Lecture 7: Model Fitting

1 Noise Models

In Lectures 3,4,5 and 6 we addressed various problems using linear formulations that approximately solve the governing algebraic equations. While this is an easy approach we can in general not expect that these equations can be solved exactly. Since the appearance of a patch changes when the viewpoint changes, exact positioning of corresponding points is not possible, see Figure 1. Additionally, since matching is done automatically we have to expect that some matches are incorrect causing large deviations from the model. Therefore, our point measurements will in practice always be corrupted by noise of various forms and levels and in general approximate solutions based on DLT will not give the "best" possible fit to the observed data.

In this lecture we will derive formulations that gives the "best" fit under the assumption of Gaussian noise. The resulting problems are in general more difficult to solve than the formulations that we have used previously. In many cases they can only be locally optimized. Therefore the linear approaches are still very useful since they provide an easy way of creating a starting solution.

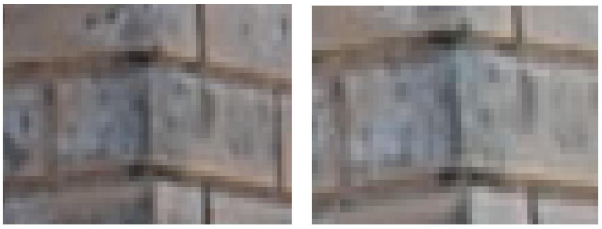

Figure 1: Two patches extracted from images with slightly different viewpoint. Exact localization of corresponding points is made difficult because of slight appearance differences and limited image resolution.

2 Line Fitting

What is meant by the "best" fit depends on the particular noise model. In this section we will consider two different noise models and show that they lead to different optimization criteria. For simplicity we will consider the problem of line fitting since this leads to closed form solutions.

2.1 Linear Least Squares

Suppose that (x_i, y_i) are measurements of 2D-points belonging to a line $y = ax + b$. Furthermore, we assume that y_i is corrupted by Gaussian noise, that is,

$$y_i = \tilde{y}_i + \epsilon_i \tag{1}$$

where $\epsilon_i \in \mathcal{N}(0, 1)$ (Gaussian noise with mean 0 and standard deviation 1) and $\tilde{y}_i$ is the true y-coordinate. Our goal is to estimate the line parameters a and b for which the measurements y_i are most likely. Since $\epsilon_i \in \mathcal{N}(0, 1)$,

its probability density function is

$$p(\epsilon_i) = \frac{1}{\sqrt{2\pi}} e^{-\epsilon_i^2/2}. \tag{2}$$

Furthermore, if we assume that the ϵ_i, $i = 1, ..., n$ are independent of each other then their joint distribution is

$$p(\epsilon) = \prod_{i=1}^{n} p(\epsilon_i), \tag{3}$$

where $\epsilon = (\epsilon_1, \epsilon_2, ..., \epsilon_n)$. Since $\epsilon_i = y_i - \tilde{y}_i$ we can compute the likelihood of the measurements by

$$p(\epsilon) = \prod_{i=1}^{n} p(\epsilon_i) = \prod_{i=1}^{n} p(y_i - \tilde{y}_i) = \prod_{i=1}^{n} p(y_i - (ax_i + b)) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-(y_i - (ax_i + b))^2/2}. \tag{4}$$

We now want to find the the line parameters a and b that make these measurements most likely. To simplify the maximization we maximize the logarithm of the likelihood

$$\log\left(\prod_{i=1}^{n} p(\epsilon_i)\right) = -\sum_{i=1}^{n} \frac{(y_i - (ax_i + b))^2}{2} + \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi}}\right). \tag{5}$$

Since the second term does not depend on a of b this is the same as minimizing

$$\sum_{i=1}^{n} (y_i - (ax_i + b))^2. \tag{6}$$

In matrix form we can write this as

$$\left\| \underbrace{\begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix}}_{=A} \begin{pmatrix} a \\ b \end{pmatrix} - \underbrace{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}}_{=B} \right\|^2. \tag{7}$$

The minimum of this expression can be computed using the normal equations

$$\begin{pmatrix} a \\ b \end{pmatrix} = (A^T A)^{-1} A^T B, \tag{8}$$

which we will derive in Lecture 9. The geometric interpretation of (6) is that under this noise model the vertical distance between the line and the measurement should be minimized, see Figure 2.

