Problem 1 (graded by Dunzhu) 30 points

(a) - 5 points

Here

$$F(m) = \sum_i \frac{(d_i - g_i(m))^2}{2\sigma_d^2}$$

$$P(m|d) \propto P(d|m) \propto \exp\left(-F(m)\right) \approx \exp\left(-F(m_0) - \frac{\partial F}{\partial m}^T (m - m_0) - \frac{1}{2}(m - m_0)^T \frac{\partial^2 F}{\partial m \partial m}(m - m_0)\right)$$

Above equation has the form of a multivariate Gaussian distribution, So the covariance matrix will be

$$cov(m) = \left(\frac{\partial^2 F}{\partial m \partial m}|_{m=m_0}\right)^{-1} = H^{-1}$$

Note we want to know covariance matrix when m_0 is the best least square model. Using hw2's convention, define

$$\hat{G}_{i,k} = \frac{\partial g_i}{\partial m_k}$$

Now

$$H \approx \frac{\hat{G}^T \hat{G}}{\sigma_d^2}$$

```
function hw4_p1()
xi=[0 10 15 6 -7 3]';
yi=[0 0 6 13 10 7]';
zi=[0 0 0 0 0 0]';
ti=[322.418 321.031 321.228 323.093 324.415 322.706]';

[M0, hess0]=do_one_ti(xi, yi, zi, ti);
sigma = 0.01;
C0=inv(hess0/sigma^2)


function [M, hess]=do_one_ti(xi, yi, zi, ti)

M=[315 30 -17 15 5]';
for step = 1:1000
    grad=zeros(5,1);
    hess=zeros(5,5);
    for i=1:length(xi)
        [tmp_grad, tmp_hess]=get_grad_hessian(xi(i), yi(i), zi(i), ti(i), M);
        grad = grad + tmp_grad;
        hess = hess + tmp_hess;
    end
    err(step) = norm(predict(xi, yi, zi, M)-ti);
    if(step > 1 && err(step) > err(step-1))
        break;
```

```
    end
        M = M- inv(hess)*grad;
end

function [grad,hess] = get_grad_hessian(xi,yi,zi,ti,M)
ts=M(1); xs=M(2); ys=M(3); zs=M(4); v=M(5);
R=sqrt((xs-xi)^2 + (ys-yi)^2 + (zs-zi)^2);
nx=(xs-xi)/R;
ny=(ys-yi)/R;
nz=(zs-zi)/R;
e=(ts+R/v-ti);

grad = e*[1 nx/v ny/v nz/v -R/v^2]';
sen = [1 nx/v ny/v nz/v -R/v^2]';
hess = sen*sen'; % approximate hess

function r=predict(xi,yi,zi,M)

ts=M(1); xs=M(2); ys=M(3); zs=M(4); v=M(5);
r=xi*0;
for i=1:length(xi)
    R=sqrt((xs-xi(i))^2 + (ys-yi(i))^2 + (zs-zi(i))^2);
    r(i)=(ts+R/v);
end
```

the output is

```
C0 =

    0.3651   -0.8059    0.7507   -1.3922    0.0480
   -0.8059    2.8090   -2.5992    1.9509   -0.0010
    0.7507   -2.5992    2.4548   -1.7962    0.0010
   -1.3922    1.9509   -1.7962    6.5912   -0.2971
    0.0480   -0.0010    0.0010   -0.2971    0.0171
```

(b)- 5 points

The covariance matrix calculated here will be similar to the Monte Carlo simulation in HW2. They should be more and more similar when $\sigma_d \to 0$.

Diagonal element shows the variance of each parameter. For example, we can notice $\sigma_t^2 = 0.3651$, $\sigma_x^2 = 2.8099 \approx \sigma_y^2 = 2.4548$. σ_z is much larger than σ_x and σ_y.

The off diagonal shows trade off between parameters. For example, $\sigma_{xy} = -2.5992$, explains the linear shape of (x_s, y_s)in HW2. Also note $\sigma_{zt} = -1.3922$ indicate strong trade off between z_s and t_s.

