The chapter provides an in-depth treatment of linear regression models and associated techniques. Linear regression is a workhorse model in statistics and machine learning for modeling the relationship between input variables (features) and a continuous output variable.

The basic linear regression model assumes a Gaussian or normal distribution on the output/response variable, with the mean being a linear combination of the input features. By introducing basis function expansions like polynomials, linear regression can also model non-linear relationships, though the model remains linear in the parameters.

Maximum likelihood estimation (MLE) is the standard method for estimating the parameters of the linear regression model. This involves minimizing the sum of squared residuals between the observed outputs and the model's predictions, known as the least squares method. The MLE solution has a geometric interpretation as the projection of the output vector onto the column space of the input data matrix.

While MLE minimizes training error, it can lead to overfitting or poor generalization performance on unseen test data, especially when the training data is noisy. To combat this, several regularization techniques are discussed. Ridge regression adds an l2 penalty term on the parameter values during estimation, shrinking them towards zero. This is equivalent to placing a Gaussian prior over the parameters in a maximum a posteriori (MAP) estimation framework. The amount of shrinkage is controlled by the regularization parameter lambda, with larger values encouraging simpler models.

The chapter shows how ridge regression can be computed efficiently by augmenting the input data with virtual samples from the Gaussian prior. It also draws connections between ridge regression and principal component analysis (PCA), illustrating how ridge automatically discards dimensions corresponding to small singular values of the input data.

For robustness to outliers in the data, the PDF covers techniques that replace the Gaussian assumption on the output noise with heavy-tailed distributions like the Laplace or Student's t distribution. Using the Laplace likelihood leads to minimizing the l1 norm of residuals, which can be solved as a linear program.

The beneficial effects of having large training datasets are also discussed. As the amount of training data increases, the effective model complexity automatically reduces, allowing more complex models to be used without overfitting. However, for small dataset sizes, simpler models with higher regularization are preferred.

From a computational perspective, the chapter provides numerical tricks like QR decomposition for stable least squares solutions and an SVD trick for efficient high-dimensional ridge regression.

The chapter comprehensively covers linear regression from theoretical foundations, to regularization and robustness extensions, computational considerations, and connections to

related techniques. Linear regression models form the basis for more complex regression models like generalized linear models and form a core component of many machine learning systems.