2.2 Total Linear Least Squares

Next we will assume that we have noise in both coordinates, that is,

$$\begin{pmatrix} x_i \\ y_i \end{pmatrix} = \begin{pmatrix} \tilde{x}_i \\ \tilde{y}_i \end{pmatrix} + \delta_i, \tag{9}$$

where $\delta_i \in \mathcal{N}(0, I)$ and $a\tilde{x}_i + b\tilde{y}_i = c$. The δ_i now belong to a two dimensional normal distribution with probability density function

$$p(\delta_i) = \frac{1}{2\pi} e^{-\|\delta_i\|^2/2}. \tag{10}$$

The log likelihood function is

$$\sum_{i=1}^{n} \log(p(\delta_i)) = -\sum_{i=1}^{n} \frac{(x_i - \tilde{x}_i)^2 + (y_i - \tilde{y}_i)^2}{2} + \sum_{i=1}^{n} \log(\frac{1}{2\pi}). \tag{11}$$

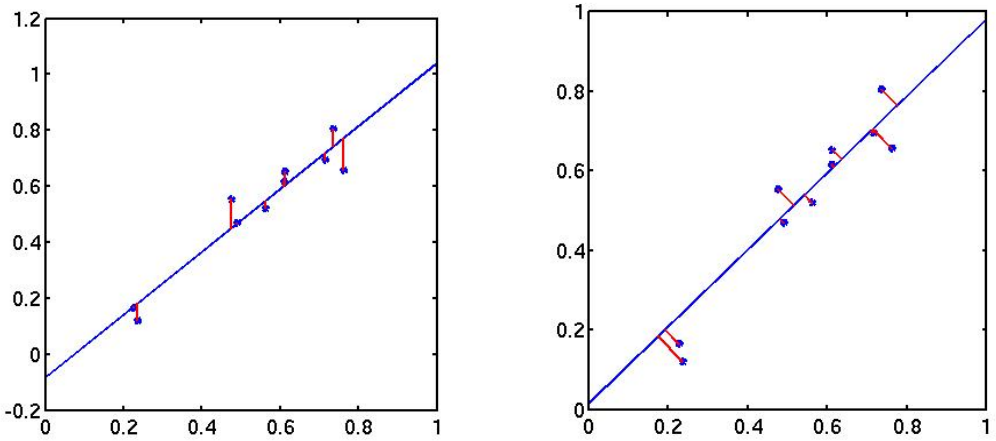

Figure 2: Left: The vertical distances between the line and the measured points are minimized in (6). In contrast, the minimal distances between the line and the measured points are minimized in (12).

Therefore, to maximize the likelihood we need to minimize

$$\sum_{i=1}^{n}((x_i - \tilde{x}_i)^2 + (y_i - \tilde{y}_i)^2), \tag{12}$$

where $a\tilde{x}_i + b\tilde{y}_i = c$. The point $(\tilde{x}_i, \tilde{y}_i)$ can be any point on the line, however since we are minimizing (12) we can restrict it to be the closest point on the line. The expression (12) then becomes the distance between (x_i, y_i) and the line. This distance can be expressed using the distance formula as

$$\frac{|ax_i + by_i + c|}{\sqrt{a^2 + b^2}}. \tag{13}$$

Without loss of generality we ca assume that $a^2 + b^2 = 1$, and therefore we need to solve

$$\min \quad \sum_{i=1}^{n}(ax_i + by_i + c)^2 \tag{14}$$

$$s.t. \quad a^2 + b^2 = 1. \tag{15}$$

This problem is often referred to as the **total linear least squares problem**.