Note also it's common to change the covariance matrix to correlation matrix. For example, $\rho_{xy} = \sigma_{xy}/(\sigma_x \sigma_y)$. In this case

```
S=diag(sqrt(diag(C0)));
rho = S^(-1)*C0*S^(-1)
rho =
    1.0000   -0.7958    0.7929   -0.8975    0.6065
```

```
-0.7958    1.0000   -0.9898    0.4534   -0.0046
 0.7929   -0.9898    1.0000   -0.4465    0.0047
-0.8975    0.4534   -0.4465    1.0000   -0.8840
 0.6065   -0.0046    0.0047   -0.8840    1.0000
```

the correlation matrix shows $\rho_{zv} = -0.8840$, also a strong tradeoff.

(c)- 5 points

Now

$$F(m) = \frac{1}{2\sigma^2}(d - Gm)^T(d - Gm)$$

where $\sigma = 40$, and

$$G = [ones(length(x), 1), x]$$
$$m = [m1, m2]^T$$

So the Hessian is

$$H = \frac{1}{\sigma^2}G^T G$$

and thus

$$cov(m) = H^{-1}$$

```
G=[ones(length(x),1), x(:)];
sigma=40;
cov0 = inv(G'*G/sigma^2)

cov0 =

       164.83        -3.084
       -3.084       0.11213
```

(d)- 5 points

Since in this case, we assume uniform prior, then $P(m|d) \propto P(d|m)$. The full parameter space calculate the likelihood $P(d|m)$, while the covariance is calculated from distribution $P(m|d)$. So the error map in full-parameter-space in HW2, forms an ellipsoid, whose axis length and rotation angle is all determined by this covariance matrix. For example, the box containing the error ellipsoid has height/width ratio about $(500 - 100)/(2 - (-8)) \approx 4/1$, which is similar to $\sqrt{164.83/0.11213}$. The off diagnoal is negative, explains the negative slope in HW2.

(e)- 5 points

First calculate the misfit $F(m)$ for different models. Then the likelihood $P(d|m_1, m_2)$ can be calculated using $\exp(-F(m))$. Then the $P(m_1, m_2|d)$ follows from Bayesian law and uniform prior. Then we can integrate $P(m_1, m_2|d)$ to get the marginal pdf. Finally, we need to make sure the result is normallized properly (it's indeed a pdf, whose integration is 1).

$$P(m_1, m_2|d) \propto \exp(-F(m))$$
$$P(m_1, m_2|d) = \frac{\exp F(m)}{\int dm_1 \int dm_2 \exp F(m)}$$

and marginal

$$P(m_1|d) = \int dm_2 \left(\frac{\exp F(m)}{\int dm_1 \int dm_2 \exp F(m)} \right)$$
$$P(m_2|d) = \int dm_1 \left(\frac{\exp F(m)}{\int dm_1 \int dm_2 \exp F(m)} \right)$$

We can use summation to approximate the integration.

```
minmis = min(misfit2(:));  % minimum misfit
Pd_m = exp(- (misfit2-minmis) / (2*sigma^2)  ); % likelihood, without normalize

Pd_m = Pd_m / sum(Pd_m(:));
Pm1_d = sum(Pd_m,2)/(m1(2)-m1(1));   %% now int (Pm1_d) dm1 = 1
Pm2_d = sum(Pd_m,1)/(m2(2)-m2(1));

subplot(121);
plot(m1,Pm1_d);
subplot(122);
plot(m2,Pm2_d);
```

(f)- 5 points

$$\sigma_{m1}^2 = E[m_1^2] - (E[m_1])^2 = \int p(m_1|d) m_1^2 dm_1 - \left(\int p(m_1|d) m_1 dm_1 \right)$$

```
Em1 =  sum(Pm1_d(:) .* m1(:) * (m1(2)-m1(1)) ) ;
Em1m1 = sum(Pm1_d(:) .* m1(:).^2 * (m1(2)-m1(1)) ) ;
Em2 =  sum(Pm2_d(:) .* m2(:) * (m2(2)-m2(1)) ) ;
Em2m2 = sum(Pm2_d(:) .* m2(:).^2 * (m2(2)-m2(1)) ) ;

sigma_m1 = sqrt(Em1m1-Em1*Em1)
sigma_m2 = sqrt(Em2m2-Em2*Em2)

sigma_m1 =

       12.838


sigma_m2 =

      0.33485
```

Note that σ_{m1}, σ_{m2} is equal to square root of the diagnoal of the covirance matrix calculated in (c).

