JOINT ENTROPY BOUND MINIMIZATION

Zois Boukouvalas

University of Maryland, College Park

1. MAXIMUM ENTROPY DISTRIBUTIONS

The maximum entropy density, subject to known moment constraints, problem can be written as the following optimization problem [**?**]:

$$\begin{aligned}
\max_{p(\mathbf{x})} H(p(\mathbf{x})) &= -\int_{\Omega(\mathbf{x})} p(\mathbf{x}) \log p(\mathbf{x}) d\mathbf{x} \\
s.t. &\int_{\Omega(\mathbf{x})} p(\mathbf{x}) d\mathbf{x} = 1, \\
&\int_{\Omega(\mathbf{x})} r_i(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} = \alpha_i, \text{for } i = 1,\dots,M,
\end{aligned} \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^K$, $p(\mathbf{x}) \geq 0$, with equality outside the support set $\Omega(\mathbf{x})$, $r_i(\mathbf{x}) \in C(\mathbb{R}^K,\mathbb{R}), \forall i$, $C(\mathbb{R}^K,\mathbb{R})$ represents the space of continue functions which have domain $\mathbb{R}^K$ and codomain $\mathbb{R}$. The first constraint need to be $\int_{\Omega(\mathbf{x})} p(\mathbf{x}) d\mathbf{x} = 1$, equivalently $r_0 = 1$ and $\alpha_0 = 1$, in order for $p(\mathbf{x})$ to be a valid pdf. Thus, p is a density on support set $\Omega(\mathbf{x})$ meeting certain moment constraints $\alpha_1,\dots,\alpha_M$. The optimization problem in (1) can be rewritten as a Lagrangian function:

$$L(p) = -\int p \log p + \lambda_0 \left(\int p - 1 \right) + \sum_{i=1}^{M} \lambda_i \left(\int r_i p - \alpha_i \right), \tag{2}$$

where λ_i, $i = 0,\dots,M$, are the Lagrangian multipliers. By setting $\partial L(p)/\partial p(\mathbf{x}) = 0$, we obtain the form of the maximizing density

$$p(\mathbf{x}) = e^{-1+\lambda_0+\sum_{i=1}^{M} \lambda_i r_i(\mathbf{x})}, \tag{3}$$

where Lagrangian multipliers are chosen so that p satisfies the constraints. We can solve for them by the following Newton iteration:

$$\begin{aligned}
\boldsymbol{\lambda}_{t+1} &= \boldsymbol{\lambda}_t - \mathbf{J}^{-1} E_{p_t}\{\mathbf{r} - \boldsymbol{\alpha}\} \\
&= \boldsymbol{\lambda}_t - \left(E_{p_t}\left\{\mathbf{r}\mathbf{r}^T\right\}\right)^{-1} E_{p_t}\{\mathbf{r} - \boldsymbol{\alpha}\},
\end{aligned} \tag{4}$$

where $\mathbf{r} = [r_0,\dots,r_M]^T \in C^{M+1}(\mathbb{R}^K,\mathbb{R})$, $\boldsymbol{\lambda} = [\lambda_0,\dots,\lambda_M]^T \in \mathbb{R}^{M+1}$, $\boldsymbol{\alpha} = [\alpha_0,\dots,\alpha_M]^T \in \mathbb{R}^{M+1}$, and $\mathbf{J} \in \mathbb{R}^{K\times K}$ is the Jacobian matrix. The ijth entry of $\mathbf{J}$ is

$$\begin{aligned}
\mathbf{J}_{ij} &= \frac{\partial \int_{\Omega(\mathbf{x})} r_i(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} - \alpha_i}{\partial \lambda_j} \\
&= \int_{\Omega(\mathbf{x})} r_i(\mathbf{x}) \frac{\partial p(\mathbf{x})}{\partial \lambda_j} d\mathbf{x} \\
&= \int_{\Omega(\mathbf{x})} r_i(\mathbf{x}) r_j(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \\
&= E_{p_t}\left\{r_i r_j\right\}
\end{aligned} \tag{5}$$

The ith entry of $E_{p_t}\{\mathbf{r} - \boldsymbol{\alpha}\}$ is

$$E_{p_t}\{\mathbf{r} - \boldsymbol{\alpha}\}_i = E_{p_t}\{r_i - \alpha_i\} = \int_{\Omega(\mathbf{x})} r_i(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} - \alpha_i \tag{6}$$