2.2.1 Solving the Total Least Squares Problem

To solve (14),(15) we first take derivatives with respect to c. This shows that the the optimal solution must fulfill

$$c = -(a\bar{x} + b\bar{y}), \tag{16}$$

where $\bar{x}$ and $\bar{y}$ are the mean values

$$(\bar{x}, \bar{y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i, y_i). \tag{17}$$

Substituting into (14) we get

$$\min \quad \sum_{i=1}^{m}(a(x_i - \bar{x}) + b(y_i - \bar{y}))^2 \tag{18}$$

$$\text{such that} \quad 1 - (a^2 + b^2) = 0. \tag{19}$$

By forming the matrix

$$M = \sum_{i=1}^{m} \begin{pmatrix} (x_i - \bar{x})^2 & (x_i - \bar{x})(y_i - \bar{y}) \\ (x_i - \bar{x})(y_i - \bar{y}) & (y_i - \bar{y})^2 \end{pmatrix}, \tag{20}$$

we can write this problem as

$$\begin{aligned} \min \quad & t^T M t \qquad (21) \\ \text{such that} \quad & 1 - t^T t = 0, \qquad (22) \end{aligned}$$

where t is a 2×1 vector containing a and b. This is a constrained optimization problem of the type

$$\begin{aligned} \min \quad & f(t) \qquad (23) \\ \text{such that} \quad & g(t) = 0. \qquad (24) \end{aligned}$$

According to Persson-Böiers, "Analys i flera variabler" and the method of Lagrange multipliers the solution of such a system has to fulfill

$$\nabla f(t) + \lambda \nabla g(t) = 0. \tag{25}$$

Therefore the solution of (21)-(22) must fulfill

$$2Mt + \lambda(-2t) = 0 \Leftrightarrow Mt = \lambda t. \tag{26}$$

That is, the solution t has to be an eigenvector of the matrix M. Furthermore, inserting into (21), and using that $t^T t = 1$ we see that it has to be the eigenvector corresponding to the smallest eigenvalue.

2.3 Outliers and Robust Loss-Functions

In structure from motion problems we typically have two types of noise. Appearance changes that makes it hard to exactly localize corresponding points giving rise to displacements that are typically in the order of at most a few pixels. This type of noise is usually modeled as Gaussian. In addition to this we often have mismatches that give rise to very large errors. When minimizing the least squares objective such measurements can severely distort the results. Figure 3 shows a line fitting example with one outlier point. Because of the quadratic growth of the least squares criterion (6) points that are far away from the estimated line will have a proportionally larger effect on the estimate than points that are close to it. This gives a poor estimate shown in the middle of Figure 3. By using a robust loss-function that essentially removes the effect of the outlier it is possible obtain a better estimate of the line as shown to the right in Figure 3.

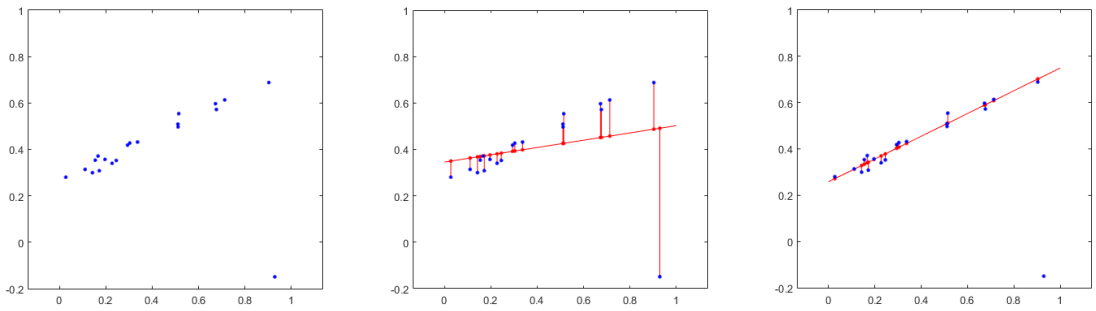

Figure 3: Line fitting with one outlier. *Left:* Measurements. *Middle:* Obtained model fit using the least squares objective (6). *Right:* Obtained model fit using a robust loss-function.