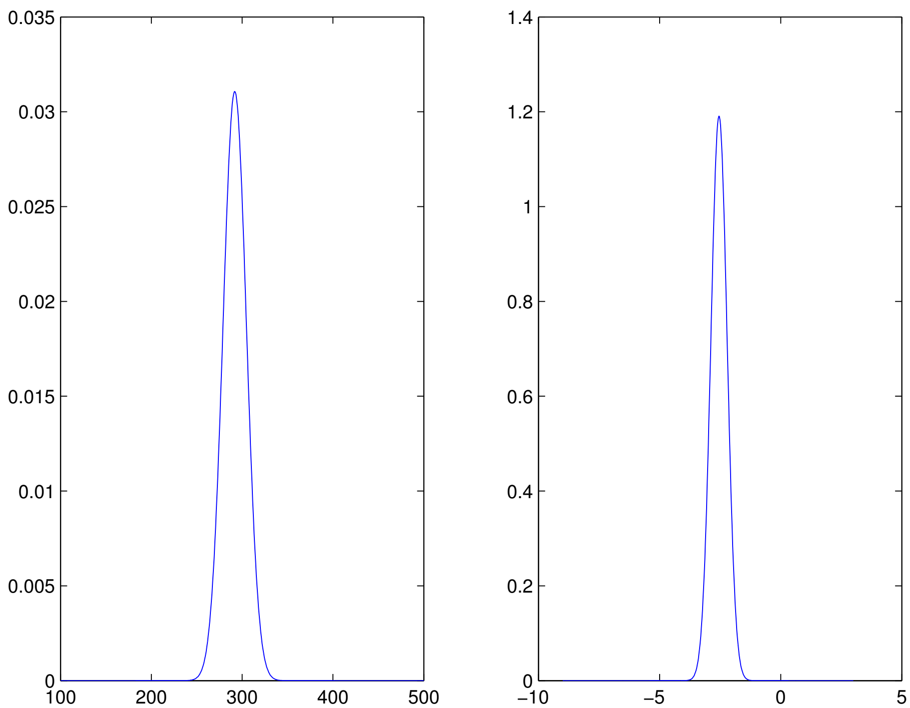

Figure 1: (e): Marginal pdf for m_1 and m_2.

Problem 2 (graded by Toby) - 15 points

(a)

The posterior probability is given by

$$p(\mu|\mathbf{x}) \propto \left(\sum_{k=1}^{N}(x_k-\mu)^2\right)^{-\frac{N-1}{2}} = \exp\left(-\frac{N-1}{2}\ln\left[\sum_{k=1}^{N}(x_k-\mu)^2\right]\right). \tag{1}$$

Minima of F maximize the probability density function. We have that

$$F(\mu) = \frac{N-1}{2}\ln\left(\sum_{k=1}^{N}(x_k-\mu)^2\right), \tag{2a}$$

$$\frac{\partial F}{\partial \mu} = -(N-1)\frac{\sum_{k=1}^{N}(x_k-\mu)}{\sum_{k=1}^{N}(x_k-\mu)^2} \tag{2b}$$

$$\frac{\partial^2 F}{\partial \mu^2} = (N-1)\frac{N\sum_{k=1}^{N}(x_k-\mu)^2 + 2\sum_{k=1}^{N}(x_k-\mu)}{\left(\sum_{k=1}^{N}(x_k-\mu)^2\right)^2}. \tag{2c}$$

