function, i.e.,

$$f(y_1,\ldots,y_n)=f(y_n,\ldots,y_2|y_1)f(y_1).$$

It follows from $\mathrm{Var}(y_1)=\frac{\sigma^2}{1-\phi^2}$ that

$$f(y_1)=(2\pi\frac{\sigma^2}{1-\phi^2})^{-\frac{1}{2}}\exp(-\frac{1-\phi^2}{2\sigma^2}y_1^2).$$