The entropy corresponding to this density function will be

$$
\begin{aligned}
H(p) &= -\int p \log p \\
&= -\int \left(-1 + \lambda_0 + \sum_{i=1}^{M} \lambda_i r_i\right) p \\
&= 1 - \lambda_0 - \sum_{i=1}^{M} \lambda_i \alpha_i. \qquad (7)
\end{aligned}
$$

2. JOINT ENTROPY BOUND FOR RANDOM VECTOR

2.1. Constrained functions

In order to evaluate the joint entropy bound for random vector $\mathbf{s} \in \mathbb{R}^K$, assuming all the marginals have the similar distribution, we can use maximum entropy distributions method. By selecting a set of constraints, we can obtain the maximum entropy distribution which satisfy these constraints, as in [**?**]. Before evaluating the maximum entropy distribution, it will make the problem easier by whitening the data, as in [**?**].

$$
\begin{aligned}
H(\mathbf{s}) &= H(\mathbf{x}) - \log|\det(\boldsymbol{\Sigma}^{-1/2})| \\
&= H(\mathbf{x}) + \frac{1}{2}\log(\det(\boldsymbol{\Sigma})), \qquad (8)
\end{aligned}
$$

where $\boldsymbol{\Sigma} = E\{\mathbf{s}\mathbf{s}^T\}$ is the covariance matrix of $\mathbf{s}$, and $\mathbf{x} = \boldsymbol{\Sigma}^{-1/2}\mathbf{s}$. Now, we focus on the joint entropy bound on $\mathbf{x}$. One way to choose the constraints is

- 1, $\int_{\Omega(\mathbf{x})} p(\mathbf{x}) d\mathbf{x} = 1$: to make sure $p(\mathbf{x})$ is a valid pdf.
- 2, $\int_{\Omega(\mathbf{x})} x_i p(\mathbf{x}) d\mathbf{x} = 0, \forall i$: to make sure $E\{\mathbf{x}\} = \mathbf{0}$.
- 3, $\int_{\Omega(\mathbf{x})} x_i x_j p(\mathbf{x}) d\mathbf{x} = \delta_{i-j}, \forall i, j$: to make sure $E\{\mathbf{x}\mathbf{x}^T\} = \mathbf{I}$.
- 4, $\int_{\Omega(\mathbf{x})} g_i(x_i) p(\mathbf{x}) d\mathbf{x} = \alpha_{g_i}, \forall i$: to have better fit on the marginals.
- 5, $\int_{\Omega(\mathbf{x})} q(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} = \alpha_q$: to capture the higher order dependency among $\mathbf{x}$,

where $g_i \in C(\mathbb{R}, \mathbb{R}), \forall i$ are some nonlinear functions as in [**?**], and $q \in C(\mathbb{R}^K, \mathbb{R})$ is some nonlinear function which depends on at the entries of $\mathbf{x}$ jointly. There are $1 + K + K^2 + K + 1 = K^2 + 2K + 2$ constraints, which means that there are too many parameters need to be estimated. However, after some thought, most of the constraints are unnecessary. We will choose the following constraints instead

- 1, $\int_{\Omega(\mathbf{x})} p(\mathbf{x}) d\mathbf{x} = 1$
- 2, $\int_{\Omega(\mathbf{x})} \mathbf{1}^T \mathbf{x} p(\mathbf{x}) d\mathbf{x} = 0$
- 3, $\int_{\Omega(\mathbf{x})} \mathbf{x}^T \mathbf{x} p(\mathbf{x}) d\mathbf{x} = K$
- 4, $\int_{\Omega(\mathbf{x})} \mathbf{1}^T (\mathbf{x}\mathbf{x}^T - \mathbf{I}) \mathbf{1} p(\mathbf{x}) d\mathbf{x} = 0$
- 5, $\int_{\Omega(\mathbf{x})} \mathbf{1}^T \mathbf{g}(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} = \alpha_g$
- 6, $\int_{\Omega(\mathbf{x})} q(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} = \alpha_q$