To find a suitable robust loss-function we can replace the assumption of normalized noise with a more general distribution with density function

$$p(\epsilon_i) = ce^{-\rho(\epsilon_i^2)}. \tag{27}$$

Here ρ is a function that should be less sensitive to outliers and c is some constant which ensures that the density function integrates to 1. As in Section 2.1 we assume that we have 2D measurements (x_i, y_i) of points belonging to a line $y = ax + b$ with noise in the y-coordinate $y_i = \tilde{y}_i + \epsilon_i$. To obtain the maximum likelihood estimate we should then minimize

$$\sum_{i=1}^{n} \rho\left((ax_i + b - y_i)^2\right). \tag{28}$$

Taking derivatives with respect to a and b we get

$$\sum_{i=1}^{n} \rho'\left((ax_i + b - y_i)^2\right) 2(ax_i + b - y_i)x_i = 0 \tag{29}$$

$$\sum_{i=1}^{n} \rho'\left((ax_i + b - y_i)^2\right) 2(ax_i + b - y_i) = 0 \tag{30}$$

respectively. In general these equations lack closed form solutions and have to be solved iteratively. None the less, we can still draw some simple conclusions from these expressions. In matrix form we can also write these two equations $M \begin{pmatrix} a \\ b \end{pmatrix} = m$, where

$$M = \sum_{i=1}^{n} \rho'(\epsilon_i^2) \begin{pmatrix} x_i^2 & x_i \\ x_i & 1 \end{pmatrix} \quad \text{and} \quad m = \sum_{i=1}^{n} \rho'(\epsilon_i^2) \begin{pmatrix} x_i y_i \\ x_i \end{pmatrix}. \tag{31}$$

Assuming that we somehow know what ϵ_i is at the optimal values of (a, b) then we have $\begin{pmatrix} a \\ b \end{pmatrix} = M^{-1}m$. We note that M and m can be seen as weighted sums of the matrices $\begin{pmatrix} x_i^2 & x_i \\ x_i & 1 \end{pmatrix}$ and $\begin{pmatrix} x_i y_i \\ x_i \end{pmatrix}$ respectively. If we use $\rho(t) = t$ then (28) reduces to the least squares solution and $\rho'(\epsilon_i^2) = 1$. By modifying the weight $\rho'(\epsilon_i^2)$ we can increase or reduce the impact of measurement i on the solution. For example selecting ρ so that its derivative is decreasing will reduce the effects of measurements with large errors. A common approach is to use a threshold τ. If we let $\rho_2(t) = \min(t, \tau^2)$ for some value τ then $\rho_2'(\epsilon_i^2)$ will be zero if $|\epsilon_i|$ larger than τ and therefore this residuals completely disappears from (31). In contrast if $|\epsilon_i|$ is smaller than τ then $\rho_2'(\epsilon_i^2) = 1$. (Note that this is the derivative of $\rho_2(t)$ with respect to t where we have inserted $t = \epsilon_i^2$). Therefore this option will essentially remove large residuals and compute the least squares solution of remaining ones, which is what is illustrated to the right in Figure 3. Figure 4 shows the truncated loss function $\rho_2(\epsilon_i^2)$.

Strictly speaking the function $\min(\epsilon_i^2, \tau^2)$ will not be differentiable at $\epsilon_i = \tau$. However, we can always find approximations that are close to $\min(\epsilon_i^2, \tau^2)$ while still being differentiable. In Figure 4 we show $\rho_3(\epsilon_i^2)$ which is one such commonly used approximation.