Because exp and ln are monotonic functions, we therefore need to find the minimum of the sum of the equared error (least-squares)

$$\sum_{k=1}^{N}(x_k-\mu)^2 = 0. \tag{3}$$

Taking the derivative with respct to μ and setting it to zero yields the minimum μ_0

$$\mu_0 = \frac{1}{N}\sum_{k=1}^{N}x_k, \tag{4}$$

which is the sample mean of the data. For the standard deviation σ_μ we expand F to second order and insert μ_0, i.e.,

$$p(\mu|\mathbf{x}) \propto e^{-F(\mu)} \approx e^{-F(\mu_0)-\frac{1}{2}F''(\mu_0)(\mu-\mu_0)^2}, \tag{5}$$

For the second derivate at the best fit solution we have

$$\frac{\partial^2 F}{\partial \mu^2} = \frac{N(N-1)}{\sum_{k=1}^{N}(x_k-\mu)^2}. \tag{6}$$

The standard deviation σ_μ then amounts to

$$\sigma_\mu = \frac{1}{F''(\mu_0)^{1/2}} = \frac{S}{\sqrt{N}}, \tag{7}$$

where

$$S = \sqrt{\frac{1}{N-1}\sum_{k=1}^{N}(x_k-\mu_0)^2} \tag{8}$$

is the sample variance of the data.

(b)

Formal general proof

We prove that an orthogonal matrix $\mathbf{Q}$ does not impact the norm of a vector $\mathbf{x}$. Then $\mathbf{Q}$ is called an isometry. If the length of an arbitrary vector is preserved, $\mathbf{Q}$ must be a rotation matrix, because only rotations leave the length of vector unchanges. To prove this, we use the $\mathcal{L}_2$-norm of an arbitrary vector x and the fact that for orthogonal matrices, we have $\mathbf{Q}^{-1} = \mathbf{Q}^T$. We can then write

$$\|\mathbf{x}\|_2 = \mathbf{x}^T\mathbf{x} = \mathbf{x}^T\mathbf{1}\mathbf{x} = \mathbf{x}^T\mathbf{Q}^{-1}\mathbf{Q}\mathbf{x} = \mathbf{x}^T\mathbf{Q}^T\mathbf{Q}\mathbf{x} = (\mathbf{Q}\mathbf{x})^T(\mathbf{Q}\mathbf{x}) = \|\mathbf{Q}\mathbf{x}\|_2. \tag{9}$$

Since the left hand side and the right hand side are equal, we have shown that the lengths are equal and unchanged under orthogonal transformations.

Simple proof in 2D

In two dimensions, we have that $\mathbf{Q} = (\mathbf{e}_1\ \mathbf{e}_2)$, where $\mathbf{e}_i$ are the column vectors of $\mathbf{Q}$. They are orthonormal. Now, operating $\mathbf{Q}$ on the column vector $\mathbf{x} = (x_1,\, x_2)^T$ gives

$$\mathbf{x}' = \mathbf{Q}\mathbf{x} = x_1\mathbf{e}_1 + x_2\mathbf{e}_2, \tag{10}$$

which is the representation of the vector $\mathbf{x}'$ in the eigenbasis of $\mathbf{H}$. Because the $\mathbf{e}_i$ have length 1, the vector $\mathbf{x}'$ has length $\sqrt{x_1^2 + x_2^2}$. But this is the length of $\mathbf{x}$ as well. So $\mathbf{x}$ and $\mathbf{x}'$ are of the same magnitude. Formally, we have

$$\|\mathbf{x}'\|_2 = \mathbf{x}'^T\mathbf{x}' = x_1^2\mathbf{e}_1^T\mathbf{e}_1 + x_2^2\mathbf{e}_2^T\mathbf{e}_2 + 2x_1x_2\mathbf{e}_1^T\mathbf{e}_2 = x_1^2 + x_2^2 = \mathbf{x}^T\mathbf{x} = \|\mathbf{x}\|_2. \tag{11}$$

This is exactly (!) the same logic as in the general proof but applied explicitly for the two-dimensional problem.

Simple explanation

Because $\mathbf{x}^T\mathbf{H}\mathbf{x} = \text{const.}$ defines an ellipse when $\mathbf{H}$ is symmetric, where the eigenvectors of $\mathbf{H}$ define the principle axes of the ellipse, we can think of $\mathbf{Q}\mathbf{x}$ as a rotation into a coordinate system aligned with these principle axes. This can be seen from $\mathbf{x}'^T\mathbf{\Lambda}\mathbf{x}' = \text{const}$, which also defines an ellipse, but with the principle axes aligned with the coordinate axes (since $\mathbf{\Lambda}$ is diagonal).