where α_g and α_q can be evaluated by using sample average, $\mathbf{g} = [g_1, \ldots, g_K]^T \in C^K(\mathbb{R}, \mathbb{R})$, and $q \in C(\mathbb{R}^K, \mathbb{R})$. These constraints can be showed to be a approximation to the previous one if $g_i \forall i$ are the same function and q is the symmetric function with respect to all entries of $\mathbf{x}$. For any subset of random vector $\mathbf{x}$, $\tilde{\mathbf{x}}$ and $\check{\mathbf{x}}$, which have same dimension, will have same distribution since

$$
p(\tilde{\mathbf{x}}) = \int_{\Omega(\tilde{\mathbf{x}}^c)} p(\mathbf{x}) d\tilde{\mathbf{x}}^c = \int_{\Omega(\check{\mathbf{x}}^c)} p(\mathbf{x}) d\check{\mathbf{x}}^c = p(\check{\mathbf{x}})
$$

where $\tilde{\mathbf{x}}^c$ and $\check{\mathbf{x}}^c$ are the complement set for $\tilde{\mathbf{x}}$ and $\check{\mathbf{x}}$, respectively. Hence, all the marginals have the same statistic. Therefore, from constraint 2, 3 and 4, we know that all the marginals are zero mean, unit variance, and all the covariance are zero. In addition, we know the integral of the distribution of any subset of x is equal to 1 since

$$\int_{\Omega(\tilde{\mathbf{x}})} p(\tilde{\mathbf{x}})\, d\tilde{\mathbf{x}} = \int_{\Omega(\tilde{\mathbf{x}})} \int_{\Omega(\tilde{\mathbf{x}}^c)} p(\mathbf{x})\, d\tilde{\mathbf{x}}^c\, d\tilde{\mathbf{x}} = \int_{\Omega(\mathbf{x})} p(\mathbf{x})\, d\mathbf{x} = 1$$

Few comments on this set of constraints:

- The first four constraints are make sure that it is a valid pdf, zero mean and identity covariance matrix.
- It will end up with a multivariate Gaussian if we only use first four constraints.
- The fifth constraint is added in order to capture the non-Gaussianity.
- It will end up with a multivariate distribution with independent marginals if we only use first five constraints since we can always write the joint pdf as a product of all marginal pdfs.
- The last constraint is added in order to capture the higher order dependency across the entries of $\mathbf{x}$.
- All the marginals will have the same distribution since we choose same g for all marginals and q is symmetric with respect to all entries of $\mathbf{x}$.
- The main difficulty of maximum entropy multivariate density estimation is the evaluation of high dimensional integral when using Newton iteration to solve for Lagrangian multipliers.

We can choose the constrained function g as in [**?**]. In order to evaluate the multidimensional integral easily, constrained function $q(\mathbf{x})$ can be define as following,

$$q(\mathbf{x}) = \begin{cases} 1, & \text{if } \max_i |x_i| < \gamma,\ \Omega_\gamma \\ 0, & \text{otherwise, } \Omega_\gamma^c \end{cases} \tag{9}$$

Let $p = e^{\lambda_5 q}\tilde{p}$. Then, for Newton iteration, the integral could be calculated as following

$$\begin{aligned} E\{r_i r_j\} &= \int_\Omega r_i r_j p \\ &= \int_{\Omega_\gamma} r_i r_j p + \int_{\Omega_\gamma^c} r_i r_j p \\ &= \int_{\Omega_\gamma} r_i r_j e^{\lambda_5 1}\tilde{p} + \int_{\Omega_\gamma^c} r_i r_j e^{\lambda_5 0}\tilde{p} \\ &= e^{\lambda_5}\int_{\Omega_\gamma} r_i r_j \tilde{p} + \left(\int_\Omega r_i r_j \tilde{p} - \int_{\Omega_\gamma} r_i r_j \tilde{p}\right) \\ &= \int_\Omega r_i r_j \tilde{p} - (1 - e^{\lambda_5})\int_{\Omega_\gamma} r_i r_j \tilde{p} \end{aligned} \tag{10}$$

where both integrals can be separate into product of a lot of one dimensional integrals.