As we noted above using decreasing derivatives generally makes the loss-function more robust to outliers. On the other hand optimization then becomes more difficult since this often leads to non-convex problems with local minimizers. If we for example use $\sum_{i=1}^{n} \rho_2\left((ax_i + b - y_i)^2\right)$ the derivatives will all be zero for any choice of a and b that makes all error residuals larger than τ. Thus an iterative algorithm, making use of the derivatives, will immediately get stuck if the starting point is not close enough to the global minimizer. Selecting ρ so that the derivatives are decreasing but never reach zero can help to achieve better convergence. The function $\rho_4(\epsilon_i^2)$ in Figure 4 shows one such example. However, without convexity there is in general no guarantee that we will reach the global minimizer. Initialization is therefore an important issue (which we will address in Lecture 8) for these methods.

The last example $h(\epsilon_i) := \rho_5(\epsilon_i^2)$ of Figure 4 is the Huber function. If we consider h as a function of ϵ_i it has the derivative

$$h'(\epsilon_i) = \begin{cases} 2\epsilon_i & |\epsilon_i| < b \\ 2b\text{sign}(\epsilon_i) & |\epsilon_i| \geq b \end{cases}. \tag{32}$$

Since h' is non-decreasing h is convex. Furthermore, since a sum of a set of convex functions is also convex the line fitting problem defined by (28) is convex. Therefore local optimization staring from any initialization is guaranteed to converge to the globally optimal solution. On the other hand the weights $\rho'(\epsilon_i^2)$ are decreasing (and tend to 0 as

$\epsilon_i \to \infty$). In that sense this is a good trade off between robustness and convexity. We will study global optimization further in Lecture 10.