(c)

We need to show that $\sigma_x^2 = \frac{B}{AB-C^2}$. We are deadling with an integral of the form

$$\sigma_x^2 = \frac{\int dxdy\, x^2 \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x}\right)}{\int dxdy\, \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x}\right)} \tag{12a}$$

$$= \frac{\int \mathrm{d}x\, \mathrm{d}y\, x^2 \exp\left(-\frac{1}{2}(Ax^2 + 2Cxy + By^2)\right)}{\int \mathrm{d}x\, \mathrm{d}y\, \exp\left(-\frac{1}{2}(Ax^2 + 2Cxy + By^2)\right)} \tag{12b}$$

$$= \frac{\int \mathrm{d}x\, \mathrm{d}y\, x^2 \exp\left(-\frac{1}{2}(A - C^2/B)x^2 - \frac{1}{2}B(y + Cx/B)^2\right)}{\int \mathrm{d}x\, \mathrm{d}y\, \exp\left(-\frac{1}{2}(A - C^2/B)x^2 - \frac{1}{2}B(y + Cx/B)^2\right)} \tag{12c}$$

$$= \frac{\int \mathrm{d}x\, x^2 \exp\left(-\frac{1}{2}(A - C^2/B)x^2\right)}{\int \mathrm{d}x\, \exp\left(-\frac{1}{2}(A - C^2/B)x^2\right)} \tag{12d}$$

If we set $\sigma^2 = 1/(A - C^2/B)$, we arrive at

$$\sigma_x^2 = \frac{\int \mathrm{d}x\, x^2 \exp\left(-\frac{1}{2}x^2/\sigma^2\right)}{\int \mathrm{d}x\, \exp\left(-\frac{1}{2}x^2/\sigma^2\right)} \tag{13}$$

, which as we have seen in previous homework sets and solutions implies

$$\sigma_x^2 = \langle x^2 \rangle = \sigma^2 = \frac{1}{A - \frac{C^2}{B}} = \frac{B}{AB - C^2}. \tag{14}$$

The hint was used in step 3. In step 4 we integrated out the y-dependence because the integration boundaries of $\pm\infty$.

Problem 3 (graded by Stephen) - 35 points

(a) - 5 points

It is not reasonable to assume that all model parameters have constant priors (-∞ to ∞). We know that t_s should occur before all of the observations (for the signal to be causal, at least to within errors on the data, i.e. $t_s < 322$ s) and velocity must be positive (thus $v > 0$). We can leave these priors from $-\infty$ to 322 s and 0 to ∞, respectively, even though we know that there is a physical upper limit (for example v cannot be faster than the speed of light). The prior for z_s should be restricted to positve or negative values, depending on our sign convention for the height coordinate, because the earthquake occured below the surface. Practically, we can leave the priors on x_s, y_s as constant priors, but in reality they also would be limited to the size of the earth.

(b) - 5 points

We can incorporate this information as a prior by multiplying our likelihood $P(\{t_k\}|m)$ by a prior for velocity $P\{v\}$. In this case we will use a Gaussian prior with $\mu = 4.8$ and $\sigma = 0.1$. Other variables get uniform priors. Our expression will look like:

$$\begin{aligned}
P\{v\} &= e^{-\frac{1}{2}(\frac{v-\mu}{\sigma})^2} \\
P(m|\{t_k\}) &= P(\{t_k\}|m)P\{v\} \\
P(m|\{t_k\}) &= e^{-F(m)}e^{-\frac{1}{2}(\frac{v-\mu}{\sigma})^2} \\
P(m|\{t_k\}) &= e^{-F(m)-\frac{1}{2}(\frac{v-\mu}{\sigma})^2}
\end{aligned}$$

Note also that the the velocity should be positive. This is only approximately the case for the Gaussian prior, but is not a problem, because the probability for negative velocities is very (!) small.