2.2. The localized constraints

2.2.1. Univariate case

The meaning of the constraints:

- 0, $\int_{\Omega(x)} p(x)dx = 1$: to make sure $p(x)$ is a valid pdf.
- 1, $\int_{\Omega(x)} xp(x)dx = 0$: to make sure $E\{x\} = 0$.
- 2, $\int_{\Omega(x)} x^2 p(x)dx = 1$: to make sure $E\{x^2\} = 1$.
- 3, $\int_{\Omega(x)} g(x)p(x)dx = \alpha_g$: to capture the HOS.
  For example, $E\{x^4\} = \alpha_g$ when $g(x) = x^4$. (Globalized constraint?)
- 4, $\int_{\Omega(x)} q(x)p(x)dx = \alpha_q$: to capture localized information of PDF.
  For example, $p(x \in \Omega_q) = \alpha_q = T_q/T$ when $q(x) = 1$ for $x \in \Omega_q$ and $q(x) = 0$ otherwise is a step function, and T_q is the number of samples belonged to Ω_q.

Examples of using $r_0 = 1$ and a step function r_4.

- Use $r_0 = 1$: $p(\mathbf{x}) = e^{-1+\lambda_0+\sum_{i=1}^{M}\lambda_i r_i(\mathbf{x})}$
- Use $r_0 = 1$ and a step function q: $p(\mathbf{x}) = e^{-1+\lambda_0+\lambda_1 q(x)}$

Maximum entropy distributions: globalized constraints and localized constraints (histogram method).

Other important points about localized constraints:

- Use Gaussian kernel
- Use multiple localized constraints
- Any function can be approximated by Taylor series, step functions, or Gaussian kernels. But using step functions or Gaussian kernels, there is no stability issue for super-Gaussian data.

2.2.2. Multivariate case

- $q(\mathbf{x})$: capture the localized information of PDF and higher order dependence among entries of $\mathbf{x}$
- Unfortunately, we cannot use Gaussian kernel, since the multidimensional integral cannot be separated into product of one dimensional integrals.
  $p(\mathbf{x}) = e^{-1+\lambda_0+\lambda_1\sum_{k=1}^{K} x_k+\lambda_2\sum_{k=1}^{K} x_k^2+\lambda_3\sum_{k=1}^{K} g(x_k)++\lambda_4 q(\mathbf{x})}$

2.3. Solve for Lagrangian multipliers

Since in the Newton iterations, we need to calculate a lot of multidimensional integrals, it could be very time consuming. In addition, there could be convergence problems dependent on the chosen initialization. In [**?**], authors proposed a computationally simple way to calculate the Lagrange multipliers by replacing the Newton iterations by linear equations (LE) for univariate case. But, this method only yields an approximate solution. In [**?**], the authors proposed to combine this LE method and Newton iterations to find Lagrange multipliers in the univariate case. Here we extend the LE method to the multivariate case and use its solution as the initial value for our Newton iterations. For the ith constrained function,

$$\int_{\Omega(\mathbf{x})} r_i(\mathbf{x})p(\mathbf{x})d\mathbf{x} = \alpha_i$$

Applying integration by parts with the following definitions,

$$
\begin{aligned}
u &= p(\mathbf{x}) \\
\nabla u &= \left[\frac{\partial p(\mathbf{x})}{\partial x_1}, \ldots, \frac{\partial p(\mathbf{x})}{\partial x_K}\right]^T = \left[\sum_{m=1}^{M} \lambda_m \frac{\partial r_m(\mathbf{x})}{\partial x_1}, \ldots, \sum_{m=1}^{M} \lambda_m \frac{\partial r_m(\mathbf{x})}{\partial x_K}\right]^T p(\mathbf{x}) \\
\mathbf{v} &= \frac{1}{K}\left[\int_{-\infty}^{x_1} r_i(\mathbf{z}) dz_1, \ldots, \int_{-\infty}^{x_K} r_i(\mathbf{z}) dz_K\right]^T \triangleq [R_{i1}, \ldots, R_{iK}]^T \\
\nabla \cdot \mathbf{v} &= \sum_{k=1}^{K} \frac{\partial R_{ik}}{\partial x_k} = r_i(\mathbf{x})
\end{aligned}
$$

The integration by parts for multivariate case:

$$
\int_\Omega u \nabla^T \mathbf{v} = \int_{\partial\Omega} u \mathbf{v}^T \mathbf{n} - \int_\Omega \nabla u^T \mathbf{v}
$$