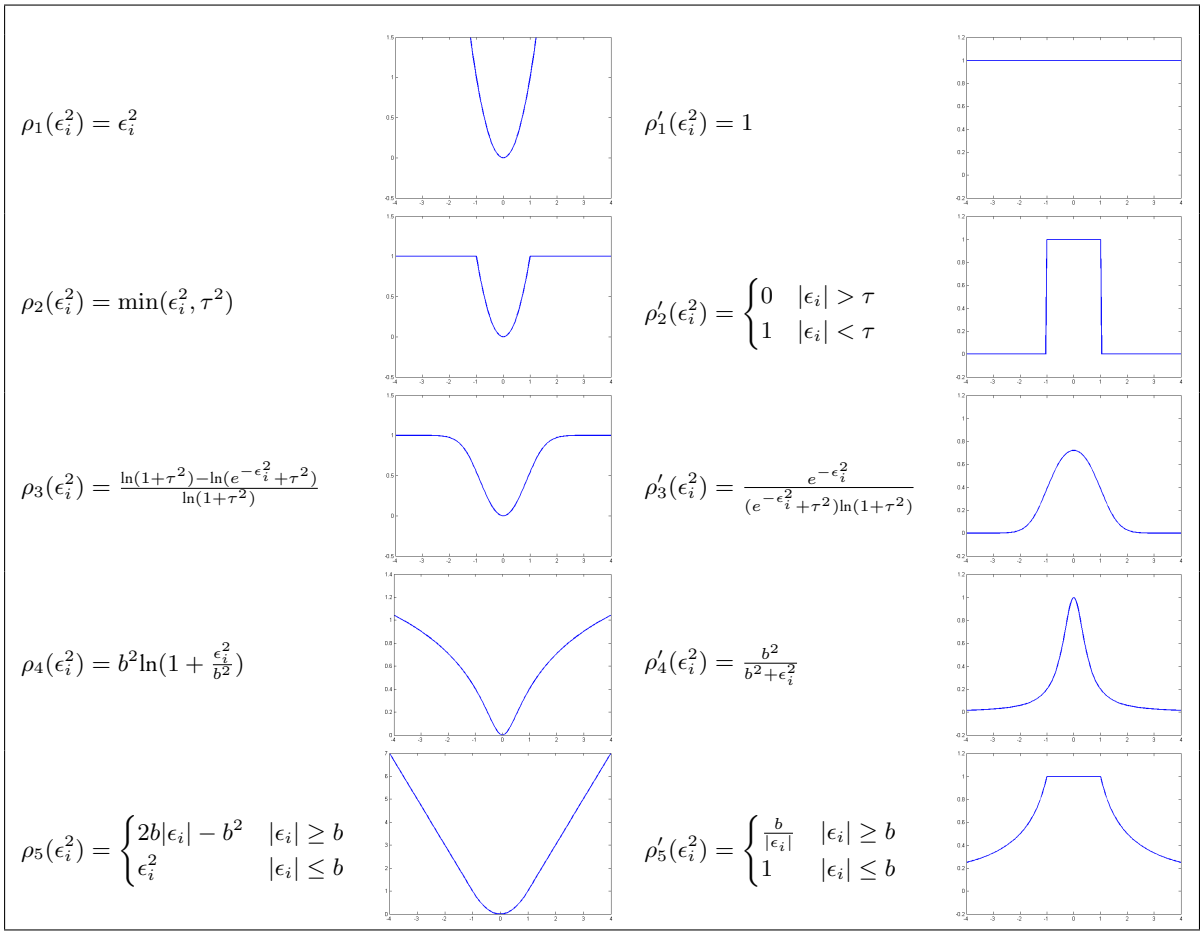

Figure 4: Some examples of loss-functions and the resulting weights. (Note that the derivative $\rho_j'(\epsilon_i^2)$ is obtained by differentiating $\rho_j(t)$ with respect to t and inserting $t = \epsilon_i^2$).

3 The Maximum Likelihood Solution for Camera Systems

In this section we derive the maximum likelihood estimator for our class projection problems. Suppose the 2D-point $x_{ij} = (x_{ij}^1, x_{ij}^2)$ is a projection in regular Cartesian coordinates of the 3D-point $\mathbf{X}_j$ in camera P_i. The projection in regular coordinates can be written

$$\left(\frac{P_i^1 \mathbf{X}_j}{P_i^3 \mathbf{X}_j}, \frac{P_i^2 \mathbf{X}_j}{P_i^3 \mathbf{X}_j} \right), \tag{33}$$

where P_i^1, P_i^2, P_i^3 are the rows of the camera matrix P_i. Also we assume that the observations are corrupted by Gaussian noise, that is,

$$(x_{ij}^1, x_{ij}^2) = \left(\frac{P_i^1 \mathbf{X}_j}{P_i^3 \mathbf{X}_j}, \frac{P_i^2 \mathbf{X}_j}{P_i^3 \mathbf{X}_j} \right) + \delta_{ij}, \tag{34}$$

and δ_{ij} is normally distributed with covariance I. The probability density function is then

$$p(\delta_{ij}) = \frac{1}{2\pi} e^{-\frac{1}{2}||\delta_{ij}||^2}. \tag{35}$$

Similarly to Section 2.2 we now see that the model configuration that maximizes the likelihood of the obtaining the observations $x_{ij} = (x_{ij}^1, x_{ij}^2)$ is obtained by solving

$$\min \sum_{i=1}^{n} \sum_{j=1}^{m} \left\| \left(x_{ij}^1 - \frac{P_i^1 \mathbf{X}_j}{P_i^3 \mathbf{X}_j},\ x_{ij}^2 - \frac{P_i^2 \mathbf{X}_j}{P_i^3 \mathbf{X}_j} \right) \right\|^2 . \tag{36}$$

where n is the number of cameras and m is the number of scene points. The geometric interpretation of the above expression is that the distance between the projection and the measured point in the image should be minimized, see Figure 5. Note that it does not matter which of the variables P_i and X_i we consider as unknowns, it is always the reprojection error that should be minimized.

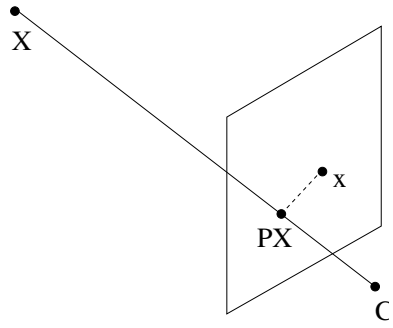

Figure 5: Geometric interpretation of the maximum likelihood estimate for projection problems. The dashed distance should be minimized.