(c.i) - 5 points

$$\begin{aligned}
\int_{v_1}^{v_2} P(v)dv &= \int_{v_1}^{v_2} \frac{1}{v} dv = \ln(v_2) - \ln(v_1) \\
\int_{kv_1}^{kv_2} P(v)dv &= \int_{kv_1}^{kv_2} \frac{1}{v} dv = \ln(kv_2) - \ln(kv_1) \\
&= \ln(k) + \ln(v_2) - \ln(k) - \ln(v_1) \\
&= \ln(v_2) - \ln(v_1)
\end{aligned}$$

(c.ii) - 5 points

We can again incorporate this information as a prior in our expression for $P(m|\{t_k\})$ by multiplying it times our likelihood (just like in part b). Our expresssion becomes:

$$P(m|\{t_k\}) = e^{-F(m)}\frac{1}{v}$$

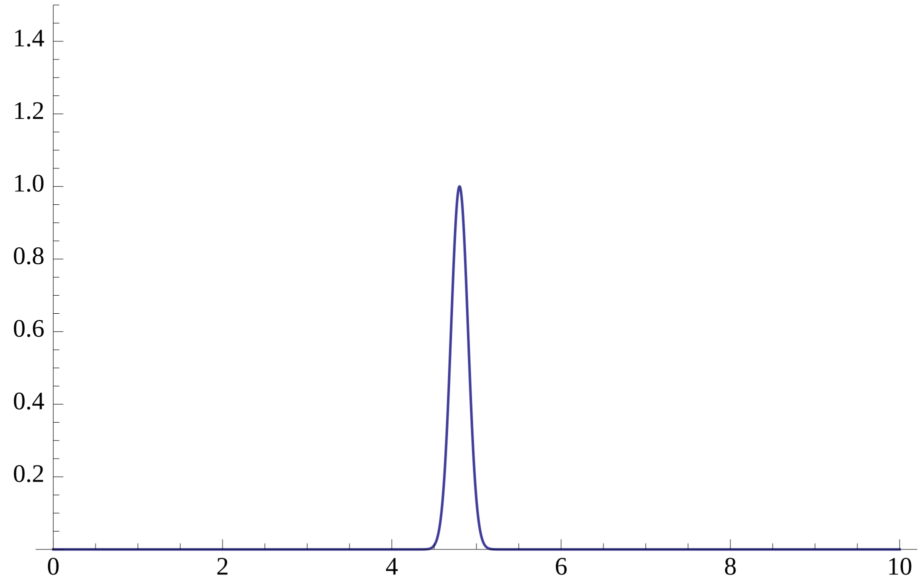

Figure 2: Plot of the Gaussian prior for velocity.

(d) - 5 points

We want to incorporate both independent pieces of information into our prior. In this case we can multiply the two priors together. Our new expression becomes:

$$P(m|\{t_k\}) = \frac{1}{v}e^{-F(m)}e^{-\frac{1}{2}(\frac{v-\mu}{\sigma})^2}$$

Let's plot the two priors separately and then together to see which information dominates the result. Figure 2 is the Gaussian prior. Figure 3 is the scale independent prior. Figure 4 is their combination. We can see that the Gaussian prior is scaled by the scale invariant prior, but still mainly retains it's shape and thus dominates the result.

(e) - 5 points

It doesn't matter in what order we do things as long as priors for analysis are not biased by our data (and they shouldn't be). If we did the experiment without talking to our friends, we would have no extra information and would use uniform priors. If we talk to them first, we use the non-uniform prior we derived in the previous step.