Hence,

$$
\begin{aligned}
\alpha_i &= \int_\Omega r_i(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \\
&= \int_{\partial\Omega} p(\mathbf{x}) \mathbf{v}^T \mathbf{n} d\mathbf{x} - \int_\Omega \left(\sum_{m=0}^{M} \lambda_m \sum_{k=1}^{K} \frac{\partial r_m(\mathbf{x})}{\partial x_k} R_{ik}\right) p(\mathbf{x}) dx \\
&= -\sum_{m=0}^{M} \lambda_m \int_\Omega \left(\sum_{k=1}^{K} \frac{\partial r_m(\mathbf{x})}{\partial x_k} R_{ik}\right) p(\mathbf{x}) dx \qquad \left[\int_{\partial\Omega} p(\mathbf{x}) \mathbf{v}^T \mathbf{n} d\mathbf{x} = 0\right] \\
&= -\sum_{m=0}^{M} \lambda_m E_{p_\lambda} \left\{\sum_{k=1}^{K} \frac{\partial r_m(\mathbf{x})}{\partial x_k} R_{ik}\right\} \\
&\approx -\sum_{m=0}^{M} \lambda_m \beta_{im}
\end{aligned}
$$

where β_{im} is the sample average of $\sum_{k=1}^{K} R_{ik} \partial r_m(\mathbf{x})/\partial x_k$. Then, the Lagrange multipliers are given by the solution of the following LE:

$$
\boldsymbol{\lambda} = -\boldsymbol{\beta}^{-1} \boldsymbol{\alpha}
$$

where the imth entry of $\boldsymbol{\beta}$ is β_{im}.

3. ON THE EQUIVALENCE OF DATA-DRIVEN MAXIMUM ENTROPY AND MAXIMUM LIKELIHOOD

3.1. Maximum Likelihood: ML

Now that the maximum entropy distribution has been fixed, finding the optimal parameters in a data-driven sense, is equivalent to solving the maximum likelihood equations, as we will show next. We choose $p_X(\mathbf{x}) = \exp(-1 + \lambda_0 + \sum_i \lambda_i r_i(\mathbf{x}))$, which corresponds to maximum entropy problem with constraints $\{r_i\}_{i=1}^M$, as the distribution family for maximum likelihood problem. Now, we will have a closer look at the single constraint that requires the integral of the density function to be one, i.e.,

$$
\begin{aligned}
&\int \exp(-1 + \lambda_0 + \sum_{i=1}^{M} \lambda_i r_i(\mathbf{x})) d\mathbf{x} = \exp(\lambda_0) \int \exp(-1 + \sum_{i=1}^{M} \lambda_i r_i(\mathbf{x})) d\mathbf{x} = 1 \\
\Leftrightarrow &\lambda_0 = -\log \int \exp(-1 + \sum_{i=1}^{M} \lambda_i r_i(\mathbf{x})) d\mathbf{x}
\end{aligned}
$$

the latter which is a function of all other $\{\lambda_i\}_{i=1}^M$, but not of $\mathbf{x}$. Now, plugging λ_0 back into $p_X(\mathbf{x})$ and putting the estimating equations of the maximum log-likelihood yield:

$$
\begin{aligned}
&\frac{\partial}{\partial\lambda_j}\log\left(\prod_{t=1}^T \exp\left(-1-\log\int \exp(-1+\sum_{i=1}^M \lambda_i r_i(\mathbf{x}))d\mathbf{x}+\sum_{i=1}^M \lambda_i r_i(\mathbf{x}(t))\right)\right)=0\\
&\Leftrightarrow \sum_{t=1}^T\left(-\frac{\int r_j(\mathbf{x})\exp(-1+\sum_{i=1}^M \lambda_i r_i(\mathbf{x}))d\mathbf{x}}{\int \exp(-1+\sum_{i=1}^M \lambda_i r_i(\mathbf{x}))d\mathbf{x}}+r_j(\mathbf{x}(t))\right)=0\\
&\Leftrightarrow \sum_{t=1}^T r_j(\mathbf{x}(t)) = T\frac{\int r_j(\mathbf{x})\exp(-1+\sum_{i=1}^M \lambda_i r_i(\mathbf{x}))d\mathbf{x}}{\int \exp(-1+\sum_{i=1}^M \lambda_i r_i(\mathbf{x}))d\mathbf{x}} && \left[\frac{\int r_j(\mathbf{x})\exp(-1+\sum_{i=1}^M \lambda_i r_i(\mathbf{x}))d\mathbf{x}}{\int \exp(-1+\sum_{i=1}^M \lambda_i r_i(\mathbf{x}))d\mathbf{x}}\ \begin{array}{c}\text{constant}\\ \text{w.r.t. }\mathbf{x}\text{(t)}\end{array}\right]\\
&\Leftrightarrow \frac{\sum_{t=1}^T r_j(\mathbf{x}(t))}{T} = \int r_j(\mathbf{x})\exp(-1+\lambda_0+\sum_{i=1}^M \lambda_i r_i(\mathbf{x}))d\mathbf{x} && \left[\exp(-\lambda_0)=\int \exp(-1+\sum_{i=1}^M \lambda_i r_i(\mathbf{x}))d\mathbf{x}\right]\\
&\Leftrightarrow \widehat{E}\{r_j(\mathbf{x})\}=E\{r_j(\mathbf{x})\}
\end{aligned}
$$

and the latter are the α_j in the maximum entropy solution. Hence, by choosing $\alpha_j = \widehat{E}\{r_j(\mathbf{x})\}$, the maximum likelihood estimates $\{\widehat{\lambda}_j\}_{j=1}^I$ are also the solutions to Equation (1).

3.2. Exponential family

Actually, the form of the distribution $p_X(\mathbf{x}) = \exp(-1+\lambda_0+\sum_i \lambda_i r_i(\mathbf{x}))$, is a natural form of the exponential family, usually expressed as

$$p_X(\mathbf{x}|\boldsymbol{\eta}) = \gamma(\mathbf{x})\exp\left(\boldsymbol{\eta}^T\mathbf{T}(\mathbf{x}) - A(\boldsymbol{\eta})\right)$$

where $T_i(\mathbf{x})$ are the sufficient statistics for the distribution. Making the analogy with our case, we have $A(\boldsymbol{\eta}) = 1-\lambda_0$. This is the normalization factor of the distribution and is fixed once all other functions have been chosen. $\boldsymbol{\eta} = \boldsymbol{\lambda}$ are called the natural parameters and are invariant under any reparametrization of the distribution. They are the only parameters for which the distribution can be written as an exponential of their linear combination. The functions that depend on the random variables are $T_i(\mathbf{x}) = r_i(\mathbf{x})$, known as the sufficient statistics of the distribution and $\gamma(\mathbf{x})$, called the

$$dF_X(\mathbf{x}|\boldsymbol{\eta}) = \exp\left(\boldsymbol{\eta}^T\mathbf{T}(\mathbf{x}) - A(\boldsymbol{\eta})\right)d\Gamma(\mathbf{x})$$

in the Lebesgue-Stieltjes sense. For a continuous distribution we have $d\Gamma(\mathbf{x}) = \gamma(\mathbf{x})d\mathbf{x}$.

3.3. Schematic representation

$$
\begin{array}{ccccc}
\text{exponential family} & & \text{ML} & & \text{most likely distribution}\\
\Updownarrow & \Longrightarrow & \Updownarrow & \Longleftrightarrow & \Updownarrow\\
\text{empirical statistical constraints} & & \text{MaxEnt} & & \text{fewest assumptions on distribution}
\end{array}
$$

A. REVIEW OF MULTIVARIATE CALCULUS

In what follows, u, w represent scalar functions and $\mathbf{v} = (v_1, v_2, v_3)$ represents a vector function. n represents the unit outward normal vector, to the boundary $\partial\mathbb{R}$.

$$\text{gradient } u \equiv \nabla u = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}\right)u = \left(\frac{\partial u}{\partial x}, \frac{\partial u}{\partial y}, \frac{\partial u}{\partial z}\right)$$

$$\text{divergence } v \equiv \nabla\cdot\mathbf{v} = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}\right)\cdot\mathbf{v} = \frac{\partial v_1}{\partial x} + \frac{\partial v_2}{\partial y} + \frac{\partial v_3}{\partial z}$$

Note: in the book, $\nabla\cdot\mathbf{v}$ is also written as