(f) - 5 points

Previous expression: $P(m|\{t_k\}) = e^{-F(m)}$
New expression: $P(m|\{t_k\}) = \frac{1}{v}e^{-F(m)}e^{-\frac{1}{2}(\frac{v-\mu}{\sigma})^2}$

Previous misfit function: $F(m)$
To find the new misfit function we will need to manipulate our expression to get everything into a single exponential. Let's start with the $\frac{1}{v}$ part:

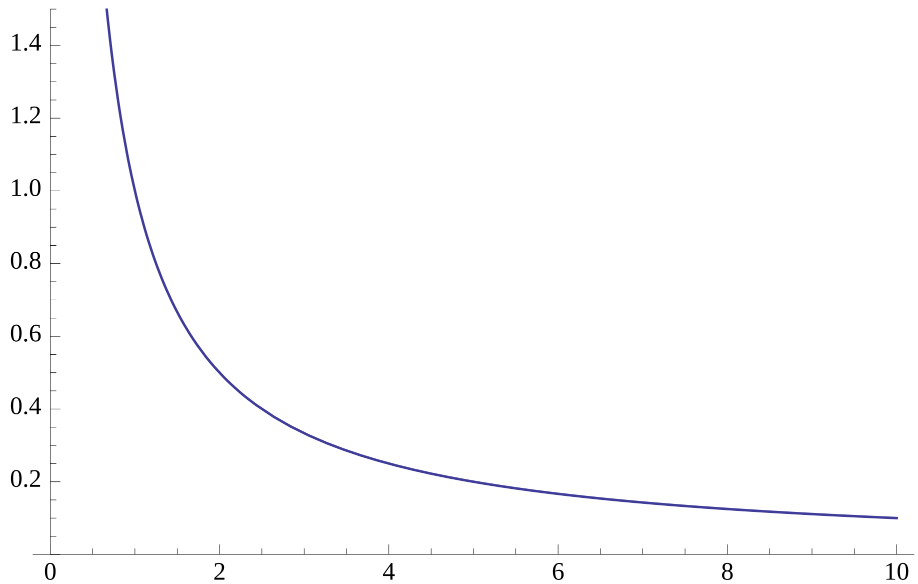

Figure 3: Plot of the scale invariant prior for velocity.

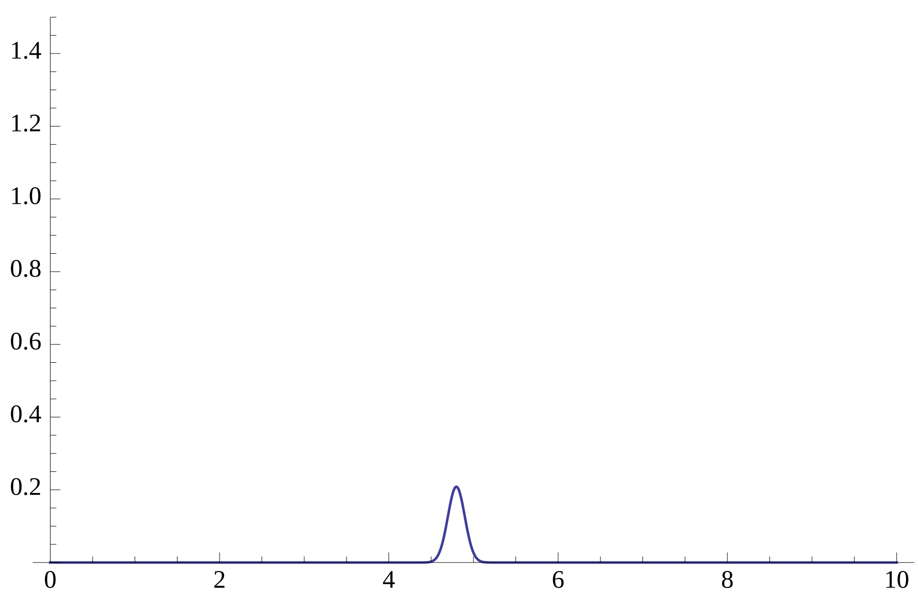

Figure 4: Plot of combination of the Gaussian and scale invariant priors for velocity. Note how the plot is scaled but retains basically the same shape as the pure Gaussian.

$$\frac{1}{v} = e^{(\ln(\frac{1}{v})} = e^{-\ln(v)}$$
$$P(m|\{t_k\}) = e^{-\ln(v)}e^{-F(m)}e^{-\frac{1}{2}(\frac{v-\mu}{\sigma})^2}$$

New misfit function: $F(m) + \ln(v) + \frac{1}{2}(\frac{v-4.8}{0.1})^2$

We only need to make a couple simple changes to our code from HW2. Instead of just using the L2 norm as our error function, we now need to add our extra two terms to our misfit function. We also need to slightly change our γ and our Hessian. Our old Hessians do not take into account the given standard deviation of the data: $\sigma = 0.01$. We can take this into account by doing a simple modification on them:

$$H_{new} = \frac{H_{old}}{0.01^2}$$
$$\gamma_{new} = \frac{\gamma_{old}}{0.01^2}$$

We will add a term to the last entries of γ and the Hessian to account for the new priors on v_s.
Add the first derivative of our new part of the misfit to γ:

$$\frac{dF}{dv} = \frac{1}{v} + \frac{v-\mu}{\sigma^2}$$

And the second derivative to the last entry of the Hessian:

$$\frac{d^2F}{dv^2} = \frac{-1}{v^2} + \frac{1}{\sigma^2}$$

Here is the full code:

```
function hw4p3()
rng(0); % this generate determined random variable
xi=[0 10 15 6 -7 3]';
yi=[0 0 6 13 10 7]';
zi=[0 0 0 0 0 0]';
ti=[322.418 321.031 321.228 323.093 324.415 322.706]';

sig = 0.1;
mu = 4.8;

M0=do_one_ti(xi,yi,zi,ti);
ti_0=predict(xi,yi,zi,M0);

function M=do_one_ti(xi,yi,zi,ti)

M=[300 20 -10 10 2]'; % initial guess
for step = 1:10
    M(5);
    grad=zeros(5,1);
    hess=zeros(5,5);
```

```
    for i=1:length(xi)
        [tmp_grad, tmp_hess]=get_grad_hessian(xi(i),yi(i),zi(i),ti(i),M);
        grad = grad + tmp_grad;
        hess = hess + tmp_hess;
    end

    %adjust the final values of gamma and the Hessian to account for the
    %velocity priors
    grad(end) = grad(end) + 1/(M(5)) + (M(5)-4.8)/(0.1)^2;
    hess(end,end) = hess(end,end) + 1/0.1^2 - 1/M(5).^2;

    err(step) = (norm(predict(xi,yi,zi,M)-ti)^2);
  %  if(step > 1 && err(step) > err(step-1))
  %      break;
  %  end
M = M- inv(hess)*grad;
end
M
plot(err);

function [grad,hess] = get_grad_hessian(xi,yi,zi,ti,M)
ts=M(1); xs=M(2); ys=M(3); zs=M(4); v=M(5);
R=sqrt((xs-xi)^2 + (ys-yi)^2 + (zs-zi)^2);
nx=(xs-xi)/R;
ny=(ys-yi)/R;
nz=(zs-zi)/R;
e=(ts+R/v-ti);

sig = 0.1;

grad = e*[1 nx/v ny/v nz/v -R/v^2]';
sen = [1 nx/v ny/v nz/v -R/v^2]';
hess = sen*sen'; % approximate hess

%need to add ajustment for sigma given for the data
hess = hess/0.01^2;
grad = grad/0.01^2;

%hess(end,end) = hess(end,end) + 1/sig^2 - 1/v^2;

function r=predict(xi,yi,zi,M)

ts=M(1); xs=M(2); ys=M(3); zs=M(4); v=M(5);
r=xi*0;
for i=1:length(xi)
    R=sqrt((xs-xi(i))^2 + (ys-yi(i))^2 + (zs-zi(i))^2);
    r(i)=(ts+R/v);
end
```

Our best fit solution is now:

$$m = \begin{pmatrix} 313.9738 \\ 30.9907 \\ -17.7781 \\ 21.7237 \\ 4.9581 \end{pmatrix}$$

Our old best fit solution:

$$m = \begin{pmatrix} 315.147372215551 \\ 30.3124982007349 \\ -17.1481612373344 \\ 15.9867180697526 \\ 5.26992730665649 \end{pmatrix}$$

We can see that the addition of these priors did change our solution for some parameters a small amount (mostly a change in the best fit for z is observed). As we would expect, the solution for v has been pushed closer to 4.8 (given our Gaussian prior around this value) and the other parameters have adjusted accordingly.